UNIVERSIDAD NACIONAL DE COLOMBIA

Relevant data representation by a kernel-based framework

Andrés Marino Álvarez-Meza

Universidad Nacional de Colombia
Faculty of Engineering and Architecture
Department of Electric, Electronic and Computing Engineering
Manizales, Colombia
2015

Relevant data representation by a kernel-based framework

Andrés Marino Álvarez-Meza

Dissertation submitted as a partial requirement to receive the grade of:
Doctor of engineering – automatics

Chair:
Prof. Germán Castellanos-Domínguez, Ph.D.

Academic Research Group:
Signal Processing and Recognition Group - SPRG

Universidad Nacional de Colombia
Faculty of Engineering and Architecture
Department of Electric, Electronic and Computing Engineering
Manizales, Colombia
2015

Un esquema núcleo para la representación relevante de datos

Andrés Marino Álvarez-Meza

Tesis sometida como requerimiento parcial para optar al título de:
Doctor en ingeniería – automática

Director:
Prof. Germán Castellanos-Domínguez, Ph.D.

Grupo de trabajo académico:
Control y Procesamiento Digital de Señales - GCPDS

Universidad Nacional de Colombia
Facultad de ingeniería y arquitectura
Departamento de ingeniería eléctrica, electrónica y computación
Manizales, Colombia
2015

Thank you, God, for being always there.

Acknowledgements

Thank you, God, for all the things that You provide me. Without Your guide my life would be meaningless. A very special thanks goes out to my parents and my brother for the support and love that they give to me in every moment of my life.

I would like to express my gratitude to the Prof. Germán Castellanos-Domínguez for his orientation during this research. Besides, I would like to thank all the Signal Processing and Recognition Group (SPRG) of the Universidad Nacional de Colombia sede Manizales for their suggestions and hours of academic discussion. Special thanks to my friends el Oso and el Mike, and to my master and undergraduate students: Luisa, Laura, Laura Ximena, Santiago, Andrés Eduardo, Sergio, David, Diego Fabian, and Yohan for their support during this research.

Furthermore, thank all the members of the Computational NeuroEngineering Laboratory (CNEL), University of Florida, Florida-USA, for their hospitality and help, and specially to Dr. José Principe by the opportunity to visit and to learn in this great research group. Alike, thank all the members of the Machine Learning Group (MLG), Université Catholique de Louvain, Louvain-la-Neuve-Belgium, specially, I would like to thank Prof. John Lee and Prof. Michel Verleysen for their hospitality and help.

Finally, I recognize that this research would not have been possible without the financial assistance of the Ph.D. studies scholarship: *Programa Nacional de Formación de Investigadores "Generación del Bicentenario"*, 2011/2012 funded by COLCIENCIAS. Moreover, some of the results of this research were supported by the project: *Evaluación asistida de potenciales evocados cognitivos como marcador del transtorno por déficit de atención e hiperactividad-TDAH-(código 20101008258)*, funded by COLCIENCIAS.

Andrés Marino Álvarez-Meza
2015

Abstract

Nowadays, the analysis of a large amount of data has emerged as an issue of great interest taking increasing place in the scientific community, especially in automation, signal processing, pattern recognition, and machine learning. In this sense, the identification, description, classification, visualization, and clustering of events or patterns are important problems for engineering developments and scientific issues, such as biology, medicine, economy, artificial vision, artificial intelligence, and industrial production. Nonetheless, it is difficult to interpret the available information due to its complexity and a large amount of obtained features. In addition, the analysis of the input data requires the development of methodologies that allow to reveal the relevant behaviors of the studied process, particularly, when such signals contain hidden structures varying over a given domain, e.g., space and/or time. When the analyzed signal contains such kind of properties, directly applying signal processing and machine learning procedures without considering a suitable model that deals with both the statistical distribution and the data structure, can lead in unstable performance results.

Regarding this, kernel functions appear as an alternative approach to address the aforementioned issues by providing flexible mathematical tools that allow enhancing data representation for supporting signal processing and machine learning systems. Moreover, kernel-based methods are powerful tools for developing better-performing solutions by adapting the kernel to a given problem, instead of learning data relationships from explicit raw vector representations. However, building suitable kernels requires some user prior knowledge about input data, which is not available in most of the practical cases. Furthermore, using the definitions of traditional kernel methods directly, possess a challenging estimation problem that often leads to strong simplifications that restrict the kind of representation that we can use on the data.

In this study, we propose a data representation framework based on kernel methods to learn automatically relevant sample relationships in learning systems. Namely, the proposed framework is divided into five kernel-based approaches, which aim to compute relevant data representations by adapting them according to both the imposed sample relationships constraints and the learning scenario (unsupervised or supervised task).

First, we develop a kernel-based representation approach that allows revealing the main input sample relations by including relevant data structures using graph-based sparse constraints. Thus, salient data structures are highlighted aiming to favor further unsupervised clustering stages. This approach can be viewed as a graph pruning strategy within a spectral clustering framework which allows enhancing both the local and global data consistencies for a given input similarity matrix.

Second, we introduce a kernel-based representation methodology that captures meaningful data relations in terms of their statistical distribution. Thus, an information theoretic

learning (ITL) based penalty function is introduced to estimate a kernel-based similarity that maximizes the whole information potential variability. So, we seek for a reproducing kernel Hilbert space (RKHS) that spans the widest information force magnitudes among data points to support further clustering stages.

Third, an entropy-like functional on positive definite matrices based on Renyi's definition is adapted to develop a kernel-based representation approach which considers the statistical distribution and the salient data structures. Thereby, relevant input patterns are highlighted in unsupervised learning tasks. Particularly, the introduced approach is tested as a tool to encode relevant local and global input data relationships in dimensional reduction applications.

Fourth, a supervised kernel-based representation is introduced via a metric learning procedure in RKHS that takes advantage of the user-prior knowledge, when available, regarding the studied learning task. Such an approach incorporates the proposed ITL-based kernel functional estimation strategy to adapt automatically the relevant representation using both the supervised information and the input data statistical distribution. As a result, relevant sample dependencies are highlighted by weighting the input features that mostly encode the supervised learning task.

Finally, a new generalized kernel-based measure is proposed by taking advantage of different RKHSs. In this way, relevant dependencies are highlighted automatically by considering the input data domain-varying behavior and the user-prior knowledge (supervised information) when available. The proposed measure is an extension of the well-known cross-correntropy function based on Hilbert space embeddings.

Throughout the study, the proposed kernel-based framework is applied to biosignal and image data as an alternative to support aided diagnosis systems and image-based object analysis. Indeed, the introduced kernel-based framework improve, in most of the cases, unsupervised and supervised learning performances, aiding researchers in their quest to process and to favor the understanding of complex data.

Keywords: signal processing, machine learning, relevant representation, kernel methods, information theoretic learning, automatics.

Resumen

Hoy en día, el análisis de datos se ha convertido en un tema de gran interés para la comunidad científica, especialmente en campos como la automatización, el procesamiento de señales, el reconocimiento de patrones y el aprendizaje de máquina. En este sentido, la identificación, descripción, clasificación, visualización, y la agrupación de eventos o patrones son problemas importantes para desarrollos de ingeniería y cuestiones científicas, tales como: la biología, la medicina, la economía, la visión artificial, la inteligencia artificial y la producción industrial. No obstante, es difícil interpretar la información disponible debido a su complejidad y la gran cantidad de características obtenidas. Además, el análisis de los datos de entrada requiere del desarrollo de metodologías que permitan revelar los comportamientos relevantes del proceso estudiado, en particular, cuando tales señales contiene estructuras ocultas que varían sobre un dominio dado, por ejemplo, el espacio y/o el tiempo. Cuando la señal analizada contiene este tipo de propiedades, los rendimientos pueden ser inestables si se aplican directamente técnicas de procesamiento de señales y aprendizaje automático sin tener en cuenta la distribución estadística y la estructura de datos.

Al respecto, las funciones núcleo (kernel) aparecen como un enfoque alternativo para abordar las limitantes antes mencionadas, proporcionando herramientas matemáticas flexibles que mejoran la representación de los datos de entrada. Por otra parte, los métodos basados en funciones núcleo son herramientas poderosas para el desarrollo de soluciones de mejor rendimiento mediante la adaptación del núcleo de acuerdo al problema en estudio. Sin embargo, la construcción de funciones núcleo apropiadas requieren del conocimiento previo por parte del usuario sobre los datos de entrada, el cual no está disponible en la mayoría de los casos prácticos. Por otra parte, amenudo la estimación de las funciones núcleo conllevan sesgos el modelo, siendo necesario apelar a simplificaciones matemáticas que no siempre son acordes con la realidad.

En este estudio, se propone un marco de representación basado en métodos núcleo para resaltar relaciones relevantes entre los datos de forma automática en sistema de aprendizaje de máquina. A saber, el marco propuesto consta de cinco enfoques núcleo, en aras de adaptar la representación de acuerdo a las relaciones impuestas sobre las muestras y sobre el escenario de aprendizaje (sin/con supervisión).

En primer lugar, se desarrolla un enfoque de representación núcleo que permite revelar las principales relaciones entre muestras de entrada mediante la inclusión de estructuras relevantes utilizando restricciones basadas en modelado por grafos. Por lo tanto, las estructuras de datos más sobresalientes se destacan con el objetivo de favorecer etapas posteriores de agrupamiento no supervisado. Este enfoque puede ser visto como una estrategia de depuración de grafos dentro de un marco de agrupamiento espectral que permite mejorar las consistencias locales y globales de los datos.

En segundo lugar, presentamos una metodología de representación núcleo que captura relaciones significativas entre muestras en términos de su distribución estadística. De este modo, se introduce una funcíón de costo basada en aprendizaje por teoría de la información para estimar una similitud que maximice la variabilidad del potencial de información de los datos de entrada. Así, se busca un espacio de Hilbert generado por el núcleo que contenga altas fuerzas de información entre los puntos para favorecer el agrupamiento entre los mismos.

En tercer lugar, se propone un esquema de representación que incluye un funcional de entropía para matrices definidad positivas a partir de la definición de Renyi. En este sentido, se pretenden incluir la distribución estadística de las muestras y sus estructuras relevantes. Por consiguiente, los patrones de entrada pertinentes se destacan en tareas de aprendizaje sin supervisión. En particular, el enfoque introducido se prueba como una herramienta para codificar las relaciones locales y globales de los datos en tareas de reducción de dimensión.

En cuarto lugar, se introduce una metodología de representación núcleo supervisada a través de un aprendizaje de métrica en el espacio de Hilbert generado por una función núcleo en aras de aprovechar el conocimiento previo del usuario con respecto a la tarea de aprendizaje. Este enfoque incorpora un funcional por teoría de información que permite adaptar automáticamente la representación utilizando tanto información supervisada y la distribución estadística de los datos de entrada. Como resultado, las dependencias entre las muestras se resaltan mediante la ponderación de las características de entrada que codifican la tarea de aprendizaje supervisado.

Por último, se propone una nueva medida núcleo mediante el aprovechamiento de diferentes espacios de representación. De este modo, las dependencias más relevantes entre las muestras se resaltan automáticamente considerando el dominio de interés de los datos de entrada y el conocimiento previo del usuario (información supervisada). La medida propuesta es una extensión de la función de cross-correntropia a partir de inmersiones en espacios de Hilbert.

A lo largo del estudio, el esquema propuesto se valida sobre datos relacionados con bioseñales e imagenes como una alternativa para apoyar sistemas de apoyo diagnóstico y análisis objetivo basado en imágenes. De hecho, el marco introducido permite mejorar, en la mayoría de los casos, el rendimiento de sistemas de aprendizaje supervisado y no supervisado, favoreciendo la precisión de la tarea y la interpretabilidad de los datos.

Palabras clave: procesamiento de señales, aprendizaje de máquina, representación relevante, métodos núcleo, aprendizaje por teoría de información, automatización.

Contents

List of Figures

List of Tables

Part I.

Preliminaries

1. Introduction

1.1. Motivation

Machine learning studies how an automated system can watch the environment, learn to distinguish patterns, and make decisions. The identification, description, classification, visualization, and clustering of events or patterns are important problems for engineering developments and scientific issues, such as: biology, medicine, economy, artificial vision, artificial intelligence, industrial production, and brain machine interfaces [132, 104, 101]. Therefore, in the last decades, machine learning community has been dedicating enormous research efforts to develop mathematical tools and methods to unfold the main patterns of a given process, allowing to the system to learn the relevant properties of the studied phenomenon.

In a local context, the Signal Processing and Recognition Group (SPRG) of the Universidad Nacional de Colombia has been working in the analysis of biosignal data, in order to propose machine learning methodologies to support the development of automatic systems for diagnostic assistance [35, 134, 2, 98, 111]. Recently, the research group is interested in the analysis of brain activity to detect cerebral pathologies and to support further rehabilitation procedures. In fact, worldwide machine learning and medical communities are interested in the treatment of such kind of diseases by using signal processing and machine learning tools to allow the user interacting with the environment from the analysis of its own brain signals [171]. Besides, the SPRG is also interested in the analysis of video and image data to support motion and biomedical image processing, for both health and interactive purposes, which are also a local and worldwide topic of interest [6, 4, 24, 154].

Nonetheless, it is difficult to interpret the available information due to its complexity and a large amount of obtained features. In addition, the analysis of the input data requires the development of methodologies that allow to unfold the relevant behaviors of the studied process, specially, when such signals contains different structures varying over the space and/or the time, e.g., nonstationary process [130, 171, 101]. When the input signals contains such kind of properties, directly applying signal processing and machine learning procedures without considering a suitable representation model that deals with both the statistical distribution and the data structure, could lead in unstable performance results.

From the local and the global context, it is necessary to continue the development of methodologies that allow to represent the input samples aiming to improve the performance of those kind of machines in terms of learning performance and data interpretability.

1.2. Problem statement

A learning system incorporates external information to improve its performance in a particular task. Commonly, the system tries to infer the rules that govern the studied phenomenon given a set of examples, e.g., the training set. Such rules can be in the form of a function that model the relation between input-output pairs or can consist of some observation about the structure of the spaces where the examples are represented. Thereby, the information provided to a machine learning system mainly comes from a set of observed inputs $\{x_n \in \mathscr{X} : n \in [1, N]\}$, where $\mathscr{X}$ is the input space and $N \in \mathbb{N}$ is the number of provided samples. It is important to note that in a machine learning system the input data can be directly processed from $\mathscr{X}$, raw data, or in some cases several measures (features) can be estimated from each provided sample x_n aiming to favor further learning procedures. Such a characterization is commonly known as the feature estimation stage. In addition, in some applications other sources of information, besides the input samples, may be given depending upon the task and the context in which the system is put, e.g., a set of output samples $\{y_n \in \mathscr{Y} : n \in [1, N]\}$, being $\mathscr{Y}$ a given output space.

Bearing this in mind, three learning scenarios can be described [144, 16]: *i)* The *supervised learning* scenario, where the goal is to find a rule of association between pairs $\{(x_n, y_n)\}$ of observed inputs and corresponding targets given an expert. As a result, learning occurs when the system effectively predicts the correct output for a non previously seen input. *ii)* The *reinforcement learning* scenario in which the system interacts with the environment by performing actions that feedback to the system in the form of rewards or punishments. In this case, the observed inputs are called states and the goal of adaptation is to estimate a suitable set of actions that will maximize the reward over time in terms of the observed state. *iii)* The unsupervised learning scenario where we can loosely say that the only available information is the observed inputs. Hence, the assumption is the presence of some data regularities since there is a process behind their generation. Then, the goal is to find such regularities and one motivation for doing so is that expressing the available information in terms of the underlying causes may favor further learning stages. So, the *unsupervised learning* view assumes a generative model for the observed data. However, one cannot argue the actual causes will be unveiled by the learning process since learning must be always accompanied by some constraints/assumptions about the input data that may or may not necessarily agree with the reality. That is, in the absence of assumptions there is no privileged a best feature representation, and that even the notion between patterns depends implicitly on assumptions that may or not be correct [39]. Regarding this, it is necessary to exploit the statistical regularities to encode the observed inputs into more compact representations and also reduce the effect of some external noise, which in absence of additional information is assumed to be unstructured, by exploiting the redundancy in the inputs.

On account of the fact that a compact representation is crucial when designing a learning system aiming to increase the accuracy while reducing the over-fitting, the input features

can be handcrafted based on the external knowledge about the domain of the application. Therefore, learning the relevant features by considering both the data regularities and the external knowledge (supervised information) provides system adaptability across different domains. Figure 1-1 illustrates the role of a suitable representation stage into a learning system.

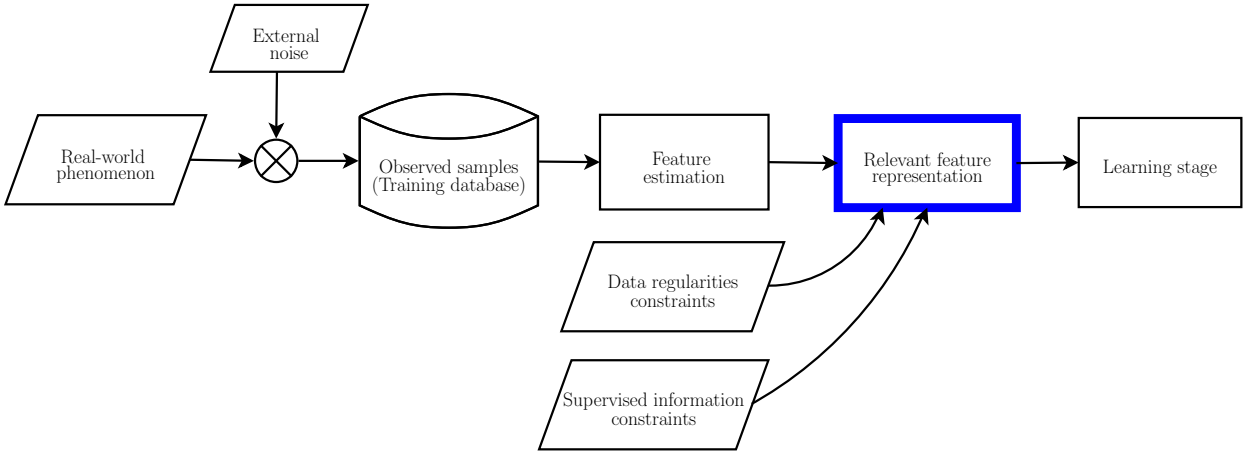


Figure 1-1.: The problem of learning and the role of a relevant data representation.

As seen, to apply the learning from examples paradigm to the problem of finding a suitable data representation, it is necessary to address two main issues: *i)* to assess the effectiveness of a representation with or without having the results on the subsequent tasks, and *ii)* to identify the desirable properties in the representation. So, when designing a learning system, it must be considered that the available information can be difficult to interpret and to process due to its complexity and the large amount of obtained features, not mentioning the noise drawbacks. Therefore, the available information, including the data regularities (structures) constraints and the user-prior knowledge regarding the studied phenomenon, must be exploited as well as possible to reveal the relevant properties of the task of interest. In turn, further learning procedures can be supported, e.g., classification, clustering, regression, prediction, etc.

1.3. Literature review on data representations

Throughout the literature, the representation problem have been approached from different perspectives leading to myriad of techniques. Overall, most of the state-of-the-art techniques overlap in terms of the criteria to assess the representation effectiveness, or in the properties/constraints that they convey. Thereby, the representation model can be studied from three main properties: *i)* the decomposition ability, *ii* the flexibility, and *iii)* the complexity.

Hence, the employed data representation can favor further learning performances with or without favoring the data interpretability and the whole learning complexity.

Traditional methods exploits solely geometrical properties of the data for clustering regardless of their occurrence. The most prominent approach is the k-means method, that clusters data points according to their minimal distance to geometrical centroid of point clouds, however, such methods can lead to inappropriate machine learning performances [39]. So, before clustering the main data patterns, a variety of methods consider the formation of features from the dimensions. Then, feature selection is the simplest approach as it consists of an inclusion or exclusion decision for each dimension. Although the relative importance of a feature is assessed, this information is only used to select the features. In addition, such a selection is commonly carried out based on a second order measure [35], which can be not appropriate to model complex data distributions. On the other hand, a simple feature selection approach is to find how informative each feature is for a given task and then select a set of informative, but not redundant features [161]. Similarly, a set of features may be obtained by backward or forward selection algorithms [139, 141]. In addition, heuristic search strategies have been increasingly used. Nonetheless, such approaches avoid conjectures about the feature interactions and evaluate sets of solutions simultaneously. Also, they are not prone to getting stuck in local minima [123].

On the other hand, linear projections can be used to find linear combinations of the input features [164, 113]. Indeed, the correlation or dependence between dimensions can also be used as features. Then, the well-known Principal component analysis (PCA) algorithm can reveal relevant patterns based on a variability criteria [75], however, the results from PCA on may be unsatisfactory, as the directions of largest variance may not contain any useful information in many real-world applications [84]. Also in the linear case, independent component analysis (ICA) optimizes a linear projection so the resulting vector has maximally independent elements [70]. The activity projected along each of these dimensions may be informative, but unlike the case for PCA, there is not a natural ordering for independent components, and the user is left to assess which components are meaningful.

In the supervised case, Fisher discriminant analysis (FDA) extensions use sample covariances from each class to form discriminative projections [133]. The optimal projection is a solution to a generalized eigenvalue problem that maximizes the spread between the means in different classes while minimizing the spread of samples within the same class. Beyond linear projections, techniques for multilineal processing exploit the intrinsic organization of dimensions along multiple modes [29].

Added to that, several data representations strategies have been proposed as nonlinear dimensionality reduction schemes. Then, the nonlinear dimensionality reduction techniques differ each other in the type of structure that they preserve, e.g., variance, dot products, dissimilarities (distances), similarities, or other local/global measures of proximity [87]. Thereby, the variety of the structure that they preserve has lead to the development of a large number of methods. The basic method is the well-known Classical Multidimensional

Scaling-(MDS), which maximize the dot product preservation [18]. Subsequently, some variants of the metric MDS appears, such as: Sammons's Nonlinear Mapping-(SNM) and the Curvilinear Component Analysis-(CCA) [64, 127]. These methods are based on notions like topology and neighborhood preservation, however, their main limitations come from the distortions between the distances measured in the input space and the distances measured in the manifold space. Some methods like Curvilinear Distances Analysis (CDA) and *Isometric Mapping*-(ISOMAP) [149, 88], are nonlinear methods derived from MDS, which use as metric the curvilinear or geodesic distance. This metric (geodesic distance) can measure good approximations of the distances along the manifold, without shortcuts as does the Euclidean distance. Nonetheless, the solution commonly exists as a global minimum, but, when the problem does not fit the model, its interpretation could be hazardous [89].

Nowadays, more developed methods aimed at preserving the data topology have been proposed from both spectral and divergence-based functions. The spectral approaches are represented by methods such as: Locally Linear Embedding-(LLE) [125], Laplacian Eigenmaps-(LEM) [14], Hessian LLE-(HLLE) [38], and Diffusion Maps-(DM) [105]. In LLE each datum is approximated by a linear combination of its neighbors in the input space and the obtained coefficients are then used to compute a low dimensional representation. LEM is a geometrically algorithm for constructing a representation for data sampled from a low dimensional manifold embedded in a higher dimensional space. HLLE achieves linear embedding by minimizing the Hessian functional on the manifold where the data set resides. DM is a based probabilistic interpretation of spectral clustering that use the eigenvectors of the normalized graph Laplacian. However, all of these methods require each input sample to be associated with only a single location in the low dimensional representation space.

With regard to the divergence-based methods, they are represented mainly by Stochastic Neighbor Embedding (SNE) [66] and its variants, i.e., t-SNE [155], and Jensen-Shannon Embedding (JSE) [83]. The main difference between spectral methods and SNE-based variants, is that SNE matches similarities that are computed both in the input feature space and the low dimensional representation, while spectral methods directly convert the pair-wise similarities defined in the feature space into inner products. Thus, SNE and its variants are based on similarity preservation instead of distance preservation and makes them robust against the phenomenon of norm concentration. Nonetheless, divergence-methods suffer from reaching distorted and overlapped latent spaces, moreover, the user must to tune for each dataset several parameters in order to obtain suitable representations and embeddings [22, 116].

On top of that, more versatile and powerful methods are related to artificial neural networks. Auto-encoders are one instance of artificial neural networks that are designed to solve non-linear dimensionality reduction problems [68]. Nevertheless, these methods require a priori selection of the architecture. Also, these models typically require extensive computation time to adapt parameters and hyper-parameters using non-convex optimization techniques. In a Bayesian modeling framework, certain models can be formulated only requiring convex optimization, particularly with certain choices in generalized linear models [117].

In the last two decades, there has been a growing interest for building different notions of similarity aiming to outperform the traditional second-order based dependence, e.g., the correlation. Thus, nonlinear notions of similarity have been proposed based on kernel functions. Indeed, the introduction of the support vector machine algorithm for pattern recognition rekindled the interest on this topic [132]. One of the major appeals of kernel methods is the ability to handle nonlinear operations on the data by indirectly computing an underlying nonlinear mapping to a space (Reproducing Kernel Hilbert Spaces (RKHS)) where linear operations can be carried out [11]. The linear solution corresponds to an universal approximation in the input space, and many of the related optimization problems can be posed as convex (no local minima) with algorithms that are still reasonably easy to compute. In this regard, powerful data representations can be built by highlighting relevant feature and/or sample dependencies within a kernel formulation.

Following the success of kernel machines and the ability to define kernels on general spaces, such as graphs [49], researchers have explored the kernel-based representations for several machine learning tasks. So, linear algorithms in the Hilbert-space can implement non-linear processing in the input space [26]. This is especially important for machine learning problems where the input space does not permit linear operations [112, 114].

Moreover, several approaches have been proposed multiple kernels representations within the machine learning contexts [82, 122, 55, 150]. Their main goal is to employ different sources of information to identify the similarities among samples, and then, a combination of these similarities is calculated by means of statistical kernel learning [132]. In this regard, a convenient approach is to consider that the calculated multiple kernel (Multi-Kernel Learning - MKL) is actually a convex combination of a basis kernel [55, 122]. Moreover, in [54] a Localized multiple kernel learning (LMKL) framework is presented, to extend the MKL framework to allow combining kernels with different weights in different regions of the input space by using a gating model. LMKL extracts the relative importance of kernels in each region whereas MKL gives their relative importance over the whole input space. Nonetheless, optimizing the MKL and LMKL functionals is not a straightforward tasks, besides, it is difficult to favor the input feature interpretability. In turn, even more flexibility can be achieved using kernel-based metric learning approaches [19, 46]. Then, such alternatives allow the kernel functions themselves to be adapted according to the studied task.

In this way, kernel based frameworks seem to be a suitable alternative to support the development of machine learning algorithms from a data representation point of view. Kernels are flexible tools that can be adapted to decompose the input data aiming to highlight relevant patterns. Nevertheless, the construction of methodologies that allow to deal with different structures requires the development of new automatic strategies, which ensure stable learning performances. Indeed, it is necessary to built a framework that allow incorporating different data constraints regarding the studied samples and the learning task to favor the data interpretability and the system performance.

1.4. Objectives

1.4.1. General objective

To develop a kernel-based representation framework that allows automatically disclosing relevant patterns from available input data. The developed framework must process the samples in a data-dependent manner to exploit their inherent structure and/or statistical distribution. In addition, such a representation framework must be matched to the type and complexity of both the signal and the learning task, which includes unsupervised or supervised scenarios. Thus, the proposed kernel representation must summarize and capture the main input patterns to support clustering and classification tasks, improving the learning performance in terms of task accuracy and data interpretability.

1.4.2. Specific objectives

- To develop a kernel-based representation approach to support automatic clustering aiming to reveal the main input sample relations by including its relevant topological structures. The proposed approach must highlight salient structures of the input data by finding the main pair-wise relations among samples to support further unsupervised clustering stages. In addition, the introduced approach must be tested in terms of both system accuracy and data interpretability.

- To develop an automatic kernel function estimation strategy to support clustering tasks based on an RKHS representation that allows including the input data statistical distribution to extract relevant pair-wise sample relationships. The methodology must adapt a kernel function to represent the samples in an RKHS favoring the separability between inherent data clusters.

- To propose a kernel-based representation strategy that consider the statistical distribution and the salient data structures from information theory-based constraints, aiming to reveal relevant input patterns in learning tasks. Particularly, the introduced approach must be useful to process high-dimensional samples by encoding both the relevant local and global input relationships.

- To develop an automatic clustering methodology based on a kernel representation that includes the input data statistical distribution and the user prior knowledge regarding the studied process, e.g., supervised information, to extract relevant pair-wise sample relationships. The methodology must highlight relevant features in terms of the studied task to favor the data interpretability and the system accuracy in supervised clustering tasks.

- To built a kernel-based representation approach that allows incorporating both the structure and the statistical distribution of the input samples taking advantage of

different RKHSs. Thus, relevant dependencies must be highlighted automatically by considering the input data domain-varying behavior and the user prior knowledge when available. The proposed approach must be tested in terms of both system accuracy and data interpretability in both unsupervised and supervised clustering tasks.

1.5. Contributions of this work

The present work is done within the kernel-based representation framework. We aim to provide some kernel strategies to learn automatically relevant data relations in machine learning systems. With this in mind, the framework can be adapted according to both the imposed sample relationships constraints and the learning scenario, including supervised and unsupervised tasks. Following, the main contributions of the work are described:

- A graph pruning approach, called *kernel alignment based graph pruning* (KAGP), is proposed within a spectral clustering framework to enhance both local and global data consistencies for a given input kernel-based similarity matrix. The proposed KAGP reveals the main input data relations by including both the data statistical distribution and its relevant structures by imposing graph-based sparse constraints. Hence, to weak all irrelevant relationships of the input kernel-based similarity matrix, KAGP quantifies the loss of information during the pruning process in terms of a kernel alignment-based function. Moreover, we encode the sample similarities using a compactly supported kernel that allows obtaining a sparse data representation to favor the graph partitioning problem. As a consequence, KAGP takes advantage of an initial guess of the relationships among points to identify all relevant connections.

- A new kernel function estimation strategy, termed *kernel function estimation based on information potential variability maximization* (KEIVP) is proposed aiming to capture meaningful data relations in terms of their statistical distribution. To this end, we make use of the intrinsic information potential variations from a Parzen-based probability density function estimator within and ITL framework. Namely, we seek for a RKHS maximizing the whole information potential variability in terms of the reproducing kernel parameters. In particular, the Gaussian kernel is study and we get a scale (kernel bandwidth) updating rule as a function of the information forces, which are induced by the kernel function applied over a finite sample set.

- An entropy-like functional on positive definite matrices based on Renyi's definition is employed to built a kernel-based representation that considers the statistical distribution and the salient data structures from information theory-based constraints. Our approach, termed *kernel-based entropy dimensionality reduction* (KEDR), is applied as a representation tool to measure the mismatch between high-dimensional and low-dimensional data representation spaces. The KEDR is a data-driven framework for

ITL based on infinitely divisible kernel functions. Therefore, the proposed approach employs estimators of entropy-like quantities for Gram matrices that can be computed by evaluating infinitely divisible kernels on pairs of samples to find a relevant representation space.

- A supervised kernel-based representation is introduced as a metric learning approach in a RKHS. In this way, our approach named *supervised kernel-based relevance analysis* (SKRA), is able to take advantage of the user-prior knowledge regarding the studied learning task. So, proposed SKRA incorporates the aforementioned ITL-based kernel function estimation strategy (KEIPV) to adapt automatically a relevant representation using both the supervised information and the input data distribution. As a result, relevant sample dependencies are highlighted by weighting the input features that mostly encode the supervised learning task.

- A new generalized kernel-based measure is proposed by representing the input samples in different RKHSs. As a result, relevant dependencies are highlighted automatically by considering the input data domain-varying behavior and the user-prior knowledge (supervised information) when available. Proposed measure, termed *generalized cross-correntropy* (GCC), can be viewed as an extension of the well-known cross-correntropy function by including Hilbert space embeddings. Moreover, the main connections between GCC and cross-covariance Hilbert space embedding are discussed. In addition, a dynamic enhancement of GCC is introduced within an adaptive learning scheme. The GCC approach can be incorporated as similarity function in clustering, classification, and prediction tasks.

fig. **1-2** describes the proposed kernel-based relevant representation framework in terms of the adapted approach according to both the imposed sample relationships constraints and the learning scenario (supervised and unsupervised tasks). Following, the mathematical preliminaries are presented in Chapter 2. Then, the proposed methodologies of the introduced framework, including the KAGP, KEIVP, KEDR, SKRA, and GCC approaches, are presented in Chapters 3, 4, 5, 6, and 7, respectively. Afterwards, the conclusion about this work as well as the possible future work are presented in Chapter 8. Finally, the academic discussion and the developed products (software prototypes) from the proposals of this thesis are presented in Chapter 9.

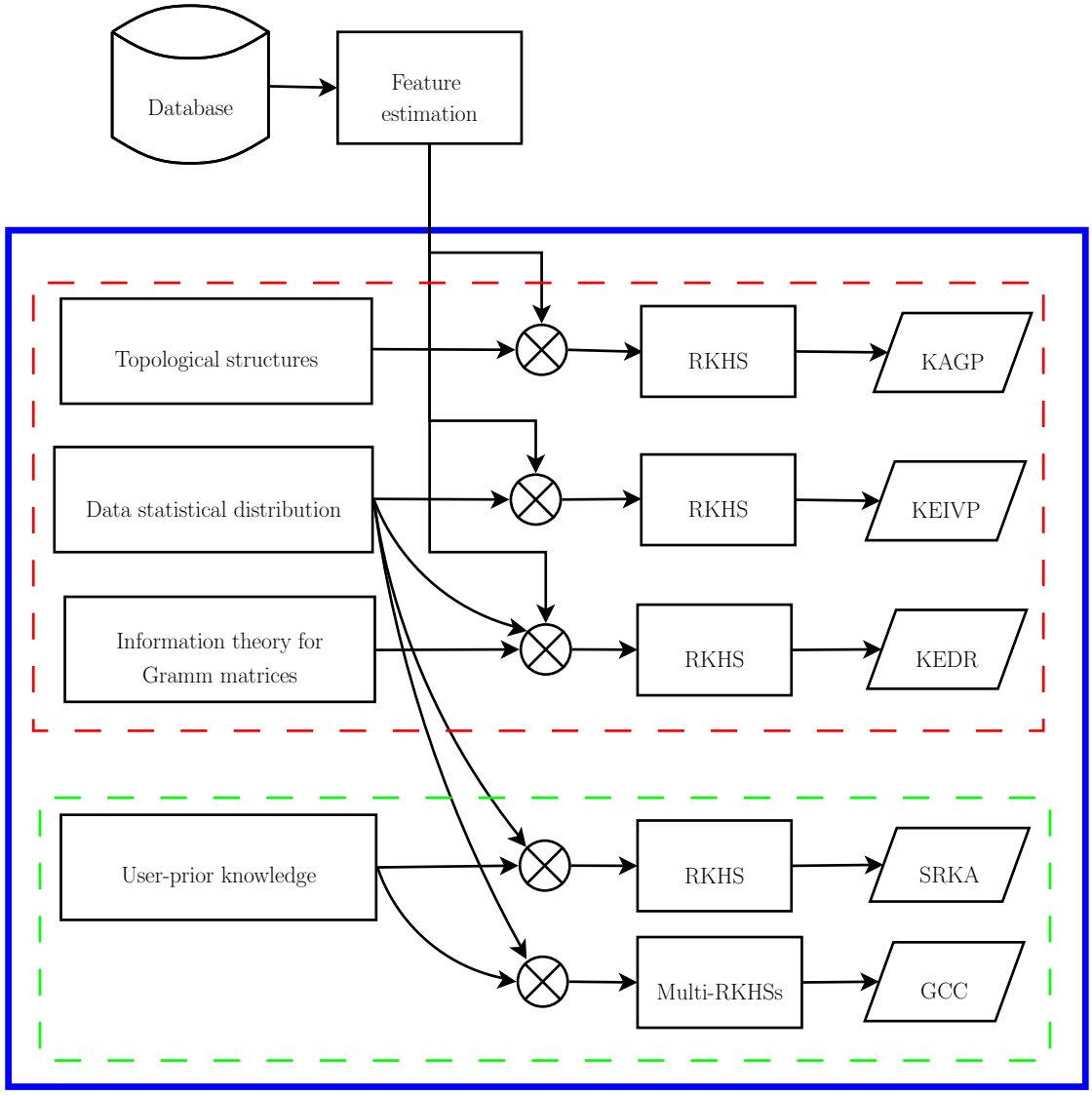


Figure 1-2.: —Kernel-based representation framework. – – Unsupervised approaches. – – Supervised approaches.

2. Mathematical preliminaries

In this chapter, we provide a brief account of the introductory concepts of the theory of kernel functions for representing the input samples in signal processing and machine learning systems. First, the formal definition and the necessary and sufficient conditions for a function to be a reproducing kernel are presented. Added to that the main concepts regarding information theoretic learning (ITL) are described, aiming to highlight the connections between kernel-based representations and information-based functions. The contents of this chapter are based on the papers and books by Aronszajn [10], Parzen [115], the Scholkopf and Smola [132], Sanchez [52] and Principe [121].

2.1. Reproducing kernel Hilbert spaces

Let $\mathscr{X}$ be a set and $\mathscr{F}$ be a vector space of functions from $\mathscr{X}$ to the field $\mathbb{F}$; in particular, let $\mathbb{F}$=$\mathbb{R}$. Then, there exits a reproducing kernel Hilbert space (RKHS) $\mathscr{H}$ on $\mathscr{X}$ over $\mathbb{R}$, if:

- $\mathscr{H}$ is a vector subspace of $\mathscr{F}$.
- $\mathscr{H}$ is endowed with an inner product, $\langle \cdot, \cdot \rangle_{\mathscr{H}}$, and is complete in the metric induced by it.
- For every $x \in \mathscr{X}$ and $f \in \mathscr{H}$, the linear evaluation functional $F_x : \mathscr{H} \to \mathbb{R}$, defined as $F_x(f) = f(x)$, is bounded.

From the Riez theorem [79], it is known that for any bounded functional H on a Hilbert space $\mathscr{H}$, there exists a unique vector $h \in \mathscr{H}$ such that: $H(f) = \langle h, f \rangle_{\mathscr{H}}$ for all $f \in \mathscr{H}$. In turn, for each evaluation functionals F_x there exist a corresponding vector $\kappa_x \in \mathscr{H}$. The bivariate function defined by:

$$\kappa(x, x') = \kappa_x(x') \tag{2-1}$$

is called a reproducing kernel for $\mathscr{H}$, with $x' \in \mathscr{X}$. So, it can be verified that

$$\kappa(x, x') = \langle \kappa_x, \kappa_{x'} \rangle_{\mathscr{H}} \tag{2-2}$$

and $\|F_x\|^2_{\mathscr{H}}=\|\kappa_x\|^2_{\mathscr{H}} = \langle \kappa_x, \kappa_{x'} \rangle_{\mathscr{H}}=\kappa(x,x)$, where $\|\cdot\|$ stands for the norm operator.

Let $\mathscr{H}$ be a RKHS on the set $\mathscr{X}$ with kernel κ. The linear span of $\{\kappa(x,\cdot) : x\in\mathscr{X}\}$ is dense in $\mathscr{H}$. This results from the fact that any function f orthogonal to the span of $\{\kappa(x,\cdot) : x\in\mathscr{X}\}$ must satisfy $\langle f, \kappa_x \rangle_{\mathscr{H}}$, and thus $f(x)=0$.

Lemma 2.1.1. *Let $\{f_n\} \subset \mathscr{H}$, being $n\in\mathbb{N}$ an index counter. If $\lim_{n\to+\infty} \|f_n - f\|_{\mathscr{H}} = 0$, then $f(x)=\lim_{n\to+\infty} f_n(x)$ for every $x\in\mathscr{X}$.*

Proof 2.1.1. *This is a simple consequence of the reproducing property and Cauchy-Schwarz inequality:*

$$|f_n(x) - f(x)| = |\langle f_n - f, \kappa_x \rangle_{\mathscr{H}}| \leq \|f_n - f\|_{\mathscr{H}} \|\kappa_x\|_{\mathscr{H}} \to 0$$

□

Proposition 2.1.1. *Let $\mathscr{H}_1$ and $\mathscr{H}_2$ be RKHS on $\mathscr{X}$ with kernels κ_1 and κ_2, respectively. If $\kappa_1(x,x')=\kappa_2(x,x')$ for all $x,x'\in\mathscr{X}$, then $\mathscr{H}_1=\mathscr{H}_2$ and $\|f\|_{\mathscr{H}_1}=\|f\|_{\mathscr{H}_2}$ for every f.*

Proof 2.1.2. *we can take $\kappa(x,x')=\kappa_1(x,x')=\kappa_2(x,x')$ and thus the $M_l=\text{span}\{\kappa_x\in M_l : x\in\mathscr{X}\}$ is dense in $\mathscr{H}_l$, and for any $f(x)=\sum_n \alpha_n\kappa_{x_n}(x)$ there is no regard about whether f belongs to either M_1 or M_2. Note that $\|f\|^2_{\mathscr{H}_1}=\sum_{n,n'} \alpha_n\alpha_{n'}\kappa(x_n,x_{n'})=\|f\|^2_{\mathscr{H}_2}$, and thus $\|f\|_{\mathscr{H}_1}=\|f\|_{\mathscr{H}_2}$ for every $f\in M_1=M_2$. If $f\in\mathscr{H}_1$, then there is a sequence of functions $\{f_n\} \subset M_1$ that converge to f in norm. Since $\{f_n\}$ is Cauchy in M_1 is also Cauchy in M_2, so by completeness of $\mathscr{H}_2$ there exist $g\in\mathscr{H}_2$ such that $f_n \to g$. Then, by Lemma 2.1.1 we have that $f(x)=\lim_{n\to+\infty} f_n(x)=g(x)$ for every $x\in\mathscr{X}$, thus every $f\in\mathscr{H}_1$ is also in $\mathscr{H}_2$ and vice versa, and $\mathscr{H}_1=\mathscr{H}_2$. Finally, we can extend $\|f\|_{\mathscr{H}_1}=\|f\|_{\mathscr{H}_2}$ to all $\mathscr{H}_1$ and $\mathscr{H}_2$.*

□

Thus, two different RKHSs do not have the same reproducing kernel. The following theorem shows an alternative way to express the reproducing kernel of a RKHS $\mathscr{H}$.

Theorem 2.1.1. *Let $\mathscr{H}$ have reproducing kernel κ. if $\{e_\lambda : \lambda\in\Lambda\}$ is an orthonormal basis of $\mathscr{H}$, then:*

$$\kappa(x,x') = \sum_{\lambda\in\Lambda} e_\lambda(x)e_\lambda(x'), \tag{2-3}$$

where the series converges point-wise.

Proof 2.1.3. *For a fixed $\{x_n\} \subseteq \mathscr{X}$, we have:*

$$\sum_{n,n'=1}^{N} \alpha_n\alpha_{n'}\kappa(x_n,x_{n'}) = \left\langle \sum_{n=1}^{N} \alpha_n\kappa_{x_n}, \sum_{n'=1}^{N} \alpha_{n'}\kappa_{x_{n'}} \right\rangle_{\mathscr{H}} = \left\| \sum_{n=1}^{N} \alpha_n\kappa_{x_n} \right\|_{\mathscr{H}} \geq 0$$

□

Added to that, the Moore's Theorem is introduced, which is the converse to the above result and provides us a characterization of a positive definite function to be a sufficient condition for the function to be the reproducing kernel of some RKHS.

Theorem 2.1.2. *Let $\mathscr{X}$ be a set and $\kappa : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ be a positive definite function. Then, there exits a RKHS $\mathscr{H}$ of functions on $\mathscr{X}$, such that, κ is the reproducing kernel of $\mathscr{H}$.*

Proof 2.1.4. *Consider the functions $\kappa_x(x')=\kappa(x,x')$ and the space W spanned by the set $\{\kappa_x : x\in\mathscr{X}\}$. The following bilinear map $B : W \times W \to \mathbb{R}$:*

$$B\left(\sum_i \alpha_n \kappa_{x_n}, \sum_{n'} \beta_{n'} \kappa_{x_{n'}}\right) = \sum_{n,n'} \alpha_n \beta_{n'} \kappa(x_n, x_{n'}),$$

where $\alpha_n\beta_{n'}\in\mathbb{R}$, is well defined on W. To support the above claim, notice that if $f(x)=\sum_n \alpha_n\kappa_{x_n}(x)$ is zero for all $x\in\mathscr{X}$, then by definition $B(f,\kappa_x)=0$ for all x. Conversely, if $B(f,w)=0$ for all $w\in W$, then by taking $w=\kappa_x$ it can be seen that $f(x)=0$. Then, B is well defined.

Since κ is positive definite $B(f,f)\geq 0$ and we see that $B(f,f)=0$ if and only if $B(w,f)=0$ for all $w\in W$, therefore $f(x)=0$ for all $\mathscr{X}$. Now we have shown that W is a pre-Hilbert space with inner product B. Let $\mathscr{H}$ denote the completion of W, we need to show that every element of $\mathscr{H}$ is function on $\mathscr{X}$. Let $h\in\mathscr{H}$ be the limit point of a Cauchy sequence $\{f_n\}\subseteq W$. By Cauchy-Schwarz inequality:

$$|f_n(x) - f_{n'}(x)| = |B(f_n - f_{n'}, \kappa_x)| \leq \|f_n - f_{n'}\|\kappa(x,x).$$

Therefore, the point-wise limit $h(x)=\lim_{n\to+\infty} f_n(x)$ is well defined. Concluding, let $\langle\cdot,\cdot\rangle_{\mathscr{H}}$ be the inner product on $\mathscr{H}$. Then, we have $\langle h,\kappa_x\rangle_{\mathscr{H}}=\lim_{n\to+\infty}\langle f_n,\kappa_x\rangle_{\mathscr{H}}=\lim_{n\to+\infty} B(f_n,\kappa_x)=h(x)$. Thus $\mathscr{H}$ is a RHKS with reproducing kernel κ.

□

Combining Proposition 2.1.1 with the Moore's Theorem (Theorem 2.1.2) shows the correspondence between RKHS's on the set $\mathscr{X}$ and positive definite functions on this set.

2.2. The covariance function

Consider a stochastic process $\{X(t) : t\in\tau\}$, where $X(t)$ are real random variables defined on a probability space $(\Omega,\mathscr{B},\mathscr{P})$ with bounded second order moments, that is:

$$\mathbb{E}_t\left\{|X(t)|^2\right\} = \int_\Omega |X(t)|^2 d\mathscr{P} < \infty, \tag{2-4}$$

where $\mathbb{E}\{\cdot\}$ stands for the expectation operator. Without loss of generality, we can consider random variables with zero mean, $\mathbb{E}_t\{X(t)\}=0$ for all $t\in\tau$. The covariance function is defined as:

$$R(t,t') = \mathbb{E}_{t,t'}\{X(t)X(t')\} = \int_{\Omega} X(t)X(t')d\mathscr{P}, \tag{2-5}$$

where $t,t'\in\tau$. It is easy to verify that $R:\tau\times\tau\to\mathbb{R}$ is a positive definite function and therefore defines a RKHS of functions on τ. A result originally due to Loeve and presented by Parzen in [115] showed a congruence map between the RKHS induced by the function R on L_2 space that corresponds to the completion of the span of the set $\{X(t):t\in\tau\}$ denoted by $L_2(X(t):t\in\tau)$.

Theorem 2.2.1. *Let $\{X(t):t\in\tau\}$ be a random process with covariance kernel R. Then $L_2(X(t):t\in\tau)$ is congruent with the RKHS $\mathscr{H}$ with reproducing kernel R. Furthermore, any linear map $\phi_R:\mathscr{H}\to L_2(X(t))$ which has the property that for any $f\in\mathscr{H}$ and any $t\in\tau$*

$$\mathbb{E}_t\{\phi_R(f)X(t)\} = f(t) \tag{2-6}$$

is the congruence from $\mathscr{H}$ onto $L_2(X(t))$, which maps $R(t,\cdot)$ into $X(t)$.

2.3. Reproducing kernel Hilbert spaces in machine learning

It is universally acknowledged that the study of positive definite kernels is a topic of interest for the machine learning community as a generalization of a well body of theory that has been developed for linear models. In this way, a positive definite kernel κ is an implicit way to represent the samples of the input space $\mathscr{X}$. Owing to there is a correspondence between κ and a RKHS of functions $\mathscr{H}$, the kernel can be understood as an indirect way to compute inner products between elements of a Hilbert space that are the result of mapping the elements of $\mathscr{X}$ to $\mathscr{H}$. So, there is a mapping function $\varphi:\mathscr{X}\to\mathscr{H}$ such that:

$$\kappa(x,x') = \langle\varphi(x),\varphi(x')\rangle_{\mathscr{H}}. \tag{2-7}$$

Regarding this, the space $\mathscr{H}$ can be viewed as a feature space and φ is called the feature map. Consequently, by performing linear operations in $\mathscr{H}$ it is possible to perform nonlinear manipulations in the input space $\mathscr{X}$, however, there is no need to perform any explicit computations in $\mathscr{H}$ (see Figure **2-1**).

Note that this idea is completely different to the congruence map introduced in Theorem 2.2.1. Then, an important property associated with the use of positive definite kernels in machine learning is the so-called representer theorem[77, 132]:

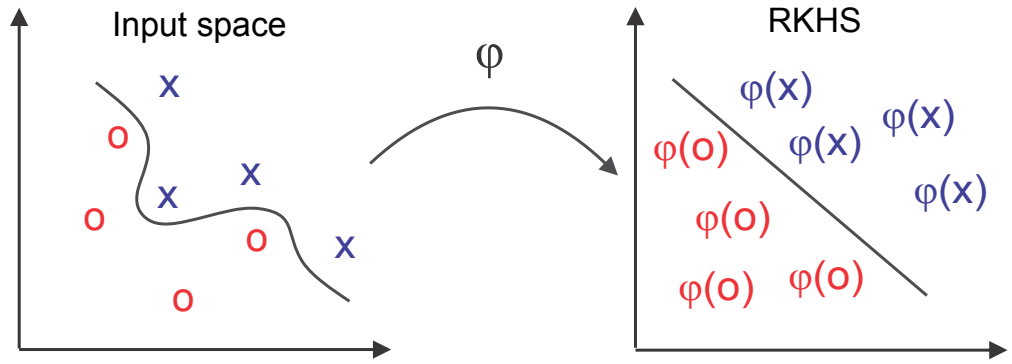


Figure 2-1.: kernel-based mapping.

Theorem 2.3.1. *Let $\Omega : [0, +\infty) \to \mathbb{R}$ be a strictly monotonic increasing function, $\mathscr{X}$ be a set, and $\epsilon : (\mathscr{X} \times \mathbb{R}^2)^N \to \mathbb{R} \cup \infty$ be an arbitrary loss function. Then, each minimizer $f \in \mathscr{H}$ of the regularized risk functional:*

$$\epsilon\left((x_1, y_1, f(x_1)), \ldots, (x_N, y_N, f(x_N))\right) + \Omega\left(\|f\|_{\mathscr{H}}^2\right), \tag{2-8}$$

admits a representation of the form:

$$f(x) = \sum_{n=1}^{N} \alpha_n \kappa(x_n, x), \tag{2-9}$$

where each $y_n \in \mathbb{R}$ is a given output associated with the input $x_n \in \mathscr{X}$.

Proof 2.3.1. *Let $S = \text{span}\{\kappa(x_n, \cdot) : x_n \in \mathscr{X}, n \in [1, N]\}$ denotes the subspace of $\mathscr{H}$ spanned by the N training samples. Consider the solution $f \in \mathscr{H}$, this solution can be written as: $f = f_S + f_{S^\perp}$, where $f_S \in S$, $f_{S^\perp} \in S^\perp$, and $\perp$ stands for the orthogonal symbol. Consequently, $f(x_n) = f_S(x_n) + f_{S^\perp}(x_n) = f_S(x_n) + 0$. Now, for the second term of the regularized risk funtional:*

$$\Omega\left(\|f\|_{\mathscr{H}}^2\right) = \Omega\left(\|f_S\|_{\mathscr{H}}^2 + \|f_{S^\perp}\|_{\mathscr{H}}^2\right),$$

since Ω is strictly monotonic increasing it is possible to see that the minimum will be achieved for $\|f_{S^\perp}\| = 0$, which implies that $f_{S^\perp} = 0$.

□

With this in mind, it is possible to conclude that the representer theorem basically states that the solution of the minimization of the regularized risk functional can be expressed in term of the so-called training sample $\{(x_n, y_n) : n \in [1, N]\}$. Therefore, it allows us to deal with problems that a first glance appear to be infinite dimensional. Nonetheless, the regularization does not prevent of having local multiple minima, such a property requires some extra conditions, namely, convexity.

Part II.

Unsupervised kernel-based representation approaches

3. Localized kernel representation based on graph pruning: a spectral clustering approach

In practice, the only presence of unlabeled data forces to develop unsupervised clustering techniques that search for hidden points of similar entities expressed in terms of a given pattern proximity measure. Regarding this, a suitable data representation is required to find out hidden patterns from input features. As a promising alternative in some disciplines –like data mining, pattern recognition, image processing, and machine learning– spectral clustering has been widely used for data grouping. So, topological restrictions are imposed into an unsupervised representation scheme aiming to extract relevant data dependencies. Mostly, these algorithms, which group points using matrix eigenvectors derived from the data, get better performance on complex datasets with non-convex clusters where traditional methods, e.g., K-means, frequently fail [28]. In fact, spectral clustering, which has its root in graph partitioning problems, can handle the optimization problem within the standard linear algebra framework for avoiding the local optima [162]. Also, spectral clustering formulations are very closed to kernel-based clustering approaches such as Kernel K-means, SOM, and Neural Gas [43]. Indeed, the objective function of a graph partitioning problem is mathematically equivalent to the weighted extension of the kernel K-means algorithm [36, 37].

As discussed in [169], spectral clustering approaches focus mainly on two issues: i) the extraction of an optimal partition and ii) choosing a suitable affinity matrix when building the graph representation. With respect to the former issue, it can be shown that the second smallest eigenvalue of the matrix estimated from the circuit netlist provides acceptable cut approximations. Nevertheless, the size of a graph subset is proportional to its number of vertices that is not always related to the within-cluster similarity. The following approaches have been proposed to cope with this drawback: the normalized cut criterion, termed NCut, measuring the total dissimilarity among groups and the overall one within clusters [135], the random-walk interpretation of spectral clustering [100], computation of the eigenvectors using matrix perturbation theory [109], among others. Nonetheless, the choice of a suitable affinity matrix for the graph building has received much less attention regardless its importance on the clustering performance [12]. To encode the pairwise relationships among samples in the affinity matrix correctly, the corresponding measure should be smooth with respect to the intrinsic data structure. Thus, samples belonging to the same group should

have high similarity each other and have enough space consistency. For this purpose, two assumptions on consistency are proposed in [169]: *i*) local consistency, meaning that nearby points in the space should have high similarity, *ii*) global consistency, meaning that samples in the same cluster should have high similarity. In spectral clustering, similarity graphs model the local neighborhood relationships between samples. Even though the choice of any of the regularly used graph constructions (ϵ-neighborhood, k-nearest graphs, and fully connected graphs) does not influence the clustering result [157], the similarity graph implementation requires fixing the ruling parameters. Since this procedure is not easy to automate, the free parameters of the similarity measure are manually adjusted in practice. Then, their corresponding partition graph matrix cannot be constructed to preserve the above assumptions on consistency. In particular for the widely-known Gaussian-based similarity, commonly known as Gaussian kernel, its bandwidth parameter is commonly set as a fraction of the pairwise distance estimated from the whole data. However, multi-scale datasets cannot be properly partitioned using a unique bandwidth value because of their local variable data density.

A strategy overcoming this restriction is to adapt the scaling measure according to each k-neighbor distance related to the local sample density [166]. Due to this approach seeks just for preservation of the local consistency, it mostly fails when dealing with outliers or noisy data. Besides, automatic tuning of the k-neighbor size remains an open issue. Clustering aggregation can be improved based on probability accumulation where the co-association matrices built from K-means clustering are weighted by the average pairwise distance of each cluster [158]. Nevertheless, some imposed constraints regarding the probability distribution of the data are not always satisfied in practice. More elaborate approaches are also in use like the locally adaptive similarity measure using neighborhood density information and the incorporation of the assumption about consistency of the similarity. Even though both approaches aim to reveal hidden data structures by imposing the density or topological constraints, each formulation requires a real user prior knowledge about the influence of the free parameters. This aspect becomes crucial to achieve a suitable trade-off between local and global consistency preservation.

Here, a graph pruning approach, called Kernel Alignment based Graph Pruning (KAGP), is proposed within a spectral clustering framework to enhances both the local and global data consistencies for a given input similarity matrix. Our approach aims to reveal the salient complex structure of the input data by finding relevant pairwise relationships among samples. To weak all irrelevant relationships of the input similarity matrix, KAGP quantifies the loss of information during the pruning process in terms of a kernel alignment-based function [34, 95]. Moreover, we encode the sample similarities using a compactly supported kernel that allows obtaining a sparse data representation to support the graph partitioning problem. As a consequence, KAGP takes advantage of an initial guess of the relationships among points to identify all relevant connections. Testing that is carried out on synthetic and real-world datasets shows that the proposed methodology allows enhancing the graph representation of different state of the art approaches, improving the clustering performance

in most of the cases. Moreover, KAGP avoids the need for a comprehensive user knowledge regarding the influence of its free parameters.

3.1. Spectral clustering fundamentals

Let $\boldsymbol{X}\in\mathbb{R}^{N\times P}$ be an input data matrix holding N samples and P features, where each row $\{\boldsymbol{x}_n\in\mathbb{R}^P : n\in[1,N]\}$ represents a single data point. The goal of clustering is to divide the data into different clusters, where samples within the same cluster are similar to each other. To discover the main topological relationships among data points, spectral clustering-based approaches build from $\boldsymbol{X}$ a weighted graph representation $\mathscr{G}(\boldsymbol{X},\boldsymbol{K})$, where each sample point, $\boldsymbol{x}$, is a vertex or node and $\boldsymbol{K}\in\mathbb{R}^{N\times N}$ is a similarity (affinity) matrix encoding all associations between graph nodes. In turn, each element of the similarity matrix, $k_{nn'}\subseteq\boldsymbol{K}$, corresponding to the edge weight between $\boldsymbol{x}_n$ and $\boldsymbol{x}_{n'}$, is commonly defined as follows [43]:

$$k_{nn'}=\kappa\left(\boldsymbol{x}_n-\boldsymbol{x}_{n'}\right), \tag{3-1}$$

where $\kappa(\cdot)\in\mathbb{R}^+$ is a symmetric kernel [132]. Commonly, the kernel function is fixed as Gaussian. Among many others available kernels (like Laplacian or polynomial), the Gaussian function has the advantages of finding Hilbert spaces with universal approximating capability and its mathematical tractability [94]. Thus, the Gaussian kernel that is defined as:

$$\kappa_G\left(\boldsymbol{x}_n-\boldsymbol{x}_{n'};\sigma_X\right)=\exp\left(\frac{-\|\boldsymbol{x}_n-\boldsymbol{x}_{n'}\|_2^2}{2\sigma_X^2}\right), \tag{3-2}$$

is preferred since it aims to find an RKHS with universal approximating capability and with a single bandwidth parameter $\sigma_X\in\mathbb{R}^+$ ($\|\cdot\|_2$ stands for the 2-norm operator).

Hence, the clustering task now relies on the statement of the conventional graph cut problem, where the goal is to partition the set of vertices $\mathscr{V}\in\boldsymbol{X}$ into $C\in\mathbb{N}$ disjoint subsets $\mathscr{V}_c$, so that:

$$\mathscr{V}=\cup_{c=1}^{C}\mathscr{V}_c, \tag{3-3}$$

and

$$\mathscr{V}_{c'}\cap\mathscr{V}_c=\emptyset,\quad\forall\, c'\neq c. \tag{3-4}$$

Since graph-cut approaches require high computational burden, relaxation of the clustering optimization problem has been developed based on the spectral graph analysis [107]. So, spectral clustering-based methods decompose the input data $\boldsymbol{X}$ into C disjoint subsets by using both spectral information and orthogonal transformations of $\boldsymbol{K}$. Algorithm 1 describes the well-known solution of the cut problem (termed NCut) that is based on the Rayleigh-Ritz theory.

Algorithm II.1: Basic spectral clustering

Input: $\boldsymbol{X}\in\mathbb{R}^{N\times D}$, $\boldsymbol{K}\in\mathbb{R}^{N\times N}$, $C\in\mathbb{N}$.
Output: $\mathscr{V}=\{\mathscr{V}_c : c=1,\ldots,C\}$.

1 initialization;
2 Compute the diagonal degree matrix $\boldsymbol{W}\in\mathbb{R}^{N\times N}$ holding elements $w_{nn}=\sum_{j\in N} k_{nn'}$.
3 Estimate the normalized Laplacian matrix: $\boldsymbol{L}=\boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{K}\boldsymbol{W}^{-\frac{1}{2}}$.
4 Calculate the eigenvalues$\{\lambda_n\in\mathbb{R}^+\}$, and eigenvectors $\{\boldsymbol{u}_n\in\mathbb{R}^N : n\in[1,N]\}$, of $\boldsymbol{L}$ and stack the C eigenvectors corresponding to the first C largest eigenvalues into $\boldsymbol{X}_A\in\mathbb{R}^{N\times C}$.
5 Assuming each row of $\boldsymbol{X}_A$ as a point of dimension $\mathbb{R}^C$, cluster them into C clusters by using the K-means algorithm.
6 Assign the original point $\boldsymbol{x}_n$ to cluster c, if and only if, the n-th row of the matrix $\boldsymbol{X}_A$ is allocated to the cluster c.

3.2. Kernel alignment-based graph pruning (KAGP)

To get available data partitioning, performance of the spectral clustering-based methods primarily resides in the choice of the similarity measure which should be smooth with respect to the intrinsic structure of the samples [162]. In fact, the quality of the graph representation, $\mathscr{G}(,)$ directly depends on the estimated similarity matrix $\boldsymbol{K}$ [76]. Moreover, an adequate similarity measure for clustering should hold the following two kinds of assumptions of consistency [169]: *i)* nearby points in the input space should have high similarity (local consistency); *ii)* points belonging to the same cluster should reach high similarity (global consistency).

Aiming to enhance the local and global consistency of the similarity matrix, we propose to employ a kernel alignment-based function that is subject to sparse constraints. As a result, a graph pruning method is developed that takes advantage of the initial guess for the relationship of the input samples, making possible to extract complex data structures. Thus, based on the properties of the Gaussian kernel, the following compactly supported kernel can be constructed [50]:

$$\kappa_\phi\left(\boldsymbol{x}_n,\boldsymbol{x}_{n'};\sigma_X\right)=\phi\left(\boldsymbol{x}_n,\boldsymbol{x}_{n'}\right)\kappa_G\left(\boldsymbol{x}_n,\boldsymbol{x}_{n'};\sigma_X\right). \tag{3-5}$$

being $\boldsymbol{K}$ the similarity matrix encoding the graph structure, $\mathscr{G}(\boldsymbol{X},\boldsymbol{K})$, and $\phi:\mathbb{R}^P\times\mathbb{R}^P\to\mathbb{R}^+$, is a compactly supported radial basis function. To preserve positive definiteness of κ_ϕ and to enhance the local and global data consistency in $\boldsymbol{K}$, operator $\phi(,)$ is chosen as a sparseness function [53]:

$$\phi\left(\boldsymbol{x}_n,\boldsymbol{x}_{n'};b,\nu\right)=\left(\max\{1-(\|\boldsymbol{x}_n-\boldsymbol{x}_{n'}\|_2)/b,0\}\right)^\nu, \tag{3-6}$$

being $b{\in}\mathbb{R}^{+}$. Notation $\max\{\cdot,0\}$ stands for the maximum value between the argument and zero. The power term that rules the degree of smoothness (i.e. differentiability) of $\phi(,)$ is adjusted as $\nu{\geq}(P+1)/2$. Therefore, the sparseness function introduces a hard threshold in eq. (3-6), making all entries having distance $\|\boldsymbol{x}_n-\boldsymbol{x}_{n'}\|_2>b$ to be zero.

Based on the operator ϕ established in eq. (3-6), a compactly supported kernel-based matrix $\boldsymbol{K}^{b,\nu}{\in}\mathbb{R}^{N\times N}$ can be computed as:

$$\boldsymbol{K}^{b,\nu}=\boldsymbol{S}^{b,\nu}\circ\boldsymbol{K}, \tag{3-7}$$

where $\boldsymbol{S}^{b,\nu}{\in}\mathbb{R}^{N\times N}$ is a sparse matrix holding elements:

$$s_{nn'}^{b,\nu}{=}\phi\left(\boldsymbol{x}_n,\boldsymbol{x}_{n'};b,\nu\right), \tag{3-8}$$

and $\circ$ stands for the Hadamard product. Nonetheless, the values of the b and ν parameters must be properly adjusted to reveal the main structures of $\boldsymbol{X}$ for facilitating further spectral clustering analysis. Yet, the latter parameter has a negligible effect in comparison with the former one when building the compactly supported kernel as discussed in [167].

Hence, to achieve a suitable local and global data structure representation, we just tune the value b by searching the sparse kernel encoding the most relevant node connections. To this end, we weaken all irrelevant relationships of the input similarity matrix $\boldsymbol{K}$, but taking into account the loss of information during the sparsification process in terms of a kernel alignment-based cost function [34]. Namely, the reciprocal relationship between $\boldsymbol{K}^{b,\nu}$ and $\boldsymbol{K}$ can be estimated using the correlation index, $\rho(b,\nu){\in}\mathbb{R}[0,1]$ as:

$$\rho(b,\nu)=\frac{\left\langle\bar{\boldsymbol{K}},\bar{\boldsymbol{K}}^{b,\nu}\right\rangle_F}{\sqrt{\left\langle\bar{\boldsymbol{K}},\bar{\boldsymbol{K}}\right\rangle_F\left\langle\bar{\boldsymbol{K}}^{b,\nu},\bar{\boldsymbol{K}}^{b,\nu}\right\rangle_F}}, \tag{3-9}$$

where matrices $\bar{\boldsymbol{K}}{=}\boldsymbol{H}\boldsymbol{K}\boldsymbol{H}$ and $\bar{\boldsymbol{K}}^{b,\nu}{=}\boldsymbol{H}\boldsymbol{K}^{b,\nu}\boldsymbol{H}$ are the centralized kernel versions of $\boldsymbol{K}$ and $\boldsymbol{K}^{b,\nu}$, respectively, and the centralization matrix $\boldsymbol{H}{\in}\mathbb{R}^{N\times N}$ is defined as $\boldsymbol{H}{=}\boldsymbol{I}-N^{-1}\boldsymbol{1}\boldsymbol{1}^{\top}$, where $\boldsymbol{1}{\in}\mathbb{R}^{N}$ is the all-ones vector and $\boldsymbol{I}{\in}\mathbb{R}^{N\times N}$ is the identity matrix. Notation $\langle,\rangle_F$ stands for the Frobenius-based matrix inner product.

It is worth noting that centered alignment-based functions have been demonstrated to correlate better than the uncentered case [32]. As a result, the higher the $\rho(b)$ value the lower information loss during the sparsifying process. Additionally, an sparsity index that defines the degree of matrix sparseness is quantified as:

$$\varrho(b,\nu)=N_0/N^2, \tag{3-10}$$

with $\varrho(b,\nu)\in\mathbb{R}[0,1]$ and being N_0 the number of zero entries of the matrix $\boldsymbol{K}^{b,\nu}$. Here, the higher the $\varrho(b,\nu)$ value – the higher degree of sparseness. In order to fix the optimal value of b, for a given fixed ν value, we introduce a regularization-based criterion as to reach a trade-off between both $\rho(b,\nu)$ and $\varrho(b,\nu)$ measures as follows:

$$b^* = \underset{b}{\operatorname{argmax}}\,(1-\gamma)(\log(\rho(b,\nu)))^2 + \gamma(\log(\varrho(b,\nu)))^2 \qquad \text{(3-11)}$$
$$\text{s.t. } \min_{n,n'} \|\boldsymbol{x}_n - \boldsymbol{x}_{n'}\|_2 < b < \max_{n,n'} \|\boldsymbol{x}_n - \boldsymbol{x}_{n'}\|_2$$

where $\gamma\in\mathbb{R}[0,1]$ rules the compromise between the local and global consistency terms. As $\gamma\to 0$, the optimization function in eq. (3-11) heavily penalizes the sparsifying process, that is, the obtained matrix $\boldsymbol{K}^{b^*,\nu}$ will try to preserve, as well as possible, all data similarities hold by $\boldsymbol{K}$. Therefore, matrix $\boldsymbol{K}^{b,\nu}$ will hold mainly global consistencies. On the contrary, as $\gamma\to 1$, the obtained compactly supported matrix will favor those sparse representations preserving local data consistency. It is worth noting that the imposed constraint in eq. (3-11) is derived from the definition of the value b, according to eq. (3-6). Thus, if b becomes lower that the minimum value of all computed input sample distances, the sparsifying function will always be zero, making useless the derived matrix. In contrast, if b becomes higher, the sparsifying function will not affect $\boldsymbol{K}$. Therefore, the provided optimization in eq. (3-11) takes advantage of the relevant input data information. Thus, the cost function allows finding the b^* value that is adequate to extract the main data structures to be encoded in $\boldsymbol{K}^{b,\nu}$. This kernel matrix is used to build a suitable input data graph representation $\mathscr{G}(\boldsymbol{X},\boldsymbol{K}^{b^*,\nu})$. Then, our approach is named kernel alignment-based graph pruning (KAGP).

3.3. Experimental set-up

Evaluation of the proposed KAGP approach is carried out by performing an unsupervised clustering task demanding estimation of the graph structure from the underlying data. For the sake of comparison, KAGP performance is applied on four different spectral clustering approaches requiring computation of the initial graph representation, namely: Adjusted Line Segment (ALS) [162], k-Nearest Neighbor Spectral Clustering (k-SC) [166], ϵ-Spectral Clustering (ϵ-SC) [43], and Common Nearest Neighbors (CNN) [169]. Parameter setting of these algorithms is as follows:

- The ALS algorithm incorporates a prior assumption about consistency of the similarity between samples, which means that nearby points and data points on the same structure are likely to share high similarities [162]. The ALS similarity measure is defined as:

$$k_{nn'} = (z_{nn'}+1)^{-1}, \qquad \text{(3-12)}$$

where $z_{nn'}\mathbb{R}^+$ is an introduced pair-wise density sensitive distance defined as:

$$z_{nn'} = \min \sum_{r=n}^{|P_{nn'}|} \iota(\boldsymbol{x}_r, \boldsymbol{x}_{r+1}). \tag{3-13}$$

Here, $P_{nn'}{=}\{\boldsymbol{x}_n, \boldsymbol{x}_r, \ldots, \boldsymbol{x}_{n'} : n \leq r < n'\}$ is the path and denotes the set of points connecting from $\boldsymbol{x}_n$ to $\boldsymbol{x}_{n'}$. In turn, the parameter $\iota(,){\in}\mathbb{R}^+$ is an adjustable line segment length between points $\boldsymbol{x}_n$ and $\boldsymbol{x}_{n'}$ computed as follows:

$$\iota(\boldsymbol{x}_n, \boldsymbol{x}_{n'}) = \left(\exp\left(\zeta \|\boldsymbol{x}_n - \boldsymbol{x}_{n'}\|_2\right) - 1\right)^{1/\zeta}, \tag{3-14}$$

where $\zeta > 1$ is a density factor parameter that squeezes distances within high-density regions while it widens them in low-density regions. The ζ value is heuristically set as 2, providing that the initial graph is built as an ϵ-neighborhood graph fixing $\epsilon{=}\xi$.

- In the k-SC algorithm, the similarity matrix is estimated as:

$$k_{nn'} = \kappa_G\left(\boldsymbol{x}_n, \boldsymbol{x}_{n'}; \sqrt{\sigma_n \sigma_{n'}}\right), \tag{3-15}$$

where the local scaling parameters $\{\sigma_n, \sigma_{n'}\}{\in}\mathbb{R}^+$ are computed in terms of the Euclidean distance as $\sigma_n{=}\|\boldsymbol{x}_n - \boldsymbol{x}_n^K\|_2$, being $\boldsymbol{x}_n^K$ the K-th neighbor of each point $\boldsymbol{x}_n$, so that each specific scaling parameter allows self-tuning of the point-to-point distances according to the local statistics of the neighborhoods surrounding points $\boldsymbol{x}_n$ and $\boldsymbol{x}_{n'}$. To avoid overfitting, the K value is adjusted (taking into account the dataset size) as $K{=}\left\lfloor\sqrt{N}\right\rceil$, where $\lfloor\cdot\rceil$ is the operator that computes the rounded value to the closest integer for its argument.

- Now, in the ϵ-SC approach the similarity is computed as:

$$k_{nn'} = \kappa_G\left(\boldsymbol{x}_n, \boldsymbol{x}_{n'}; \sigma_\epsilon\right), \tag{3-16}$$

where the $\sigma_\epsilon{\in}\mathbb{R}^+$ value encodes the average neighborhood size of the input data. As suggested in [3], we choose the median operator.

- The pairwise similarity of the CNN algorithm becomes adaptive in dependence to the neighborhoods of the correlative points, that is, if a couple of points are located within the same cluster, both points are assumed as belonging to a high density region. Therefore, the CNN local density adaptive similarity measure can be written as:

$$k_{nn'}{=}\kappa_G\left(\boldsymbol{x}_n, \boldsymbol{x}_{n'}; (\varsigma_{nn'} + 1)^{1/2}\sigma_\varsigma\right), \tag{3-17}$$

where $\sigma_\varsigma{=}\sigma_\epsilon$, $\varsigma_{nn'}{\in}\mathbb{N}$ is a local density parameter, which is employed to distinguish points within the same cluster from others located at different clusters as:

$$\varsigma_{nn'} = \left|\{r : \|\boldsymbol{x}_n - \boldsymbol{x}_r\| < \xi \text{ and } \|\boldsymbol{x}_{n'} - \boldsymbol{x}_r\| < \xi\}\right|,$$

being $\xi{\in}\mathbb{R}^+$ a neighborhood radius parameter that is adjusted experimentally as the 10-percentile of the input data Euclidean distances. Notation $|\cdot|$ denotes the cardinality operator.

Once the similarity matrix is computed for each method, the proposed KAGP is carried out to prune all irrelevant pair-wise relationship values, where the optimal value of the parameter b is computed by solving the cost function in eq. (3-11) using a Particle Swarm Optimization-based solver. Besides, the regularization parameter γ is heuristically set as 0.5. Lastly, the sparse matrix $\boldsymbol{K}^{b^*,\nu}$ is estimated to perform further the well-known spectral clustering algorithm [157]. Fig. **4-2** outlines the main sketch of the used spectral clustering task used for validation of the proposed KAGP approach.

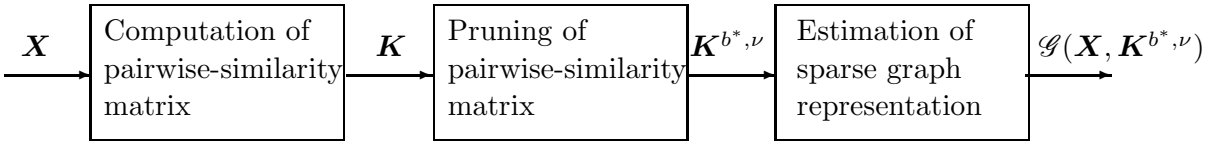


Figure 3-1.: KAGP block scheme

The KAGP method is validated as a suitable tool for pruning graph representations to improve unsupervised clustering performance. To this end, we employ both, synthetic and real-world, databases.

In the former case, we have visual insight of the performed clustering for the following well-known datasets: *Bull's eye 3 circles*, *Happy face* [166], *Half moon* [72], and *Bull's eye with outliers*. Two main reasons account for selecting these concrete databases: *i*) they represent a challenging clustering task due to their complex structures, and *ii*) each ground-truth is either known or can be visually inspected.

Regarding the real-world data, a subset of the UCI Machine Learning Repository that is widely used for quantifying clustering performance [162]. The selected data collection provides a variety of real-world clustering scenarios with different conditions in terms of number of features, number of samples, number of clusters, and data distribution complexity. Besides, validation is carried out on a representative subset of 30 images chosen randomly from the free access Berkeley Segmentation dataset [97]. This data collection provisions hand-labeled segmentation of every sample, making supervised testing suitable. All testing images are rescaled at 15% and characterized by five features: RGB color space and the spatial position of each pixel. Thus, each image is represented by the input matrix $\boldsymbol{X}{\in}\mathbb{R}^{3577\times 5}$, where $N{=}73\times 49$ is the obtained image resolution after resizing.

Further, we assess the clustering performance in terms of the following commonly used indexes of quality [8, 27]:

Adjusted Rand Index (ARI), $\rho_{ARI}\in\mathbb{R}[-1,1]$: This index measures the agreement between two compared partitions, namely, the ground truth (noted as $\mathscr{U}$) and the estimated by the tested clustering approach ($\mathscr{V}$), as follows:

$$\rho_{ARI}=\frac{a_{11}-(a_{11}+a_{01})(a_{11}+a_{10})/a_{00}}{(a_{11}+a_{01})+(a_{11}+a_{10})/2-(a_{11}+a_{01})(a_{11}+a_{10})/a_{00}} \tag{3-18}$$

where $a_{11}\in\mathbb{N}$ is the number of sample pairs belonging to the same subset in $\mathscr{U}$ and in $\mathscr{V}$, a_{10} is the number of sample pairs belonging to the same subset in $\mathscr{U}$ and to different subsets in $\mathscr{V}$, $a_{01}\in\mathbb{N}$ is the number of sample pairs belonging to different subset in $\mathscr{U}$ and to the same one in $\mathscr{V}$, and $a_{00}\in\mathbb{N}$ is the number of sample pairs belonging to different subsets in $\mathscr{U}$ and in $\mathscr{V}$.

Purity Index (PUR) $\rho_{PUR}\in\mathbb{R}[0,1]$: This measure matches the clustering partition $\mathscr{V}$ with the ground truth $\mathscr{U}$ as a weighted sum of the maximal precision values for each subset. That is:

$$\rho_{PUR}=\sum_{c=1}^{C}\frac{|\mathscr{V}_c|}{N}\max_{c'}\rho_{PRE}\left(\mathscr{V}_c,\mathscr{U}_{c'}\right); \tag{3-19}$$

where $\rho_{PRE}\left(\mathscr{V}_c,\mathscr{U}_j\right)=|\mathscr{V}_c\cap\mathscr{U}_{c'}|/|\mathscr{V}_c|$.

Accuracy Index (ACC) $\rho_{ACC}\in\mathbb{R}[0,1]$: This index is the total fraction of samples belonging to the same subset in $\mathscr{U}$ and $\mathscr{V}$, and is expressed as:

$$\rho_{ACC}=\frac{1}{|\mathscr{V}|}\sum_{c=1}^{C}|\mathscr{V}_c\cap\mathscr{U}_c| \tag{3-20}$$

where $|\cdot|$ is the cardinality operator over a given set.

Jaccard Index (JAC), $\rho_{JAC}\in\mathbb{R}[0,1]$: This index matches the similarity among two sets, $\mathscr{U}$ and $\mathscr{V}$, as follows:

$$\rho_{JAC}=\frac{a_{11}}{a_{11}+a_{10}+a_{01}} \tag{3-21}$$

Probabilistic Rand Index (NPR), $\rho_{NPR}\in\mathbb{R}[0,1]$: This measure allows comparing the performed partition under testing $\mathscr{V}$ against T available ground truths, $\Psi=\{\mathscr{U}^t : t\in[1,T]\}$, through an introduced soft nonuniform weighting of all sample pairs as function of the ground truth variability. We assume that each $\mathscr{U}^t$ is the t-th partition of $\boldsymbol{X}$ according to the t-th expert, so that $\mathscr{U}^t=\cup_{c=1}^{C}\mathscr{U}_c^t$, being $\mathscr{U}_c^t$ a disjoint subset in $\mathscr{U}^t$ and $\mathscr{U}_{c'}^t\cap\mathscr{U}_c^t=\emptyset$, $\forall\, c'\neq c$. Therefore, the NPR, is defined as [153, 102]:

$$\rho_{NPR} = \frac{1}{T\binom{N}{2}} \sum_{t=1}^{T} \sum_{n,n'=1}^{N} \Big[\delta\left(l_n^{\mathscr{V}} = l_{n'}^{\mathscr{V}}\right) \delta\left(l_n^{\mathscr{U}^t} = l_{n'}^{\mathscr{U}^t}\right) + \delta\left(l_n^{\mathscr{V}} \neq l_{n'}^{\mathscr{V}}\right) \delta\left(l_n^{\mathscr{U}_t} \neq l_{n'}^{\mathscr{U}_t}\right) \Big] \tag{3-22}$$

where $l_n^{\mathscr{V}} = \{c : \boldsymbol{x}_n \in \mathscr{V}_c\}$, $l_n^{\mathscr{U}^t} = \{c : \boldsymbol{x}_n \in \mathscr{U}_c^t\}$, and notation $\delta(\cdot,\cdot)$ stands for the delta function.

To evaluate the proposed pruning method, we firstly carry out the visual inspection of the performed data grouping for all tested databases. However, the indexes of quality above-explained are used depending on the available information. Thus, the ρ_{NPR} index is the only one estimated for the Berkeley dataset due to the hand-labeled segmentation is for every testing image. Otherwise, we calculate the ρ_{ACC}, ρ_{PUR}, and ρ_{ARI} indexes in the remaining databases.

3.4. Results and discussion

fig. **3-2** shows the accomplished grouping results on the synthetic datasets before and after applying the proposed KAGP method to each compared spectral clustering technique. The first method we compare is the ALS that aims to reveal complex data structures by encoding topological and density properties of neighboring samples. And yet, the use of a unique density parameter does not allow squeezing properly the distances among points. Moreover, the plain ALS (i.e. without the proposed graph pruning) gets the worst connectivity graphs on every single dataset (see fig. 3.3a, fig. 3.3i, fig. 3.3q, and fig. 3.3y), giving rise to poor splitting data. Still, the use of the proposed KAGP method remarkably improves clustering performance on all databases (fig. 3.2e, fig. 3.2m, and fig. 3.2ac) except for the *Half moon* dataset that is the one having the simpler structure (fig. 3.2u).

The following clustering method, k-SC technique, handles correctly on *Bull's eye 3 circles* and *Happy face* data collections as seen in fig. 3.2b and fig. 3.2j, respectively. However, the clusters are wrongly split in the *Half moon* (fig. 3.2r) and the *Bull's eye with outliers* (fig. 3.2z) datasets. This drawback is explained since the k-SC method mainly seeks just for local consistency preservation, making some global connections supply wrong clusters mostly at handling complex circumstances, e.g., a multi-scale data set [166]. In fact, the estimated connectivity graphs show that the plain k-SC graphs incorrectly assign some pairwise connections, namely, when both moons get close enough to each other (see fig. 3.3r) or when dealing with outliers (fig. 3.3z) that are mistakenly assigned to the structure of the inner rings. Conversely, the proposed KAGP algorithm performs better clustering since it produces graphs (fig. 3.3v and fig. 3.3ad) that evidently prune irrelevant connections from the k-SC similarity matrix, but jointly preserving both the local and global consistencies.

The improved version, ϵ-SC, is assumed to better describe local relationships through an introduced ϵ value that rules the size of neighboring vicinities within the connectivity graph estimation framework. Nonetheless, this strategy may be not enough to tackle the problem inasmuch as the ruling ϵ parameter barely handles multi-scale, noisy or complex structures, yielding connection graphs (see fig. 3.3c-to-fig. 3.3ae) that are quite similar to the ones estimated by the k-SC method. As a result, the ϵ-SC reaches the same clustering performance as the k-SC does, that is, the method does not benefit from the KAGP algorithm in the cases of *Bull's eye 3 circles* (see fig. 3.2c and fig. 3.2g) and *Happy face* (fig. 3.2k and fig. 3.2o) while the pruning method improves performance on the *Half moon* (fig. 3.2s and fig. 3.2w) and the *Bull's eye with outliers* (fig. 3.2aa and fig. 3.2ae) datasets.

Lastly, the compared CNN method provides a more elaborate strategy to improve characterization of local neighborhoods by adapting the pairwise edge weight between points, allowing to handle more information about multi-scale data. Detailed inspection of the CNN graphs shows that the use of the KAGP method removes more accurately irrelevant connections (see fig. 3.3l vs fig. 3.3p and fig. 3.3t vs fig. 3.3x). Thus, the CNN method reaches almost the same performance in either case of preprocessing consideration as seen in fig. 3.2d-to-fig. 3.2af.

table **3-1** displays the clustering quality assessments estimated on the examined synthetic datasets. As perceived, the proposed KAGP approach allows improving all considered indices of clustering quality. In particular, KAGP-based graph pruning remarkably enhances the clustering quality of the ALS, k-SC and ϵ-SC techniques in all tested synthetic data. Yet, the KAGP does not exhibit a notable quality enhancement of the CNN algorithm, though it does not decrease the system performance.

Table **3-1**.: Clustering quality assessment results (synthetic datasets).

Dataset	Method	ALS		k-SC		ϵ-SC		CNN	
	Measure	No KAGP	KAGP	No KAGP	KAGP	No KAGP	KAGP	No KAGP	KAGP
be3	ARI	0.31	1.00	1.00	1.00	0.51	1.00	0.51	0.51
$N = 299$	Purity	0.61	1.00	1.00	1.00	0.84	1.00	0.84	0.84
$D = 2$	Accuracy	0.61	1.00	1.00	1.00	0.64	1.00	0.63	0.63
$C = 3$	Jaccard	0.39	1.00	1.00	1.00	0.56	1.00	0.56	0.56
happy	ARI	1.00	1.00	1.00	1.00	0.85	1.00	1.00	1.00
$N = 266$	Purity	1.00	1.00	1.00	1.00	0.95	1.00	1.00	1.00
$D = 2$	Accuracy	1.00	1.00	1.00	1.00	0.95	1.00	1.00	1.00
$C = 3$	Jaccard	1.00	1.00	1.00	1.00	0.83	1.00	1.00	1.00
hm2	ARI	-0.08	0.01	0.37	1.00	0.59	1.00	0.97	0.98
$N = 373$	Purity	0.80	0.66	0.81	1.00	0.89	1.00	0.99	0.99
$D = 2$	Accuracy	0.54	0.60	0.81	1.00	0.89	1.00	0.99	0.99
$C = 2$	Jaccard	0.44	0.41	0.56	1.00	0.70	1.00	0.97	0.98
tar	ARI	0.38	1.00	0.38	1.00	1.00	1.00	1.00	1.00
$N = 770$	Purity	0.40	1.00	0.40	1.00	1.00	1.00	1.00	1.00
$D = 2$	Accuracy	0.40	1.00	0.40	1.00	1.00	1.00	1.00	1.00
$C = 6$	Jaccard	0.38	1.00	0.38	1.00	1.00	1.00	1.00	1.00
Average	ARI	0.40	**0.75**	0.69	**1.00**	0.74	**1.00**	0.87	0.87
	Purity	0.70	**0.92**	0.80	**1.00**	0.92	**1.00**	0.96	0.96
	Accuracy	0.64	**0.90**	0.80	**1.00**	0.87	**1.00**	0.91	0.91
	Jaccard	0.55	**0.85**	0.74	**1.00**	0.77	**1.00**	0.88	**0.89**

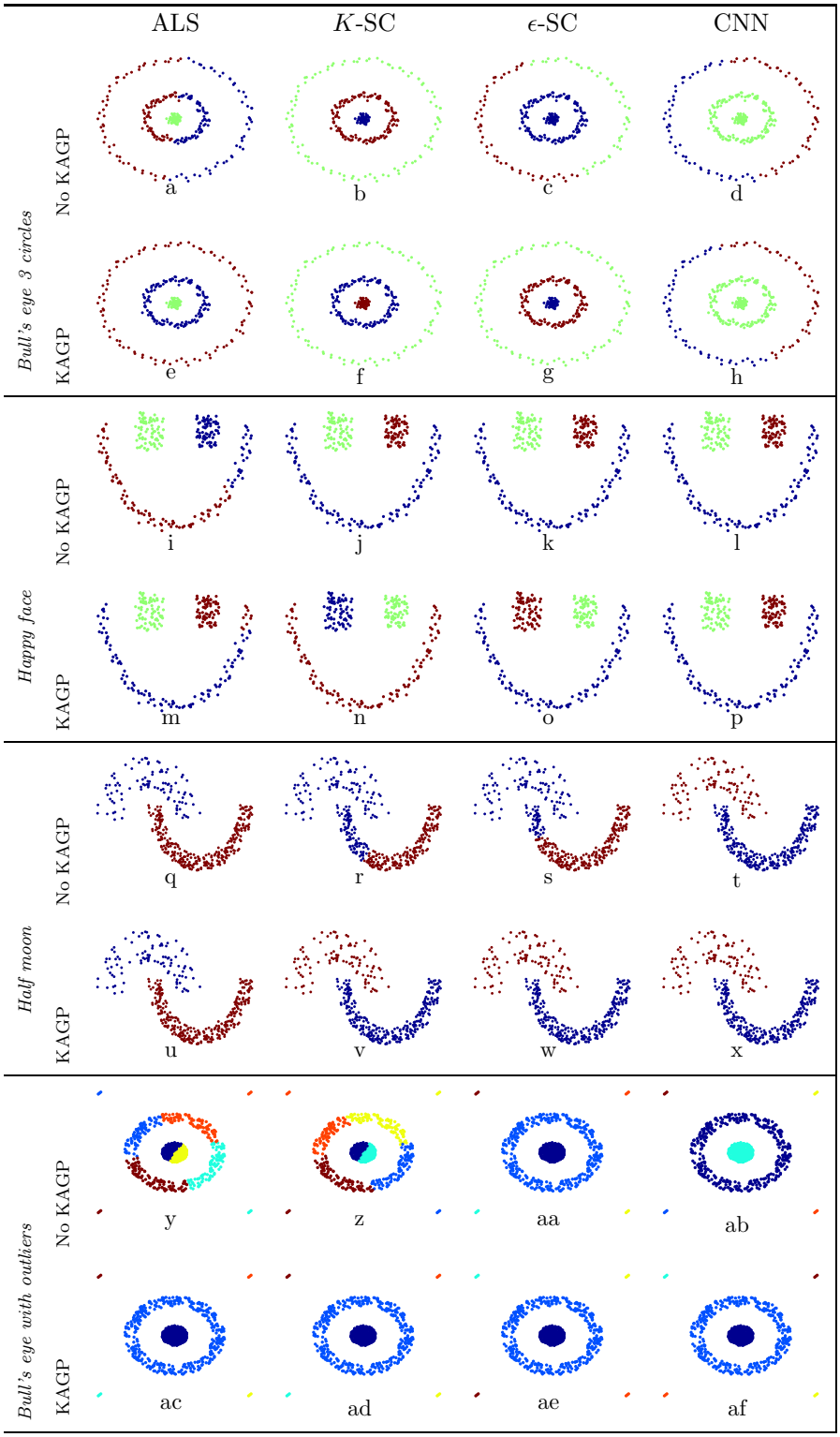


Figure 3-2.: Clustering results carried out on synthetic data sets.

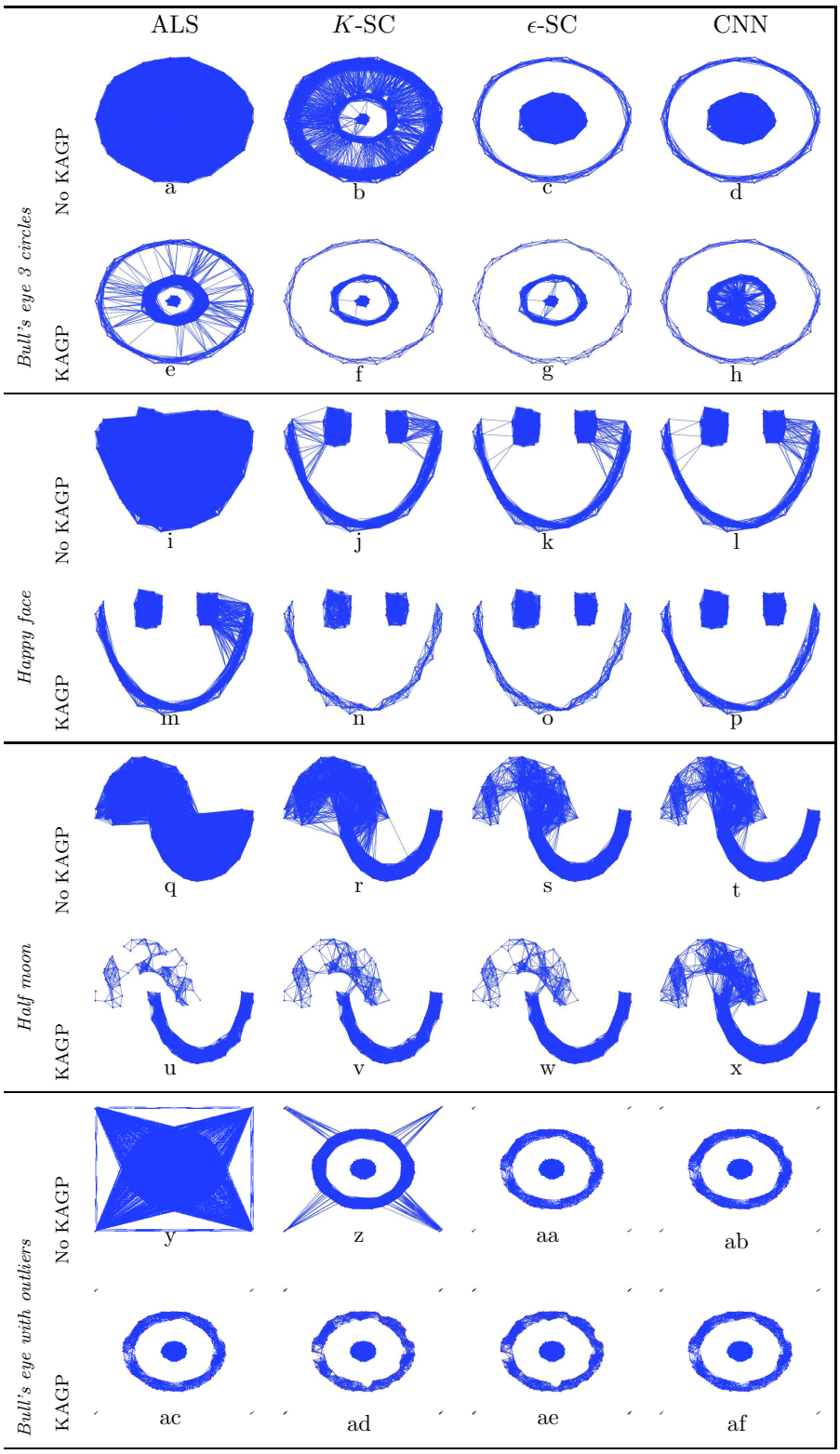


Figure 3-3.: Graph representation for synthetic data sets.

As seen intable **3-2** showing the quality assessments calculated for the UCI repository real-world datasets, the proposed KAGP enhances (at least, does not degrade) the performed clustering. In particular, KAGP remarkably enhances the purity index for the k-SC, the ϵ-SC, and the CNN algorithms. Thus, pruning of irrelevant connections using KAGP favors the clustering robustness against noisy and/or complex data distributions. Concerning the ARI, the Accuracy, and the Jaccard indexes, KAGP achieves comparable results in comparison to the benchmark approaches.

The image segmentation results are shown in fig. **3-4**, where it is clear how the proposed KAGP enhances the segmentation performance. Particularly, as seen in figs. 3.4b to 3.4al, it is possible to notice how after applying KAGP over the k-SC similarity matrix facilitates the discrimination among objects into the scene. Alike, KAGP is able to enhance the ϵ-SC-based graph representation for facilitating the image clustering, e.g., see figs. 3.4c to 3.4am. Regarding this, only considering local consistencies for computing the relationships among pixels, e.g., k-SC and ϵ-SC algorithms, leads to noisy segmentation results. Moreover, it is remarkably how the image segmentation results for CNN and ALS approaches are improved after carrying out the proposed KAGP (see figs. 3.4d to 3.4an, and figs. 3.4a to 3.4ak). Even though CNN and ALS approaches estimate pair-wise similarities by considering local neighborhood properties complex data structures related to object textures and shapes are not suitable highlighted. In terms of clustering assessment, fig. **3-5** shows the corresponding boxplots obtained for the tested images. As seen, the use of the KAPG makes more stable the estimated clustering performance since it makes the results be less variable with less outliers. Consequently, the KAGP is able to find a trade-off between local and global consistency preservation to build a graph representation without irrelevant pair-wise connections, favoring the clustering robustness against outliers, noisy data, and overlapped groups.

Sensitivity analysis of free parameters. Due to all compared spectral clustering approaches have free parameters to be specified manually, we evaluate their influence on the estimation of the initial pairwise-similarity matrix, $\boldsymbol{K}$. Thus, the following free parameters are studied: *i)* The k-th neighboring value for k-SC, *ii)* the average neighborhood size value σ_ϵ for the ϵ-SC approach, *iii)* the neighborhood radius parameter ξ for the CNN approach, and *iv)* the density factor parameter ζ for ALS. Thereby, the synthetic and the Berkeley image segmentation datasets are studied. Thus, the ARI and the NPR indexes are analyzed, respectively. fig. **3-6** shows the results of evaluation that is carried out in terms of the ARI measure before and after the KAGP operation. As seen in figs. 3.6a to 3.6c for ALS method, the lack of the proposed pruning makes worse the achieved validation ARI measure regardless the fixed value of the ruling ζ parameter, except, again, for the *Half moon*. In the cases of associated k-SC and ϵ-SC methods (see figs. 3.6e to 3.6g figs. 3.6i to 3.6k, figs. 3.7a to 3.7c), their plain versions perform clustering that rapidly worsens as the corresponding free parameter slightly varies. In contrast, the use of the KAPG-based pruning makes both techniques improve the validation outcomes over all tested data sets.

With respect to the CNN method, its plain version is the one that holds clustering perfor-

mance within the longest interval of the ξ parameter variation in comparison with the other plain versions. Yet, its provided clustering sharply decreases, at the moment of parameter imbalance, and becomes the worst as seen in figs. 3.6m to 3.6o and figs. 3.7d to 3.7f. By contrast, the insertion of the pruning operation significantly compensates for this negative effect, extending the range within the free parameter can change.

3.5. Summary

We propose a graph pruning approach, termed KAGP, that makes use of a kernel function to support grouping tasks based on spectral clustering. Here, the kernel matrix learning is based on an introduced alignment function to measure the similarity between two kernel matrices, enhancing their local and global consistencies. So, our approach takes advantage of an initial guess of the relationships between points to identify relevant connections by encoding then by means of a compactly supported kernel function. Besides, a regularization-based criterion is introduced as to reach a trade-off between the local and the global consistency preservation during the graph pruning process. For the sake of comparison, KAGP is validated on synthetic and real-world datasets using visual inspection and clustering quality measures. Performance is contrasted with four competitive graph-based spectral clustering approaches, namely: k-SC, ϵ-SC, CNN, and ALS. So, once the initial graph representation is computed for each method, the proposed KAGP is carried out before clustering the data to prune irrelevant pair-wise relationships. Obtained results of quality show that KAGP can handle complex data structures, yielding better clustering performance in comparison to the baselines. Moreover, the KAGP promotes clustering performance less sensitive to outliers, noisy data, and overlapped groups.

Due to all compared spectral clustering approaches have free parameters that commonly are to be specified manually, we also evaluate experimentally their influence during the graph pruning process. Attained results demonstrate that the insertion of KAGP significantly compensates adverse effects when the corresponding free parameter is not fixed correctly for each clustering approach. Moreover, in most of the cases, KAGP allows extending the range within the free parameter can change. Therefore, proposed approach is a suitable alternative to support clustering tasks related to graph representations, achieving appropriate performances while avoiding the need for a comprehensive user knowledge regarding the influence of its free parameters. As future work, the authors plan to validate KAGP on other different machine learning tasks as dimensionality reduction, classification, and regression. Furthermore, an extension of KAGP to support kernel-based clustering approaches will be studied. Finally, it would be of benefit to include alternatives to measure the local and global consistency preservation, i.e., information theory.

Table **3-2**.: Clustering quality assessment results (UCI repository datasets).

Dataset	Method	ALS		k-SC		ϵ-SC		CNN	
	Measure	No KAGP	KAGP	No KAGP	KAGP	No KAGP	KAGP	No KAGP	KAGP
iris	ARI	0.47	0.47	0.44	0.55	0.63	0.55	0.63	0.56
$N = 150$	Purity	0.91	0.91	0.87	0.97	0.84	0.97	0.84	0.98
$D = 4$	Accuracy	0.58	0.57	0.53	0.68	0.84	0.68	0.84	0.69
$C = 3$	Jaccard	0.52	0.51	0.50	0.58	0.60	0.58	0.60	0.59
wine	ARI	0.03	0.10	0.93	0.43	0.93	0.43	0.90	0.43
$N = 178$	Purity	0.57	0.54	0.98	0.94	0.98	0.94	0.97	0.94
$D = 13$	Accuracy	0.44	0.46	0.98	0.62	0.98	0.62	0.97	0.62
$C = 3$	Jaccard	0.24	0.27	0.91	0.50	0.91	0.50	0.87	0.50
sonar	ARI	-0.00	0.03	0.02	0.00	0.00	0.01	-0.00	0.01
$N = 208$	Purity	0.97	0.92	0.57	0.86	0.55	0.84	0.57	0.84
$D = 60$	Accuracy	0.50	0.60	0.57	0.55	0.54	0.56	0.52	0.56
$C = 2$	Jaccard	0.48	0.47	0.34	0.43	0.34	0.42	0.34	0.43
biomed	ARI	0.04	-0.00	0.47	0.11	0.09	0.10	0.09	0.09
$N = 194$	Purity	0.60	0.53	0.85	0.93	0.95	0.94	0.95	0.94
$D = 5$	Accuracy	0.60	0.51	0.85	0.71	0.70	0.71	0.70	0.70
$C = 2$	Jaccard	0.37	0.35	0.60	0.55	0.55	0.55	0.55	0.55
diabetes	ARI	-0.00	-0.00	0.16	0.01	0.16	0.00	0.01	-0.00
$N = 768$	Purity	0.99	0.99	0.70	0.99	0.70	0.99	0.95	0.99
$D = 8$	Accuracy	0.64	0.64	0.70	0.65	0.70	0.65	0.65	0.65
$C = 2$	Jaccard	0.54	0.54	0.43	0.54	0.43	0.54	0.52	0.54
glass	ARI	0.08	0.02	0.17	0.16	0.16	0.16	0.16	0.16
$N = 214$	Purity	0.42	0.41	0.51	0.89	0.87	0.89	0.87	0.89
$D = 9$	Accuracy	0.42	0.38	0.46	0.49	0.50	0.50	0.50	0.50
$C = 4$	Jaccard	0.20	0.18	0.25	0.34	0.33	0.34	0.34	0.34
x80	ARI	0.00	0.00	0.63	0.01	0.36	0.01	0.31	0.01
$N = 45$	Purity	0.96	0.96	0.87	0.87	0.76	0.89	0.71	0.89
$D = 8$	Accuracy	0.36	0.36	0.87	0.42	0.58	0.40	0.58	0.40
$C = 3$	Jaccard	0.31	0.31	0.60	0.29	0.42	0.30	0.38	0.30
ecoli	ARI	0.40	0.69	0.37	0.57	0.48	0.72	0.57	0.71
$N = 336$	Purity	0.63	0.75	0.56	0.72	0.66	0.80	0.71	0.77
$D = 7$	Accuracy	0.54	0.75	0.55	0.72	0.66	0.79	0.71	0.77
$C = 8$	Jaccard	0.36	0.62	0.32	0.51	0.42	0.66	0.50	0.65
heart	ARI	-0.00	0.00	0.31	0.02	0.37	0.41	0.39	0.02
$N = 297$	Purity	1.00	1.00	0.78	0.93	0.80	0.82	0.81	0.95
$D = 13$	Accuracy	0.54	0.54	0.78	0.58	0.80	0.82	0.81	0.59
$C = 2$	Jaccard	0.50	0.50	0.49	0.47	0.52	0.55	0.54	0.48
liver	ARI	0.03	0.02	-0.00	-0.01	-0.01	-0.01	-0.01	-0.01
$N = 345$	Purity	0.59	0.59	0.65	0.94	0.98	0.95	0.97	0.97
$D = 6$	Accuracy	0.59	0.58	0.52	0.56	0.57	0.56	0.57	0.56
$C = 2$	Jaccard	0.35	0.35	0.36	0.48	0.50	0.48	0.50	0.49
ionosphere	ARI	-0.00	-0.00	0.15	0.35	0.13	0.28	0.12	0.27
$N = 351$	Purity	1.00	1.00	0.70	0.83	0.68	0.86	0.68	0.87
$D = 34$	Accuracy	0.64	0.64	0.70	0.81	0.68	0.78	0.68	0.77
$C = 2$	Jaccard	0.54	0.54	0.44	0.60	0.41	0.58	0.41	0.58
soybean2	ARI	0.28	0.27	0.56	0.28	0.29	0.26	0.30	0.28
$N = 136$	Purity	0.77	0.75	0.82	0.76	0.78	0.76	0.79	0.77
$D = 35$	Accuracy	0.58	0.57	0.82	0.57	0.59	0.57	0.60	0.58
$C = 4$	Jaccard	0.35	0.34	0.52	0.35	0.36	0.34	0.36	0.35
Average	ARI	0.11	**0.13**	**0.35**	0.21	**0.30**	0.24	**0.29**	0.21
	Purity	0.78	0.78	0.74	**0.89**	0.80	**0.89**	0.82	**0.90**
	Accuracy	0.54	**0.55**	**0.69**	0.61	**0.68**	0.64	**0.68**	0.61
	Jaccard	0.40	**0.42**	**0.48**	0.47	0.48	**0.49**	**0.49**	0.48

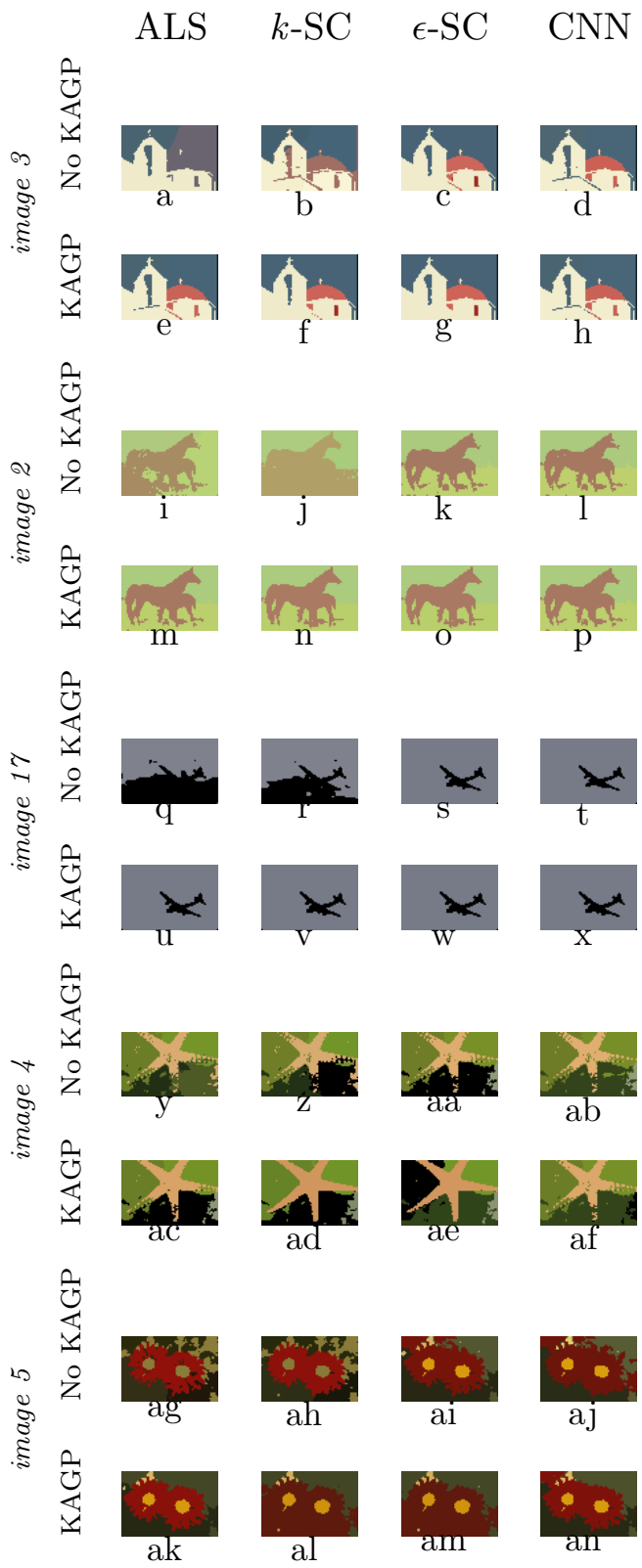


Figure 3-4.: Clustering results for the Berkeley Segmentation dataset.

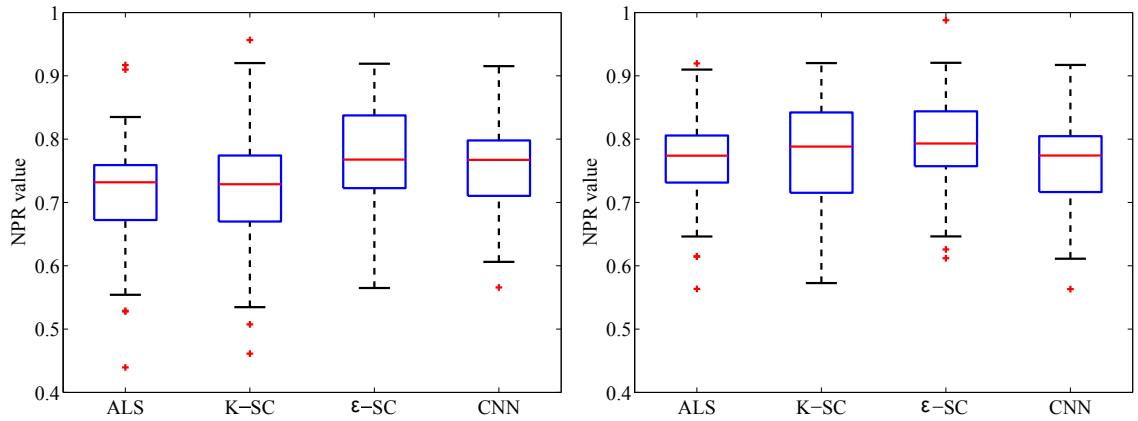


Figure 3-5.: Berkeley Segmentation dataset results (statistical analysis of the NPR).

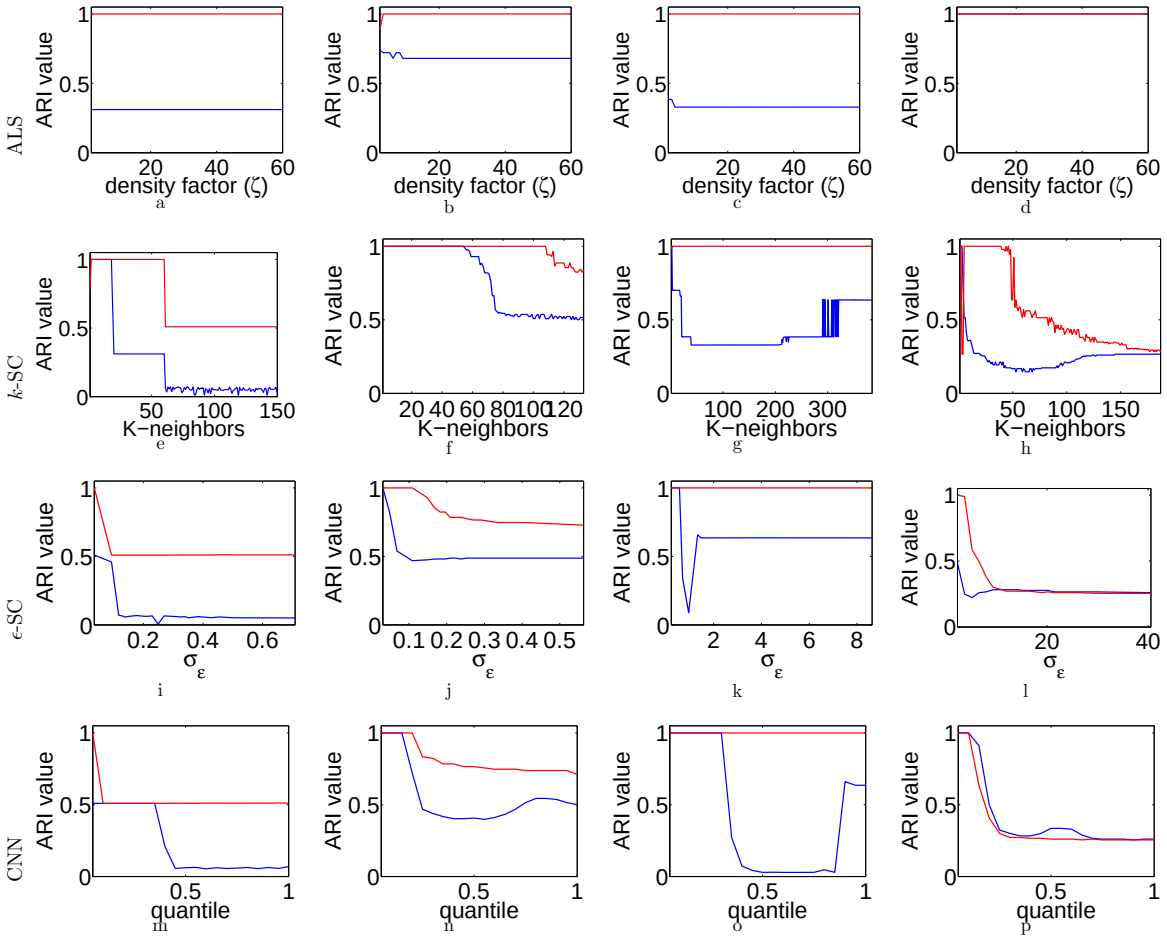


Figure 3-6.: Free parameter analysis over synthetic datsets based on the ARI measure. — No KAGP — after applying KAGP.

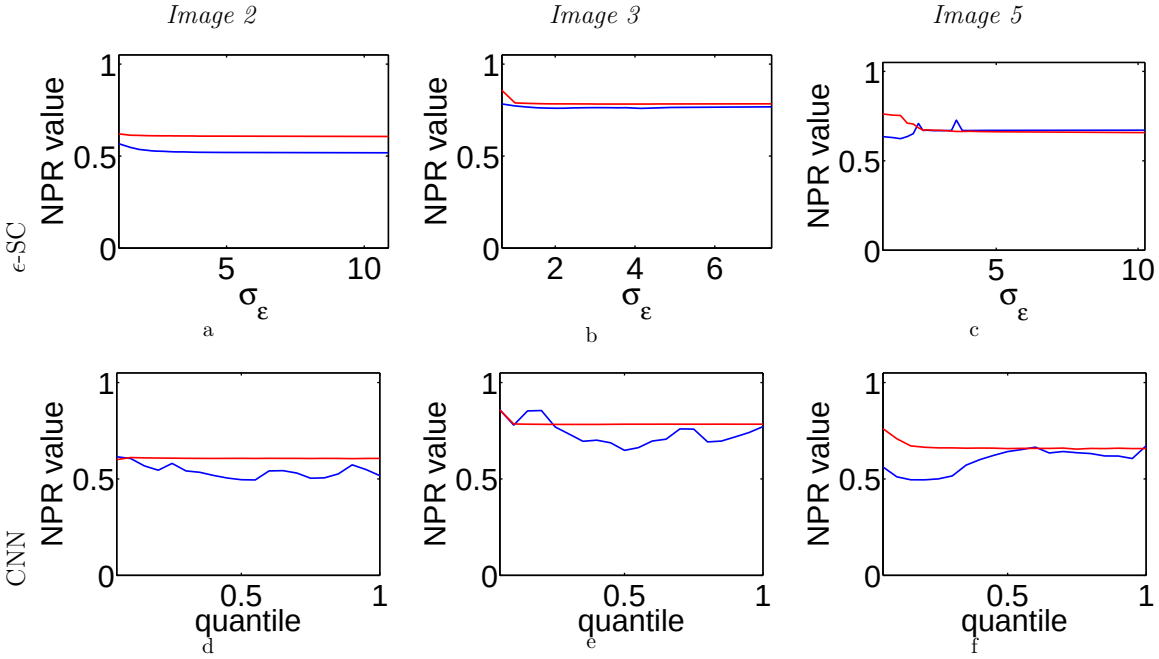


Figure 3-7.: ϵ-SC and CNN free parameter analysis over Berkeley image dataset based on the NPR measure. — No KAGP — after applying KAGP.

4. Relevant data representation based on information theoretic learning: a kernel function estimation approach

Kernel functions allow enhancing random data representation for supporting machine learning systems. Moreover, kernel-based methods are powerful tools for developing better performing solutions by adapting the kernel to a given problem, instead of learning data relationships from explicit raw vector representations. The kernel function is a very flexible container to express knowledge about the problem as well as to capture meaningful data relationships [13]. However, building suitable kernels requires some user prior knowledge about input data, which is not available in most of the practical cases; this situation becomes worse when handling unsupervised inferring tasks.

Among many feasible kernels, the Gaussian function is preferred since it aims to find an RKHS with universal approximating capability [94]. However, its use highly relies on the appropriate selection of the kernel parameters that are not easy to fix when dealing with complex data structures. In fact, the Gaussian Kernel bandwidth (scale) must be accurately tuned as to estimate an RKHS that should hold the main data relationships; otherwise, an unappropriate scale value leads to distinct RKHS not fulfilling the learning task. To cope with this issue and specifically devoted to unsupervised tasks, authors in [166, 169] propose to adjust the kernel parameter by making use of local scales instead of a global one allowing to exploit the spatial-varying data properties. Yet, these methods do not guarantee the Mercer's properties required for building kernel functions [118].

Nonetheless, most of of kernel estimation approaches are limited to the conventional concepts of second order statistics (mainly L2 distances). Instead, some information theoretic learning (ITL) frameworks have been developed based on information theoretic underpinnings, which more generally quantify data uncertainty. In fact, information-based approaches can improve interpretation of random data structures, making salient connections between information measures and RKHS [128]. In ITL methods, the kernel building task reduces to estimation of the probability density function (pdf) that is rarely known due to the only available information comes from data samples at hand. Here, the kernel estimator involves a symmetrical window sliding along a sequence with its weighted values being smoothed inside. In particular, author in [121] proposes to estimate the pdf using the Renyi's entropy along with a Gaussian kernel Parzen estimator. However, both the pdf estimation success

and the learning performance are highly dependent on the kernel parameter, namely, the bandwidth value. Some ITL-based approaches have been also proposed to fix the kernel scale value by optimizing information quantities [103, 138, 170], nevertheless, supervised data is required.

In this chapter, we propose a new kernel function estimation strategy to build a suitable Gaussian kernel-based RKHS oriented towards clustering. To this end, we make use of the intrinsic information potential variations from a Parzen-based pdf estimator. Namely, we seek for an RKHS maximizing the whole information potential variability in terms of the global kernel parameter. As a result, we get a scale updating rule as a function of the information forces, which are induced by a kernel function applied over a finite sample set. Thus, our approach allows revealing relevant sample relationships into an unsupervised kernel-based strategy. We provide testing of our proposal on two classical machine learning tasks (clustering and classification) using both synthetic and real data. Obtained results show that presented approach allows building an RKHS kernel favoring data groups separability and reaching suitable clustering performance in comparison with other state-of-the-art algorithms.

4.1. Gaussian-based Renyi's information metrics fundamentals

The basis of the ITL framework is the Renyi's information quadratic metric. A method to estimate the well-known Renyi's entropy directly from a sample set $X=\{x_n\in\mathscr{X} : \forall n\in[1,N]\}$, being $\mathscr{X}$ a given representation space, can be achieved by using the Parzen's nonparametric pdf estimation for x, defined as:

$$p(x) \approx p_X(x|\sigma_X) = \mathbb{E}_n\left\{\kappa\left(x - x_n\right)\right\} \quad , \tag{4-1}$$

where $\kappa\left(\cdot\right)\in\mathbb{R}^+$ is a symmetric kernel function and notation $\mathbb{E}\left\{\cdot\right\}$ stands for averaging operator. Though there are many feasible functions, the Gaussian is commonly preferred. In this case, the Gaussian kernel can be defined for the input domain $\mathscr{X}$ as:

$$\kappa_G\left(x - x'; \sigma_X\right) = \exp\left(\frac{-\|x - x'\|_{\mathscr{X}}^2}{2\sigma_X^2}\right), \tag{4-2}$$

where $\|\cdot\|_{\mathscr{X}}$ is a given norm in $\mathscr{X}$).

Provided the observation set X and based on the Parzen's estimation of eq. (4-1), we get the following estimator of the Renyi's α-order entropy [121]:

$$H_\alpha(X) = \frac{1}{1-\alpha}\log\left(\mathbb{E}_x\left\{p(x)^{\alpha-1}\right\}\right) \approx \hat{H}_\alpha(X|\sigma)$$
$$= \frac{1}{1-\alpha}\log\left(V_\alpha(X|\sigma)\right),$$

where the so termed information potential (IP) $V_\alpha(X|\sigma)$ of the set X is defined as follows:

$$V_\alpha(X|\sigma) = \mathbb{E}_n\left\{v_\alpha(x_n|\sigma_X)\right\}, \tag{4-3}$$

being $v_\alpha(x_n|\sigma)$ the IP of the sample x_n, which can be computed as:

$$v_\alpha(x_n|\sigma_X) = \frac{1}{N^{\alpha-1}}\sum_{n'=1}^{N}\left(\kappa_G(x_n - x_{n'};\sigma_X)\right)^{\alpha-1}. \tag{4-4}$$

4.2. Kernel function estimation from information potential variability - (KEIPV)

From eq. (4-4) we can infer that the IP yields an entropy estimate that is based on the summation of pairwise sample interactions through the Gaussian kernel function [103]. Also, the Information Force (IF), $F_n \in \mathscr{X}$, is defined as the force acting on particle x_n due to all other particles in X and is given by the derivative of the IP with respect to x_n. Particularly, for the case of α=2, the well-known quadratic Renyi's entropy leads to the following estimation of the IF:

$$F_n = \frac{\partial}{\partial x_n}V_2(X|\sigma_X) = -(N\sigma_X)^{-2}\sum\nolimits_{x_{n'}\in X}\kappa_G(x_n - x_{n'};\sigma_X)(x_n - x_{n'})$$
$$= \mathbb{E}_{n'}\left\{F(x_n|x_{n'})\right\} \quad , \tag{4-5}$$

where

$$F(x_n|x_{n'}) = (N\sigma_X^2)^{-1}\kappa(x_n - x_{n'};\sigma_X)(x_n - x_{n'}) \tag{4-6}$$

corresponds to the conditional IF acting on x_n due to $x_{n'}$. Generally, the IFs can be interpreted in light of inner products in a high dimensional feature space [74]. Some important facts have to be highlighted from eq. (4-5):

- On one hand, given that $\boldsymbol{X}$ is fixed and the factor $(x_n - x_{n'})$ points towards x_n, all IF directions are also fixed and attracting-natured.

- On the other hand, since F_n turns out to be dependent on the free parameter σ_X, the IP and all IF magnitudes become functions of the Gaussian kernel bandwidth. In fact, the IP follows a monotonically decreasing behavior over σ_X.

- At the same time, the conditional IF magnitude tends to zero as σ_X goes either to zero or infinite and reaching its maximum at some value in $\mathbb{R}^+$.

Hence, the importance of an adequate Gaussian kernel bandwidth tuning becomes clear. In this sense, we seek for an RKHS maximizing the overall IP variability with respect to the kernel bandwidth parameter so that all IF magnitudes spread the most widely on $\mathscr{X}$. To this end, the variability of the estimated IP is maximized in terms of the kernel bandwidth parameter as follows:

$$\sigma_X^* = \arg\max_{\sigma_X} \operatorname{var}\{v_2(x|\sigma_X)\}, \tag{4-7}$$

where

$$\operatorname{var}\{v_2(x|\sigma_X)\} = \mathbb{E}_x\left\{\left(\operatorname{var}\{v_2(x|\sigma_X)\} - \mathbb{E}_x\{\operatorname{var}\{v_2(x|\sigma_X)\}\}\right)^2\right\}. \tag{4-8}$$

Deriving with respect to σ_X, the optimal parameter value can be rewritten in terms of the before introduced Gaussian-based Renyi's Information Metrics as follows:

$$\begin{aligned}\frac{d}{d\sigma_X}\operatorname{var}\{v_2(x|\sigma_X)\} &= \frac{2}{N^2\sigma_X^3}\left(1+\frac{1}{N}\right)\left(\sum_{n,n'=1}^{N}\kappa_G^2\left(x_n - x_{n'};\sigma_X\right)\|x_n - x_{n'}\|_{\mathscr{X}}^2\right.\\ &\quad\left. - \left(\sum_{n,n'=1}^{N}\kappa_G\left(x_n - x_{n'};\sigma_X\right)\right)\left(\sum_{n,n'=1}^{N}\kappa_G\left(x_n - x_{n'};\sigma_X\right)\|x_n - x_{n'}\|_{\mathscr{X}}^2\right)\right),\\ &= \frac{2(N^2+N)}{\sigma_X}\left(\sigma_X^2\sum_{n,n'=1}^{N}F^2(x_n|x_{n'}) - V_2(X)\sum_{n,n'=1}^{N}\left(F(x_n|x_{n'})\right)^\top(x_n - x_{n'})\right)\end{aligned}$$

Lastly, equating the above equation to zero, a fixed point or a gradient descent update rule can be employed to find a suitable σ_X value. As a result, we get a scale updating rule as a function of the IFs, which are induced by a kernel function applied over a finite sample set. Thereby, a Gaussian kernel-based RKHS coding the most spread out IF magnitudes can be estimated using the introduced approach, termed as: *Kernel Function Estimation from Information Potential Variability* - KEIPV.

4.3. Experimental set-up

Given an input representation sample matrix $\boldsymbol{X}\in\mathscr{X}\subset\mathbb{R}^{N\times P}$, being N and P the number of provided samples and features, respectively, we test the proposed KEIVP approach as a tool to find relevant data relationships. For the concrete case, a clustering task is considered. Thus, both synthetic and real-world datasets are studied. The former is a toy set holding two multivariate Gaussian distributions (see fig. 4.1a): $f_1\left(x\right)=\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma_1})$ and $f_2\left(x\right)=\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma_2})$, with parameters $\boldsymbol{\mu}_1=\mathbf{0}$, $\boldsymbol{\mu}_2=\mathbf{1}$, $\boldsymbol{\Sigma}_1=0.5\boldsymbol{I}$ and $\boldsymbol{\Sigma}_2=0.25\boldsymbol{I}$, with $\boldsymbol{\mu}_1,\boldsymbol{\mu}_2\in\mathbb{R}^2$ and being $\boldsymbol{I}\in\mathbb{R}^{2\times2}$ the identity matrix. To get the input sample set $\boldsymbol{X}\in\mathbb{R}^{200\times2}$, one hundred samples are randomly drawn from each of both simulated pdfs.

Also, to provide visual inspection on unsupervised clustering, three well-known synthetic databases are used that represent challenging clustering tasks due to their complex structures: *Bull's eyes*, *Circle with squares*, and *Noisy squares* (see fig. **4-2** rows one, two, and three, respectively). Here, three baseline approaches for estimating the Gaussian kernel bandwidth parameter are considered:

- The *Sylverman's* rule criterion that computes the scale value as:

$$\sigma_S=\sigma_X\left(4N^{-1}(2P+1)^{-1}\right)^{1/(P+4)}, \tag{4-9}$$

  with $\sigma_X=\sum_{n\in N}s_{nn}$ and being s_{nn} the diagonal elements of the sample covariance matrix [137].

- The *Self-Tuning Spectral Clustering* (STSC) estimator that calculates a local scale parameter for each pair of row sample vectors $(\boldsymbol{x}_n,\boldsymbol{x}_{n'}\in\mathbb{R}^P)$, $n\neq n'$, by considering nearest neighbor distances as:

$$\sigma_{sc}^{nn'}=\|\boldsymbol{x}_n-\boldsymbol{x}_n^K\|_2\|\boldsymbol{x}_{n'}-\boldsymbol{x}_{n'}^K\|_2, \tag{4-10}$$

  being $\boldsymbol{x}_K^n$ the K-th nearest neighbor of $\boldsymbol{x}_n$ in terms of the Euclidean distance and $\|\cdot\|_2$ stands for the 2-norm [166].

- The local density adaptive band-width is also tested, which computes a local scale parameter as function of *Common Near Neighbors* (CNN) between points $(\boldsymbol{x}_n,\boldsymbol{x}_{n'})$, $n\neq n'$, as:

$$\sigma_{cnn}^{nn'}=\sigma_o\left(\gamma\left(\boldsymbol{x}_n,\boldsymbol{x}_{n'}\right)+1\right)^{1/2}, \tag{4-11}$$

  where $\sigma_o\in\mathbb{R}^+$ and $\gamma\left(\boldsymbol{x}_n,\boldsymbol{x}_{n'}\right)=|\Gamma_n\cap\Gamma_{n'}|$, being $\Gamma_n=\{\boldsymbol{x}_n^k:k=1,\ldots,K\}$ the set holding the K nearest neighbors of $\boldsymbol{x}_n$ according to the Euclidean distance and $|\cdot|$ stands for the cardinality operator [169]. Here, $\sigma_o=\text{median}\{\sigma_{sc}^{n,n'}\}$, $n<n'$.

For each of above introduced bandwidth selection approaches, namely Sylverman, STSC, CNN, and KEIVP, the resulting Gaussian kernel is employed to perform the unsupervised clustering learning by means of the well-known Spectral Clustering technique [110]. Additionally, the number of neighbors is fixed as $K=\sqrt{N}$ in cases of STSC and CNN. For concrete testing, the number of groups $C\in\mathbb{N}$ is fixed as three, three, and five, respectively. Furthermore, for fair comparison purposes, the KEIVP approach is calculated only considering data relationships (distances) belonging to connected samples according to a K-nearest graph.

Finally, the real-world databases from the Machine Learning UCI Repository [1] are tested as supervised clustering task (see table **4-1**). In this case, each computed kernel is used as similarity representation to learn a classification boundary based on the well-known k-nearest-neighbors classifier. A 10-folds-cross-validation strategy is carry out to validate the stability of each kernel function estimation approach. Furthermore, the k parameter is fixed from the set $\{1, 3, 5, 7, 9, 11\}$ according to the training error.

4.4. Results and discussion

As seen in figs. 4.1b to 4.1h, the IP variability cost function allows identifying different IF configurations in the two multivariate Gaussian distributions dataset. Particularly, for a narrow bandwidth value, particles are forced to apart each other due to the kernel function strongly reduces the scaling of the Euclidean-based distance between particles. Hence, low similarities between pair-wise samples and low magnitude IFs are estimated, as shown in figs. 4.1c and 4.1f. In contrast, employing a wide bandwidth value yields to an RKHS where all particles are attracted each other. Namely, the Euclidean distance scaling is strongly increased, which leads to a data representation space where all samples are closed similar, as seen in fig. 4.1e. Such a fact is shown in the IF distribution in fig. 4.1h, where red cluster particles are more attracted to the green particle. Note that low IP variability values are achieved for both narrow and wide bandwidths because, in either case, all the IFs tend to share the same magnitude regardless their direction. Therefore, the proposed KEIVP finds an RKHS where data samples share widely spread IF magnitudes, that is, close particles according to the Euclidean distance get high pairwise similarities and IFs while far ones have low similarities and IFs (see figs. 4.1d and 4.1g).

Regarding the other synthetic datasets, figs. 4.2b to 4.2d show that both local scaling-based strategies (STSC and CNN) as well as the proposed KEIPV are able to deal with the *Bull's eyes* structure. Such approaches also correctly perform grouping of the *Noisy squares* dataset, as seen in figs. 4.2j to 4.2l. That is, local scaling-based techniques are able to approximate nonlinear structures from linear analysis over each sample neighborhood. Nonetheless, STSC performs wrong clustering for the *Circle with squares* (see fig. 4.2f). These results can be explained by the fact that local scaling approximations lead to wrong

[1] `http://archive.ics.uci.edu/ml/`

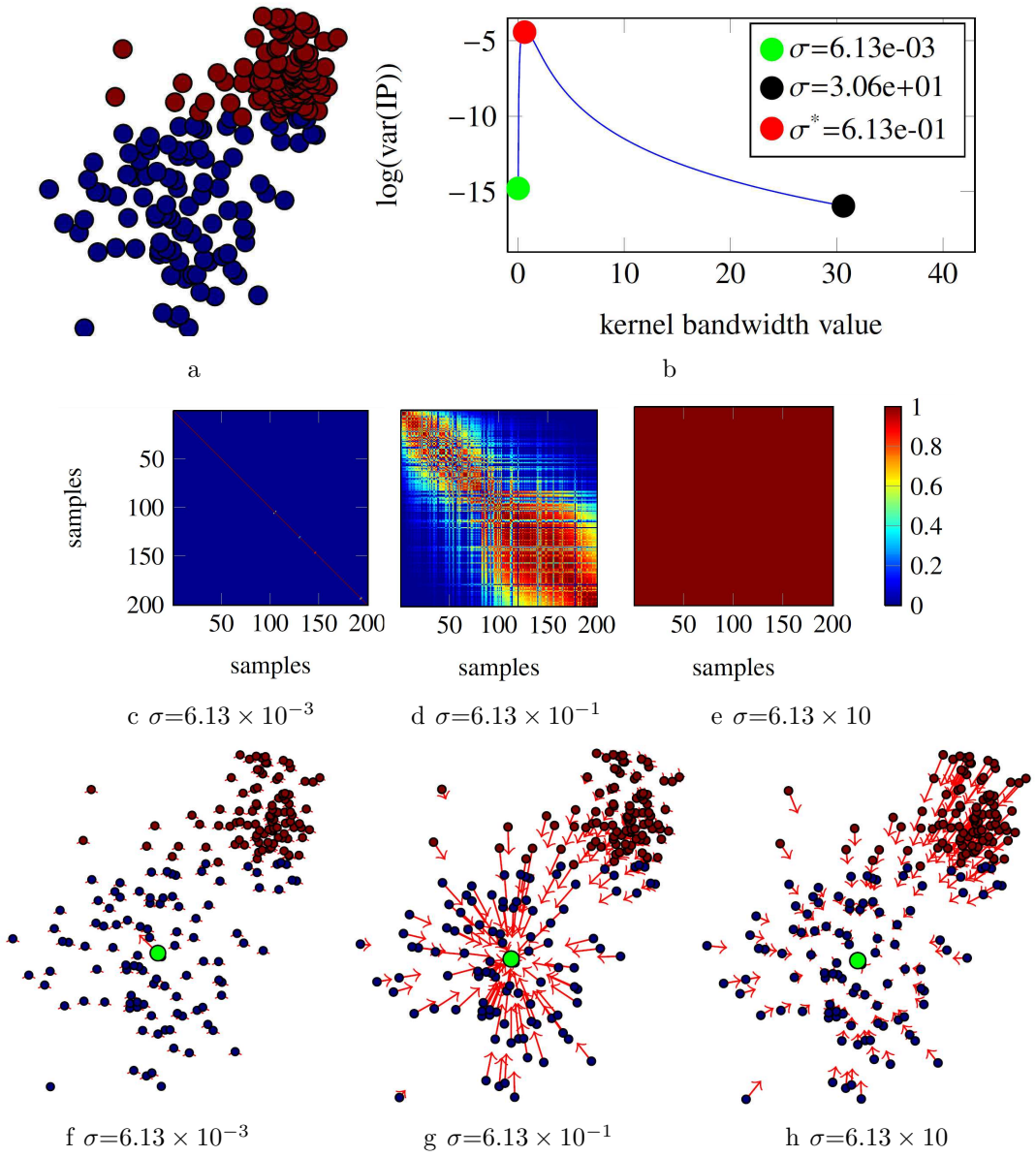


Figure 4-1.: KEIVP illustrative example. a) Multivariate Gaussian toy set. b) log of IP variability versus bandwidth. **2nd row**: Gaussian kernel for the toy set. **3rd row**: IFs acting on a fixed particle (green). Narrow (**1st column**), KEIVP (**2nd column**) and wide (**3rd column**) bandwidth values.

cluster connections when dealing with data structures with highly varying densities. Similarly, CNN suffers of the same drawback, but the σ_o parameter can deal with it if properly fixed. Nonetheless, finding a suitable neighborhood size is a difficult task for the user, not mentioning that using different bandwidth values for each pair-wise sample similarity when estimating a Gaussian kernel does not guarantee a positive definite kernel function, violat-

ing the Mercer's conditions [118]. Regarding to the Sylverman-based estimation results, this method generally yields a biased RKHS due to its statistical assumptions, resulting in wrong clustering performances (see figs. 4.2a, 4.2e and 4.2i). In turn, KEIPV is able to find an RKHS coding widely spread IF magnitudes, allowing to close samples belonging to a similar structure while repelling distant points (see fourth column of fig. **4-2**).

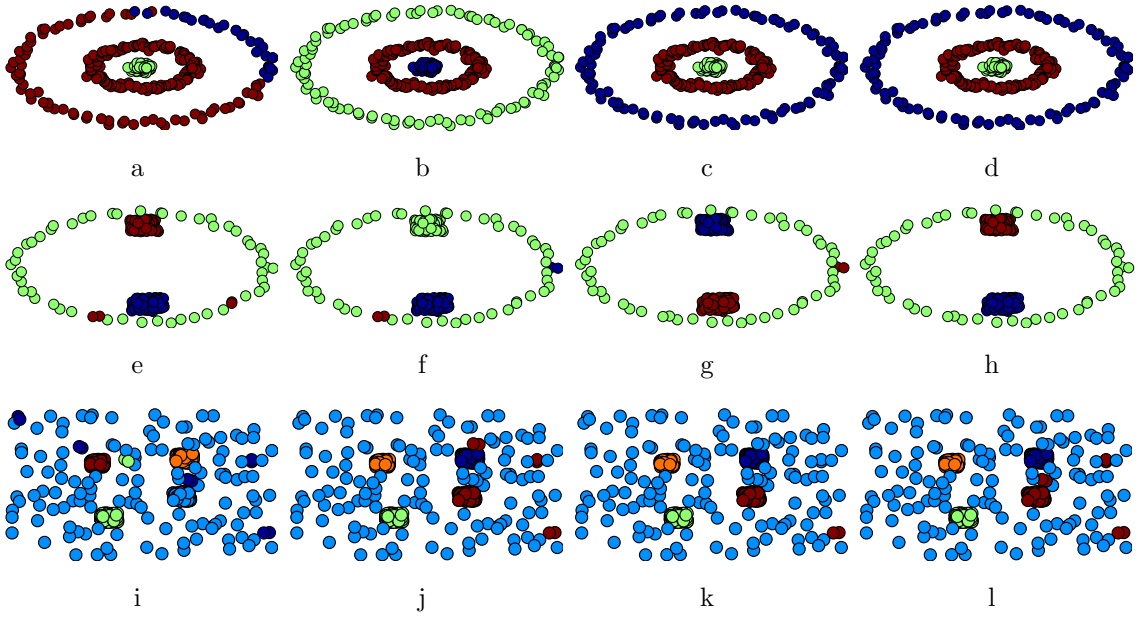


Figure 4-2.: Synthetic datasets clustering results. **1st row**: Bull's eyes. **2nd row**: Circle with squares. **3rd row**: Noisy squares. **1st column**: Sylverman's rule. **2nd column**: STSC. **3rd column**: CNN. **4rd column**: KEIPV.

Finally, with respect to the real-world databases, as seen in fig. **4-3**, the proposed KEIPV allows to compute an RKHS favoring the cluster separability. STSC and CNN algorithms get competitive results in terms of classification accuracy. Nonetheless, they need a suitable graph representation, which practically can be difficult to estimate. Moreover, their local scaling approximation of the Gaussian kernel can not be correct theoretically as mentioned before. Again, the Sylverman's rule estimation suffers of biased kernel representations, particularly, when the input dimensionality P is considerably high (see obtained results by the `mnist` and `orl` datasets).

Table **4-1**.: Employed UCI dataset description

Dataset	`iris`	`sonar`	`mnist`	`orl`	`diabetes`	`breast`	`arrhythmia`	`ionosphere`	`heart`	`wine`	`glass`
N	150	208	1000	400	768	699	420	351	303	178	214
P	4	60	784	10304	8	9	278	34	13	13	9
C	3	2	10	40	2	2	13	2	2	3	4

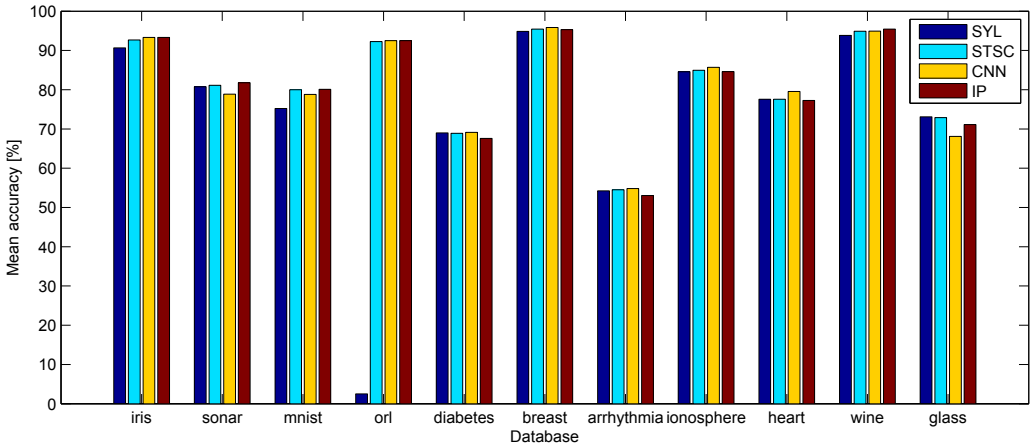


Figure 4-3.: Classification results using the fourth bandwidth selection approaches.

4.5. Summary

In this study, a new kernel function estimation based on an information potential variability framework is presented. Our approach, termed KEIPV, aims to estimate an RKHS to span the most widely information force magnitudes among data points. Particularly, KEIPV relates different kernel functions with the intrinsic information potential variations in Parzen-based pdf estimations [121]. Thereby, we seek for an RKHS that maximizes the overall information potential variability with respect to the global kernel parameter. In this sense, an automatic kernel-based relevant representation is computed by including the input data statistical distribution within an ITL framework. As a case of interest, an updating rule for estimating the Gaussian kernel bandwidth parameter is proposed as a function of the forces induced by the distances among samples. Proposed strategy is tested on both unsupervised and supervised clustering tasks. Performed results show that the presented approach allows computing RKHS's favoring data groups separability in comparison with other state-of-the-art alternatives.

5. Kernel representation based on information theoretic learning for Gramm matrices: a dimensionality reduction approach

The world is essentially described and represented by multidimensional data which reside in *High-Dimensional*-(HD) spaces, e.g., image and video, neural activity, biological signals, weather forecasting, economy, etc. Regarding this, methods of *Dimensionality Reduction*-(DR) provide a way to understand and visualize the structure of HD [84]. So, DR aims at producing meaningful representations of the input HD data to a Low-Dimensional-(LD) space. Regarding this, the general intuition that drives DR is that close or similar data items should be represented near each other, whereas dissimilar ones should be represented far from each other [83]. Namely, DR preserves as much of the relevant structure of the HD data as possible in the LD representation. Hence, the successful of the DR approach resides in two main issues: defining a notion of pair-wise relations in both the HD and the LD spaces, and measuring the mismatch between HD and LD spaces according to the imposed data relations.

Several DR techniques have been proposed, which differ each other in the type of structure that they preserve, e.g., variance, dot products, dissimilarities (distances), similarities, or other local/global measures of proximity [87]. Thereby, the variety of the structure that they preserve has lead to the development of a large number of methods. The well-known *Principal Component Analysis*-(PCA) algorithm can be considered as the oldest DR method [84]. PCA finds a linear projection of the original data which captures as much variance as possible. In other words, new features are generated by linear combinations of the original ones by optimizing a maximum/minimum loss of information criterion. Another linear DR method is the *Classical Multidimensional Scaling*-(MDS), which maximize the dot product preservation [18]. Nonlinear variants of the metric MDS appears, such as: *Sammons's Nonlinear Mapping*-(SNM) [127] and the *Curvilinear Component Analysis*-(CCA) [64]. These methods are based on notions like topology and neighborhood preservation, however, their main limitations come from the distortions between the distances measured in the input space and the distances measured in the manifold space, leading to a biased mismatch between the HD and the LD data relations.

Additionally, some methods like *Curvilinear Distances Analysis* (CDA) [88], and *Isometric Mapping*-(ISOMAP) [149], are nonlinear methods derived from MDS, which use as metric the curvilinear or geodesic distance. This metric (geodesic distance) can measure good approximations of the distances along the manifold, without shortcuts as does the Euclidean distance. The goal of the geodesic distance on these algorithms consists in computing distance along an object making possible the projection of nonlinear manifolds. A difference between CDA and ISOMAP is that while ISOMAP relies on algebraical methods resulting from the reformulation of the PCA problem as a distance preservation problem, CDA works by optimizing a criterion that explicitly measures the preservation of the pair-wise distances. One of the main advantages of a global approach like ISOMAP, is that a solution always exists for any problem in the framework of the considered model. However, to price to pay is often unrealistic or too constraining model. In other words, the solution always exists as a global minimum, but, when the problem does not fit the model, its interpretation could be hazardous [89]. On the contrary, a local approach like CDA does not offer any theoretical guarantee. Even when the problem perfectly fits the model, an unexperienced user might badly parameterize the algorithm. Lastly, another disadvantage is that several parameters need to be tuned as well for ISOMAP as for CDA.

Nowadays, more developed methods aimed at preserving the data topology have been proposed from both spectral and divergence-based functions. On the one hand, the spectral approaches are represented by methods such as: *Locally Linear Embedding*-(LLE) [125], *Laplacian Eigenmaps*-(LEM) [14], *Hessian LLE*-(HLLE) [38], and *Diffusion Maps*-(DM) [105]. In LLE each datum is approximated by a linear combination of its neighbors in the HD space and the obtained coefficients are then used to compute its LD coordinates. In other words, LLE attempts to preserve the local geometry. LEM is a geometrically algorithm for constructing a representation for data sampled from a low dimensional manifold embedded in a higher dimensional space. The algorithm provides a computationally efficient approach to nonlinear dimensionality reduction that has locality preserving properties and a natural connection to clustering. HLLE achieves linear embedding by minimizing the Hessian functional on the manifold where the data set resides. Furthermore, the conceptual HLLE may be viewed as a modification of the LEM framework. Finally, DM is a based probabilistic interpretation of spectral clustering and DR algorithms that use the eigenvectors of the normalized graph Laplacian. All of these methods, however, require each HD object to be associated with only a single location in the LD space. This makes it difficult to unfold "many-to-one" mappings in which a single ambiguous object really belongs in several disparate locations in the LD space [66].

On the other hand, the divergence-based methods, are represented mainly by Stochastic Neighbor Embedding (SNE) [66] and its variants, i.e, t-SNE [155], Jensen-Shannon Embedding (JSE) [83], among others. The main difference between spectral methods and SNE-based variants, is that SNE matches similarities that are computed both in the HD and the LD space, while spectral methods directly convert the pair-wise similarities defined in the HD

space into inner products . Thus, SNE and its variants are based on similarity preservation instead of distance preservation and makes them robust against the phenomenon of norm concentration. Nonetheless, divergence-methods suffer from reaching distorted and overlapped latent spaces, moreover, the user must to tune for each dataset several parameters in order to obtain suitable representations and embeddings [22, 116].

Recently, ITL-based quantities have been employed in machine learning as descriptors of the data distributions that go beyond second order statistics. The use of information theoretic quantities as descriptors of data requires the development of suitable probability law estimators. Regarding this, ITL and kernel-based methods have been studied in order to connect each other as tool to introduce useful high-order statics in a data-driven way [121]. Namely, previous approaches define estimators of a conventional information theoretic quantity, such as Shannon entropy, to build quantities from the data that satisfies similar axiomatic properties to those of well establish definitions such as Renyi's definition of entropy [128]. In particular, Gram matrix obtained from evaluating a positive definite kernel on samples can be used to define a quantity based on the data with properties similar to those of an entropy without assuming that the probability density is being estimated [52].

In this study, we introduce a kernel-based representation that considers the statistical distribution and the salient data structures from information theory-based constraints. For such a purpose, an entropy-like functional on positive definite matrices based on Renyi's definition is employed to discover relevant data regularities. Due to the connections with topology-based and divergence-based DR techniques, our approach, termed *kernel-based entropy dimensionality reduction* (KEDR), is applied as a representation tool to measure the mismatch between HD and LD data representation spaces. Therefore, the proposed approach employs estimators of entropy-like quantities for Gram matrices that can be computed by evaluating infinitely divisible kernels on pairs of samples to find a relevant representation space. Testing is carried out on synthetic and real-world datasets in terms of both visual inspection and neighborhood preservation (rank-based criteria). Overall, the introduced KEDR is competitive in comparison to the state-of-the-art methods, being able to encode HD data relationships by computing LD representations where the local and the global structures are preserved.

5.1. Gram matrix estimation of Renyi's α-entropy

ITL uses the conventional learning and adaptation methodologies of adaptive filters, neural networks, and kernel learning. Instead of the commonly used Shannon definition, this framework expresses the optimality criteria in terms of the following Renyi's entropy-like functional [121]:

$$H_\alpha(X) = \frac{1}{1-\alpha} \log\left(\int_{\mathscr{X}} p^\alpha(x) dx\right), \tag{5-1}$$

where $p(x)$ is the probability density function (pdf) of a random variable $X{\in}\mathscr{X}$, $\mathscr{X}$ is the support, $x{\in}X$ is a given sample, and $\alpha{\in}\mathbb{R}^+$ is a parameter providing a family of entropy functionals, where the conventional Shannon's entropy is the asymptotical case of $\alpha \to 1$. Provided $\{x_n{\in}\mathscr{X} : n{\in}[1, N]\}$ as an *i.i.d.* sample of N realizations of X, an effective plug-in estimator of eq. (5-1) can be derived for $\alpha{=}2$ using the Parzen window approximation, $\hat{p}(x_n){=}\mathbb{E}_{nn'}\left\{\kappa(x_n, x_{n'}) : \forall n, n'{\in}[1, N]\right\}$, as follows:

$$\hat{H}_2\left(X\right) = -\log\left(\mathbb{E}_{nn'}\left\{h(x_n, x_{n'})\right\}\right) \approx -\log\left(\frac{1}{N^2}\sum_{n,n'=1}^{N} h(x_n, x_{n'})\right), \tag{5-2}$$

For a given Gram matrix $\boldsymbol{A}{\in}\mathbb{R}^{N\times N}$ with elements $a_{ij}{=}\kappa\left(x_n, x_{n'}\right)$, with $a_{nn'}{\in}\mathbb{R}^+$, then, eq. (5-2) can be rewritten as:

$$\hat{H}_2\left(X\right) = -\log\left(\mathrm{tr}\left(\boldsymbol{A}\boldsymbol{A}\right)/N^2\right) + c_\kappa, \tag{5-3}$$

where $c_\kappa{\in}\mathbb{R}^+$ is a constant that accounts for the normalization factor of the Parzen window and notation $\mathrm{tr}\left(\cdot\right)$ stands for the matrix trace. As a result, the entropy estimator in eq. (5-3) can be related to the norm of the Gram matrix $\boldsymbol{A}$, i.e., $\|\boldsymbol{A}\|_2{=}\mathrm{tr}\left(\boldsymbol{A}\boldsymbol{A}\right)$.

Recently, a generalized entropy functional for a given Gram matrix set, $\mathscr{M}{=}\{\boldsymbol{A}_k{\in}\mathbb{R}^{N\times N}\}$, arises as a new information-theoretic interpretation that can be employed as objective functions of ITL [52]. Regarding this, let $f(\boldsymbol{A}_k){=}\boldsymbol{A}_k^\alpha$ be a continuous scalar-valued matrix function defined over all the positive definite matrices $\boldsymbol{A}_k{\in}\mathscr{M}$ $\left(\mathrm{tr}\left(\boldsymbol{A}_k\right){\leq}1\right)$ based on the spectral decomposition theorem [69]. So, a matrix-based functional analogue to Renyi's α-entropy can be defined as [129]:

$$S_\alpha(\boldsymbol{A}_k) = \frac{1}{1-\alpha}\log\left(\mathrm{tr}\left(\boldsymbol{A}_k^\alpha\right)\right), \tag{5-4}$$

where it holds that $\mathrm{tr}\left(\boldsymbol{A}_k\right){=}\mathrm{tr}\left(\boldsymbol{A}_l\right){=}1$, $\forall k{\neq}l$.

For $\alpha{\neq}1$, the functional $S_\alpha(\boldsymbol{A}_k)$ satisfies some properties attributed to entropy under the following conditions [52]:

- $S_\alpha(c\boldsymbol{A}_k)$ is a continuous function for $0 < c \leq 1$.
- $S_\alpha(\boldsymbol{A}_l\boldsymbol{A}_k\boldsymbol{A}_l^*){=}S_\alpha(\boldsymbol{A}_k)$ for any orthonormal matrix $\boldsymbol{A}_l{\in}\mathscr{M}$.
- $S_\alpha(\boldsymbol{A}_k) \leq S_\alpha\left(\boldsymbol{I}_N/N\right){=}\log\left(N\right)$, where $\boldsymbol{I}_N{\in}\mathbb{R}^{N\times N}$ is the identity matrix.
- $S_\alpha(\boldsymbol{A}_k \otimes \boldsymbol{A}_l){=}S_\alpha(\boldsymbol{A}_k) + S_\alpha(\boldsymbol{A}_l)$, where $\otimes$ stands for the tensor product operator [91].

- If $\boldsymbol{A}_k\boldsymbol{A}_l{=}\boldsymbol{A}_l\boldsymbol{A}_k{=}\boldsymbol{0}_N$, then for the strictly monotonic continuous function $g(\boldsymbol{x}){=}2^{(\alpha-1)\boldsymbol{x}}$ it holds that:

$$S_\alpha(c\boldsymbol{A}_k+(1-c)\boldsymbol{A}_l){=}g^{-1}\left(cg(S_\alpha(\boldsymbol{A}_k))\right)+(1-c)g(S_\alpha(\boldsymbol{A}_l)).$$

- If the rank of $\boldsymbol{A}_k$, $\rho(\boldsymbol{A}_k)$, is equal to 1, then the entropy $S_\alpha(\boldsymbol{A}_k){=}0$ for any $\alpha{\neq}0$.

Additionally, from the functional in Eq. (5-4), the Gramm matrix estimation of joint entropy, conditional entropy, and mutual information can be defined as follows:

– *Joint entropy:*

$$S_\alpha\left(\boldsymbol{A}_k,\boldsymbol{A}_l\right)=S_\alpha\left(\frac{\boldsymbol{A}_k\circ\boldsymbol{A}_l}{\mathrm{tr}\left(\boldsymbol{A}_k\circ\boldsymbol{A}_l\right)}\right), \tag{5-5}$$

where both $\boldsymbol{A}_k$ and $\boldsymbol{A}_l$ are assumed to be positive-definite matrices with unit-trace and nonnegative entries, and $\circ$ stands for the Hadamard product operator. Since the following inequality must be satisfied (see eqs. (5-4) and (5-5)): $S_\alpha\left(\boldsymbol{A}_k,\boldsymbol{A}_l\right){\geq}S_\alpha(\boldsymbol{A}_r)$, with $r{=}\{k,l\}$, the joint entropy should never be smaller than any of the constituent entropies.

– *Conditional entropy* [147]:

$$S_\alpha(\boldsymbol{A}_k|\boldsymbol{A}_l)=S_\alpha\left(\boldsymbol{A}_k,\boldsymbol{A}_l\right)-S_\alpha(\boldsymbol{A}_l), \tag{5-6}$$

where the conditional entropy in eq. (5-6) is nonnegative and upper bounded as:

$$S_\alpha(\boldsymbol{A}_k|\boldsymbol{A}_l)=S_\alpha\left(\boldsymbol{A}_k,\boldsymbol{A}_l\right)-S_\alpha(\boldsymbol{A}_l)\leq S_\alpha(\boldsymbol{A}_k).$$

– *Mutual information*:

$$I_\alpha(\boldsymbol{A}_k;\boldsymbol{A}_l)=S_\alpha(\boldsymbol{A}_k)+S_\alpha(\boldsymbol{A}_l)-S_\alpha\left(\boldsymbol{A}_k,\boldsymbol{A}_l\right), \tag{5-7}$$

where $I(\boldsymbol{A}_k;\boldsymbol{A}_l){\geq}0$ and $S_\alpha(\boldsymbol{A}_k){\geq}I_\alpha(\boldsymbol{A}_k;\boldsymbol{A}_k)$.

For all provided Renyi's α-entropy functionals, the imposed normalization is an important property of the involved matrices. Namely, if $\boldsymbol{A}_k$ and $\boldsymbol{A}_l$ are normalized to have unit trace, then, for $\alpha{\in}[0,1]$ the product $\boldsymbol{A}_k^{\circ\alpha}\circ\boldsymbol{A}_l^{\circ(1-\alpha)}$ is also normalized, where $\boldsymbol{A}_k^{\circ\alpha}$ denotes the entry-wise α-th power of $\boldsymbol{A}_k$. However, it is not always true that the resulting matrix is positive definite. Indeed, this product can be seen as a weighted geometric average for which the resulting matrix will give more emphasis to either one of the matrices. Consequently, $\boldsymbol{A}_k$ and $\boldsymbol{A}_l$ must be infinitely divisible kernels to guarantee the product to be positive definite [52].

5.2. Kernel-based entropy dimensionality reduction (KEDR)

Let $\boldsymbol{X}=\{\boldsymbol{x}_n\in\mathbb{R}^P : n\in[1,N]\}$, be a *High-Dimensional*-(HD) finite sample set, and let $\boldsymbol{K}\in\mathbb{R}^{N\times N}$ be a kernel matrix with unit trace coding non-linear data relations in $\mathbb{R}^P$ as:

$$k_{nn'} = \kappa_X\left(\boldsymbol{x}_n, \boldsymbol{x}_{n'}\right), \tag{5-8}$$

where $\kappa_X : \mathbb{R}^P \times \mathbb{R}^P \to \mathbb{R}$ is a positive-definite and infinitely divisible kernel function.

Now, let $\boldsymbol{Y}=\{\boldsymbol{y}_n\in\mathbb{R}^M : n\in[1,N]\}$ be a *Low-Dimensional*-(LD) representation of $\boldsymbol{X}$ ($M \leq P$), the corresponding elements of the LD similarity matrix $\boldsymbol{L}\in\mathbb{R}^{N\times N}$ can be computed as:

$$l_{nn'} = \kappa_Y\left(\boldsymbol{y}_n, \boldsymbol{y}_{n'}\right), \tag{5-9}$$

being $\kappa_Y : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$ a positive-definite and infinitely divisible kernel function. Based on the Gaussian kernel, the HD similarities are estimated as follows:

$$\kappa_X\left(\boldsymbol{x}_n, \boldsymbol{x}_{n'}; \sigma_X\right) = \exp\left(\frac{-\|\boldsymbol{x}_n - \boldsymbol{x}_{n'}\|_2^2}{2\sigma_X^2}\right), \tag{5-10}$$

where $\sigma_X\in\mathbb{R}^+$ is the kernel bandwidth. Similarly, the kernel function κ_Y in LD space can be estimated as:

$$\kappa_Y\left(\boldsymbol{y}_n, \boldsymbol{y}_{n'}; \sigma_Y\right) = \exp\left(\frac{-\|\boldsymbol{y}_n - \boldsymbol{y}_{n'}\|_2^2}{2\sigma_Y^2}\right), \tag{5-11}$$

where $\sigma_Y\in\mathbb{R}^+$.

Taking into account that the Shannon entropy corresponds to a particular case of the Renyi's α-entropy functional, and exploiting the ITL-based extension of the entropy measures for Gram matrices (see eq. (5-4)), here, we introduce a DR framework by means of a matrix-based Renyi's α-entropy. Particularly, due to the probabilistic interpretation of the similarities between HD and LD samples encoded in matrices $\boldsymbol{K}$ and $\boldsymbol{L}$, respectively, the entropy-like functional on positive definite matrices based on α-Renyi's axiomatic is employed to quantify the DR mismatch.

In this sense, we introduced a *Kernel-based Entropy Dimensionality Reduction*-(KEDR) approach, where the mismatch between the HD and the LD similarities are quantified as:

$$J(\boldsymbol{Y}) = S_\alpha(\boldsymbol{K}|\boldsymbol{L}) = S_\alpha\left(\boldsymbol{K}, \boldsymbol{L}\right) - S_\alpha(\boldsymbol{L}). \tag{5-12}$$

Thereby, the KEDR optimization problem can be written as:

$$\boldsymbol{Y}^* = \arg\min_{\boldsymbol{Y}} J(\boldsymbol{Y}).$$

According to the KEDR formulation in Eq (5-12), the term $S_\alpha(\boldsymbol{L}) \leq S_\alpha\left(\frac{1}{N}\boldsymbol{I}\right)$ encourages a sparse LD-space representation while $S_\alpha(\boldsymbol{K}, \boldsymbol{L})$ aims to match the pair-wise relationships k_{ij} and l_{ij}. Thereby, the sparsity and the pair-wise relationships matching are equally penalized in KEDR. Furthermore, based on the analogy between Shannon and Renyi's α-entropies, different extensions of the KEDR cost function introduced in Eq. (5-12) can be aimed out to build entropy-based functionals for Gram matrices that leads to flexible penalties in DR mapping. In consequence, such functionals are described as follows:

i) Type 1 mixture of Renyi's α-conditional entropies-KEDR-(T1KEDR). In this case, the KEDR functional can be extended as a mixture of two conditional entropies as follows:

$$J(\boldsymbol{Y}) = (1-\gamma)S_\alpha(\boldsymbol{K}|\boldsymbol{L}) + \gamma S_\alpha(\boldsymbol{L}|\boldsymbol{K}), \tag{5-13}$$

where $\gamma \in [0, 1]$. Then, writing eq. (5-13) in terms of marginal and joint entropies yields:

$$\begin{aligned} J(\boldsymbol{Y}) =& (1-\gamma)\left(S_\alpha(\boldsymbol{K}, \boldsymbol{L}) - S_\alpha(\boldsymbol{L})\right) + \gamma\left(S_\alpha(\boldsymbol{K}, \boldsymbol{L}) - S_\alpha(\boldsymbol{K})\right) \\ =& S_\alpha(\boldsymbol{K}, \boldsymbol{L}) - (1-\gamma)S_\alpha(\boldsymbol{L}) - \gamma S_\alpha(\boldsymbol{K}). \end{aligned} \tag{5-14}$$

As seen in eq. (5-14) the higher the γ value ($\gamma \to 1$) the higher the tendency of the DR algorithm for preserving the input space entropy encoded into the HD kernel matrix $\boldsymbol{K}$, that is, the embedding only cares about keeping, as well as possible, the relationships between pair-wise elements $l_{nn'}$ and $k_{nn'}$ in $S_\alpha(\boldsymbol{K}, \boldsymbol{L})$. Moreover, as $\gamma \to 0$, T1KEDR approximates the KEDR solution. Hence, T1KEDR can be seen as a regularized version of KEDR, where the trade-off between sparsity and pair-wise relationships matching is controlled by the γ parameter value.

ii) Type 2 mixture of Renyi's α-conditional entropies-KEDR-(T2KEDR). Based on the Jensen-Shannon divergence, the KEDR formulation can be also computed as:

$$J(\boldsymbol{Y}) = \gamma S_\alpha(\boldsymbol{K}|\boldsymbol{Z}) + (1-\gamma)S_\alpha(\boldsymbol{L}|\boldsymbol{Z}), \tag{5-15}$$

where $\boldsymbol{Z} \in \mathbb{R}^{N \times N}$ is a positive-definite kernel matrix with elements:

$$z_{nn'} = \gamma k_{nn'} + (1-\gamma)l_{nn'}. \tag{5-16}$$

Rewriting in terms of marginal and joint entropies:

$$\begin{aligned} J(\boldsymbol{Y}) =& \gamma \left(S_\alpha(\boldsymbol{K}, \boldsymbol{Z}) - S_\alpha(\boldsymbol{Z})\right) + (1-\gamma)\left(S_\alpha(\boldsymbol{L}, \boldsymbol{Z}) - S_\alpha(\boldsymbol{Z})\right) \\ =& (1-\gamma) S_\alpha(\boldsymbol{L}, \boldsymbol{Z}) + \gamma S_\alpha(\boldsymbol{K}, \boldsymbol{Z}) - S_\alpha(\boldsymbol{Z}). \end{aligned} \tag{5-17}$$

T2KEDR also looks for a trade-off between sparsity and pair-wise relationships matching as a function of γ. However, the variable change $z_{nn'}$ gives extra information about the matching between $k_{nn'}$ and $l_{nn'}$ as a weighting average between the similarities in HD and LD spaces.

Gradient descend-based optimization of KEDR and variants. By definition, the matrix entropy function presented in eq. (5-4) fall into the family of matrix functions known as spectral functions. Therefore, these functions only depend on the matrix eigenvalues [45]. Accordingly, the derivative of the kernel-based Renyi's α–entropy in eq. (5-4) at $\boldsymbol{A}$ gives:

$$\nabla S_\alpha(\boldsymbol{A}) = \frac{\alpha}{(1-\alpha)\mathrm{tr}\left(\boldsymbol{A}^\alpha\right)} \boldsymbol{A}^{\alpha-1}, \tag{5-18}$$

where $\boldsymbol{A}{=}\boldsymbol{U}\Delta\boldsymbol{U}^*$. Thus, the derivative of both the T1KEDR and the T2KEDR schemes will be studied by keeping in mind that the T1KEDR strategy approximates the KEDR algorithm when $\gamma{=}0$.

Regarding this, the derivative of $J(\boldsymbol{Y})$ in T1KEDR and T2KEDR approaches at $\boldsymbol{y}_n$ yields:

$$\frac{\partial J(\boldsymbol{Y})}{\partial \boldsymbol{y}_n} = \sum_{nn'=1}^{N} \frac{\partial J(\boldsymbol{Y})}{\partial \kappa_Y\left(\boldsymbol{y}_n, \boldsymbol{y}_{n'}\right)} \frac{\partial \kappa_Y\left(\boldsymbol{y}_n, \boldsymbol{y}_{n'}\right)}{\partial\, \mathrm{d}_Y(\boldsymbol{y}_n, \boldsymbol{y}_{n'})} \frac{\partial\, \mathrm{d}_Y(\boldsymbol{y}_n, \boldsymbol{y}_{n'})}{\partial \boldsymbol{y}_n}. \tag{5-19}$$

where $\mathrm{d}_Y(\boldsymbol{y}_n, \boldsymbol{y}_{n'}) = \|\boldsymbol{y}_n - \boldsymbol{y}_{n'}\|_2$.

Based on eq. (5-18), let us define the matrix $\boldsymbol{G}{\in}\mathbb{R}^{N\times N}$ holding elements:

$$g_{nn'} = \frac{\partial J(\boldsymbol{Y})}{\partial \kappa_Y\left(\boldsymbol{y}_n, \boldsymbol{y}_{n'}\right)} \frac{\partial \kappa_Y\left(\boldsymbol{y}_n, \boldsymbol{y}_{n'}\right)}{\partial\, \mathrm{d}_Y(\boldsymbol{y}_n, \boldsymbol{y}_{n'})}. \tag{5-20}$$

In particular, for the T1KEDR framework the $\boldsymbol{G}$ can be calculated as follows:

$$\boldsymbol{G} = \frac{1}{2}\boldsymbol{L} \circ \left(\frac{\nabla S_\alpha(\boldsymbol{L})}{(1-\gamma)^{-1}} - \frac{\boldsymbol{K}}{\mathrm{tr}\left(\boldsymbol{K}\circ\boldsymbol{L}\right)} \circ \nabla S_\alpha\left(\frac{\boldsymbol{K}\circ\boldsymbol{L}}{\mathrm{tr}\left(\boldsymbol{K}\circ\boldsymbol{L}\right)}\right)\right). \tag{5-21}$$

Alike, for the T2KEDR approach:

$$\boldsymbol{G} = \frac{\gamma-1}{2}\boldsymbol{L} \circ \left(\frac{((1-\gamma)\boldsymbol{L}+\boldsymbol{Z})}{\mathrm{tr}\left(\boldsymbol{L}\circ\boldsymbol{Z}\right)} \circ \nabla S_\alpha \left(\frac{\boldsymbol{L}\circ\boldsymbol{Z}}{\mathrm{tr}\left(\boldsymbol{L}\circ\boldsymbol{Z}\right)} \right) + \frac{\gamma\boldsymbol{K}}{\mathrm{tr}\left(\boldsymbol{K}\circ\boldsymbol{Z}\right)} \circ \nabla S_\alpha \left(\frac{\boldsymbol{K}\circ\boldsymbol{Z}}{\mathrm{tr}\left(\boldsymbol{K}\circ\boldsymbol{Z}\right)} \right) - \nabla S_\alpha \left(Z\right) \right),$$

Additionally:

$$\frac{\partial \mathrm{d}_Y(\boldsymbol{y}_n, \boldsymbol{y}_{n'})}{\partial \boldsymbol{y}_n} = 2\left(\boldsymbol{y}_n - \boldsymbol{y}_{n'}\right), \tag{5-22}$$

yielding:

$$\frac{\partial J(\boldsymbol{Y})}{\partial \boldsymbol{y}_n} = 2\sum_{n'=1}^{N} g_{nn'}\left(\boldsymbol{y}_n - \boldsymbol{y}_{n'}\right), \tag{5-23}$$

where $g_{nn'} = g_{n'n}$.

The partial derivative in eq. (5-23) provides a search direction that can be plugged in many gradient-based optimization algorithms. Therefore, a generic update for the LD coordinates, given an initial guess $\boldsymbol{y}_n^{(0)}$, can be written as:

$$\boldsymbol{y}_n^{(t+1)} = \boldsymbol{y}_n^{(t)} - \mu_n^{(t)} \frac{\partial J(\boldsymbol{Y})}{\partial \boldsymbol{y}_n}, \tag{5-24}$$

where $\mu_n^{(t)} \in \mathbb{R}^+$ is a step size. To accelerate convergence, $\mu_n^{(t)}$ should be a quotient of a gain factor divided by the magnitude of the second derivative $\partial^2 J(\boldsymbol{Y})/\partial \boldsymbol{y}_n^2$.

5.3. KEDR as a kernel enhancement of stochastic-based dimensionality reduction

Since the Renyi's α-entropy tends to the Shannon entropy in the limit case when $\alpha \to 1$, the well-known *Stochastic Neighbor Embedding*-(SNE) DR algorithm and its variants can be viewed as particular cases of the introduced KEDR and extensions. Namely, in the SNE technique a shift-invariant softmax HD similarity is defined as:

$$k_{nn'} = \frac{\exp\left(-\mathrm{d}_X(\boldsymbol{x}_n, \boldsymbol{x}_{n'})/(2\sigma_{x_n}^2)\right)}{\sum_{r,r\neq n}\exp\left(-\mathrm{d}_X(\boldsymbol{x}_n, \boldsymbol{x}_r)/(2\sigma_{x_n}^2)\right)}. \tag{5-25}$$

where $\mathrm{d}_X(\boldsymbol{x}_n, \boldsymbol{x}_{n'}) = \|\boldsymbol{x}_n - \boldsymbol{x}_{n'}\|_2$.

Since the similarity in eq. (5-25) is Gaussian, each local bandwidth $\sigma_{x_n} \in \mathbb{R}^+$ can be seen as a soft neighborhood radius. Similarly, for a the LD space $\boldsymbol{Y}$, the corresponding elements of the LD similarity matrix are estimated as follows:

$$l_{nn'} = \frac{\exp\left(-\mathrm{d}_Y(\boldsymbol{y}_n, \boldsymbol{y}_{n'})/(2\sigma_{y_n}^2)\right)}{\sum_{r,r\neq n} \exp\left(-\mathrm{d}_Y(\boldsymbol{y}_n, \boldsymbol{y}_r)/(2\sigma_{y_n}^2)\right)}, \tag{5-26}$$

where $\sigma_{y_n} \in \mathbb{R}^+$.

In a more elaborate version of the SNE algorithm, the *t-distributed* SNE (*t*-SNE), the LD similarities differ from the HD space by introducing an unnormalized probability mass function of a Student t distribution with δ degrees of freedom as follow:

$$l_{nn'} = \frac{\left(1 + \mathrm{d}_Y(\boldsymbol{y}_n, \boldsymbol{y}_{n'})^2/\delta\right)^{-(\delta+1)/2}}{\sum_{r,r\neq n} \left(1 + \mathrm{d}_Y(\boldsymbol{y}_n, \boldsymbol{y}_r)^2/\delta\right)^{-(\delta+1)/2}}. \tag{5-27}$$

As compared to the Gaussian case, the heavier tail of the Student t function the more prominent exponential transformation is induced between both spaces. In other words, the longer the distance in the HD space – the stronger the LD space stretches [86].

Due to the positive and normalization properties of the similarity vectors

$$\boldsymbol{k}_n = \{k_{nn'} : n' \in [1, N]\} \tag{5-28}$$

and

$$\boldsymbol{l}_n = \{l_{nn'} : n' \in [1, N]\}, \tag{5-29}$$

they can be seen as discrete probability distributions. Moreover, the SNE algorithm and most of its variants (SNE, *t*-SNE, *Neighbor Retrieval Visualizer*-(NeRV) [156], and *Jensen Shannon Embedding*-(JSE) [83]) make use of the Kullback-Leibler-based divergences to quantify the mismatch between both similarity vectors. Consequently, their cost function can be generalized in the form:

$$J(\boldsymbol{Y}) = \sum_{n=1}^{N} \xi(\boldsymbol{k}_n) + \psi(\boldsymbol{l}_n; \boldsymbol{k}_n). \tag{5-30}$$

For SNE and *t*-SNE approaches, it holds that:

$$\xi(\boldsymbol{k}_n) = -\hat{H}_S(\boldsymbol{k}_n) \tag{5-31}$$

$$\psi(\boldsymbol{l}_n; \boldsymbol{k}_n) = \hat{H}_S(\boldsymbol{k}_n, \boldsymbol{l}_n) \tag{5-32}$$

where

$$\hat{H}_S(\boldsymbol{k}) = -\sum\nolimits_{n'} k_{n'} \log\left(k_{n'}\right) \tag{5-33}$$

and

$$\hat{H}_S(\boldsymbol{k}, \boldsymbol{l}) = -\sum\nolimits_{n'} k_{n'} \log\left(l_{n'}\right) \tag{5-34}$$

are the empirical estimators of the Shannon's entropy and the Shannon's joint entropy, respectively. Now, regarding the NeRV algorithm, each term of cost function in eq. (5-30) can be rewritten as:

$$\xi(\boldsymbol{k}_n) = -(1-\gamma)\hat{H}_S(\boldsymbol{k}_n) \tag{5-35}$$

$$\psi(\boldsymbol{l}_n; \boldsymbol{k}_n) = (1-\gamma)\hat{H}_S(\boldsymbol{k}_n, \boldsymbol{l}_n) + \gamma\left(\hat{H}_S(\boldsymbol{l}_n, \boldsymbol{k}_n) - \hat{H}_S(\boldsymbol{l}_n)\right), \tag{5-36}$$

Finally, with respect to JSE algorithm, the terms in eq. (5-30) can be rewritten as:

$$\xi(\boldsymbol{k}_n) = -\gamma\hat{H}_S(\boldsymbol{k}_n) \tag{5-37}$$

$$\psi(\boldsymbol{l}_n; \boldsymbol{k}_n) = -(1-\gamma)\hat{H}_S(\boldsymbol{l}_n) + \hat{H}_S(\boldsymbol{z}_n), \tag{5-38}$$

with $\boldsymbol{z}_n = \gamma\boldsymbol{k}_n + (1-\gamma)\boldsymbol{l}_n$.

As seen, the introduced KEDR and variants are close to the SNE-based formulations. Nonetheless, this assertion must be interpreted only in terms of the DR cost functional. In fact, the main difference between SNE variants and the introduced KEDR lies in the nature of the employed HD and LD similarities. As stated above, SNE similarities are shift-invariant functions based on the well-known Gaussian distribution. However, due to its normalization, SNE similarity is not symmetric (see eqs. (5-25) and (5-26)). Therefore, computed matrices can be negative definitive. So, SNE similarity matrices are not well-defined kernels and cannot be directly employed into the KEDR functional.

5.4. Experimental set-up

In order to validate the introduced KEDR as a relevant data representation approach, a dimensionality reduction scheme is studied. Namely, the proposed KEDR and its variants are tested some synthetic and real-world datasets are used.

Three datasets are used in our experiments: *i) Swiss-Roll*: 3D synthetic dataset with 500 samples sharing nonlinear structures, which allows obtaining an input space with D=3 and N=500 (see fig. **5-1**). The goal is to re-embed the two dimensional manifold in a two-dimensional space, however, the challenge here lies in how to cut the manifold in the most

appropriate way to reveal the main nonlinear data structures. *ii) Object Image Library (COIL-100)* [108]: It contains 72 RGB-color images of size 128×128 for 20 objects in PNG format. Pictures are taken while the object is rotated 360 degrees in intervals of 5 degrees. The images are transformed to gray scale, obtaining an input space with P=16384 and N=1440 (see fig. 5.2a). The two most prominent characteristics of this dataset are its very high-dimensionality and the presence of 20 one-dimensional manifolds [21]. *iii) Olivetti faces* [126]: It contains 400 intensity-value pictures of 40 individuals with small variations in view point, large variation in expression, and occasional addition of glasses. The dataset contains 10 images per person of size 112×92 (see fig. 5.2b). So, an input with P=10304 and N=400 is obtained.

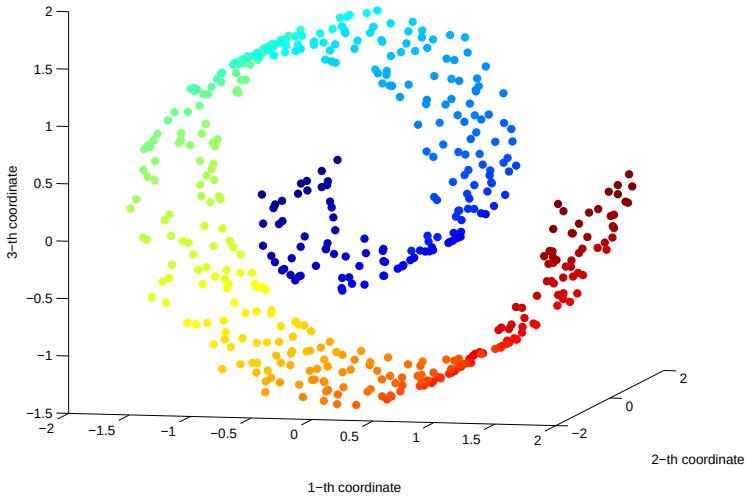


Figure 5-1.: Synthetic dataset (Swiss-Roll).

To assess the quality of the embeddings a criterion that evaluates the preservation of K-ary neighborhoods is employed [85]. In this sense, the rank of $\boldsymbol{x}_n$ with respect to $\boldsymbol{x}_{n'}$ in the HD space is computed as:

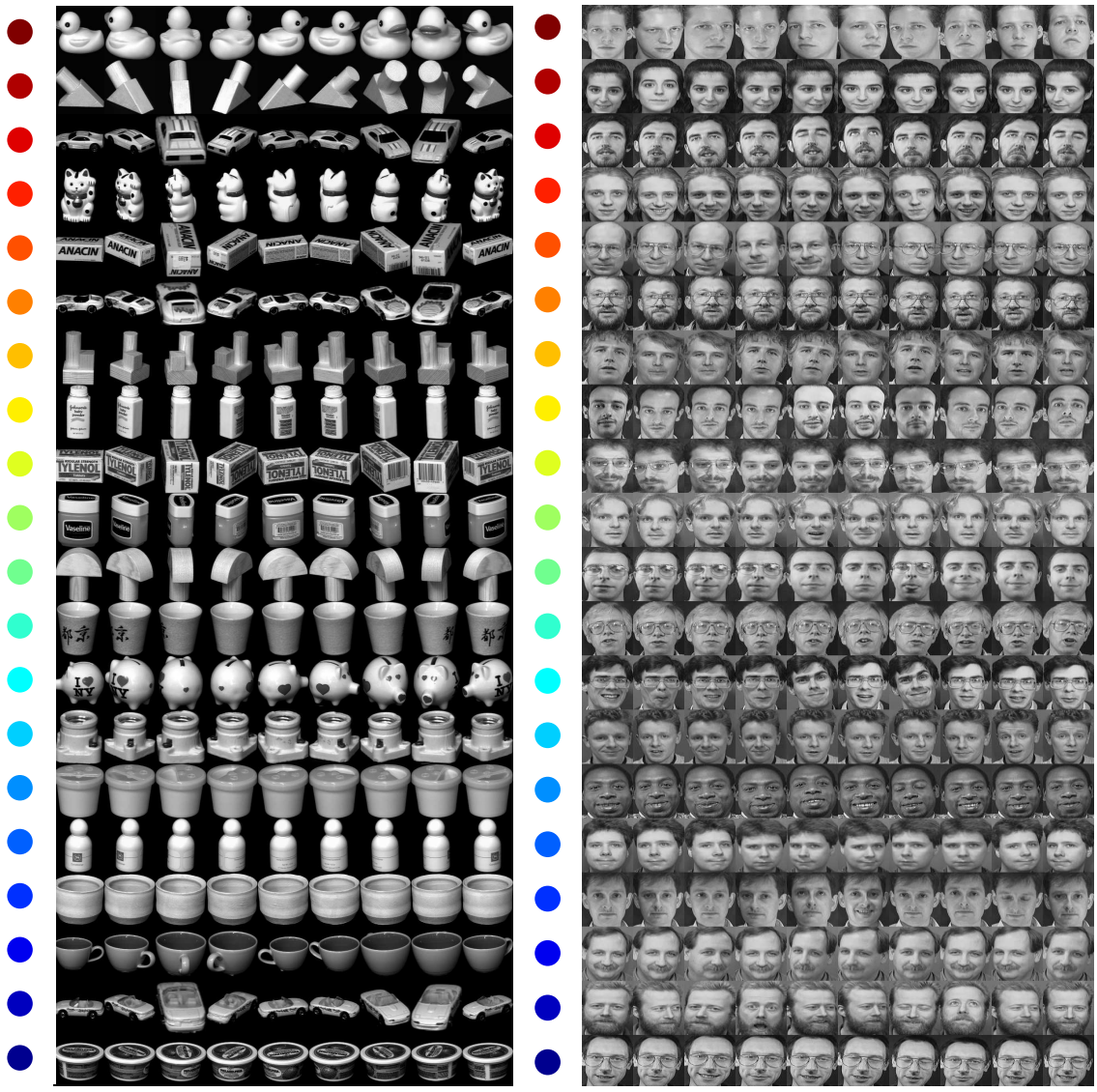
$$\rho_{RANK}\left(\boldsymbol{x}_n, \boldsymbol{x}_{n'}\right) = \left|\left\{r : \tau_{nr} < \tau_{nn'} \text{ or } \left(\tau_{nr}{=}\tau_{nn'} \text{ and } 1 \leq r < n' \leq N\right)\right\}\right|, \tag{5-39}$$

where

$$\tau_{nn'} = \mathrm{d}_X(\boldsymbol{x}_n, \boldsymbol{x}_{n'}) \tag{5-40}$$

and denotes the cardinality operator. Equivalently, the rank of $\boldsymbol{y}_n$ with respect to $\boldsymbol{y}_{n'}$ in the LD space is estimated as:

$$\pi_{nn'}{=}\left|\left\{r : \zeta_{nr} < \zeta_{nn'} \text{ or } \left(\zeta_{nr}{=}\zeta_{nn'} \text{ and } 1 \leq r < n' \leq N\right)\right\}\right|, \tag{5-41}$$



a COIL-100 images b Olivetti faces

Figure 5-2.: Exemplary of the real-world datasets.

where $\zeta_{nn'}$=$\mathrm{d}_Y(\boldsymbol{y}_n, \boldsymbol{y}_{n'})$.

Hence, the K-ary neighborhoods of $\boldsymbol{x}_n$ and $\boldsymbol{y}_n$ can be defined as:

$$\nu_i^K = \{n' : 1 \leq \rho_{RANK}\left(\boldsymbol{x}_n, \boldsymbol{x}_{n'}\right) \leq K\} \tag{5-42}$$

$$\eta_n^K = \{n' : 1 \leq \pi_{nn'} \leq K\}. \tag{5-43}$$

Then, a performance index can be written as:

$$Q_{NX}\left(K\right)=\sum_{n=1}^{N}\frac{\left|\nu_{n}^{K}\bigcap\eta_{n}^{K}\right|}{KN}, \tag{5-44}$$

where $Q_{NX}\left(K\right)\in[0,1]$ measures the average normalized agreement between corresponding K-ary neighborhoods in the HD and LD spaces. Moreover, a normalized criterion with respect to a random embedding can be obtained as:

$$R_{NX}\left(K\right)=\frac{(N-1)Q_{NX}\left(K\right)-K}{N-1-K}, \tag{5-45}$$

for $1\leq K\leq N-2$. In this case, $R_{NX}\left(K\right)=0$ in eq. (5-45) corresponds to a random embedding with $Q_{NX}\left(K\right)\approx K/(N-1)$, whereas $R_{NX}\left(K\right)=1$ means a perfect K-ary neighborhood agreement ($Q_{NX}\left(K\right)=1$).

Eight state-of-the-art DR algorithms are compared in our experiments. The first one is the well-known *Principal Component Analysis*-(PCA), which is equivalent to the Torgerson-Gower classical metric *Multidimensional Scaling*-(MDS) [39]. Namely, the linear projection along the principal directions are found by spectral decomposition of the covariance matrix or the Gram matrix in classical MDS. The second approach is the *Shepard-Kruskal non-metric MDS*-(NMDS) [80], which combines gradient descent and isotonic regression during the optimization procedure. Moreover, the KPCA algorithm is also tested, which aims to find the LD coordinates based on a variability analysis in a RKHS [132]. In addition, the *Laplacian Eigenmaps*-(LEM) technique is employed as comparison method, which is based on preserving the intrinsic geometric structure of the input data by assuming that it can be modeled as a manifold [14]. The last four methods are based on similarity preservation: SNE, t-SNE, NeRV [156], and JSE [83].

With regard to the free parameters setting, the size of the LD matrix $\boldsymbol{Y}$ is fixed as $N\times 2$, $M=2$, for visualization purposes. Moreover, the perplexity value for SNE, t-SNE, NeRV, and JSE algorithms is fixed as $N/20=25$ for the *Swiss-Roll*, $N/20=72$ for the *COIL-100* images (the cluster size is aim at 36), and $N/40=10$ for the *Olivetti faces* (the cluster size is aim at 5), as suggested in [83].

Furthermore, in KEDR, LEM, and KPCA algorithms, the HD kernel bandwidth value is estimated by means of the introduced *Kernel Function Estimation from Information Potential Variability*-(KEIPV) in Chapter 2. In addition, for all considered algorithms based on similarity preservation, the LD kernel bandwidth is fixed to one to constrain the scale of the LD space [67].

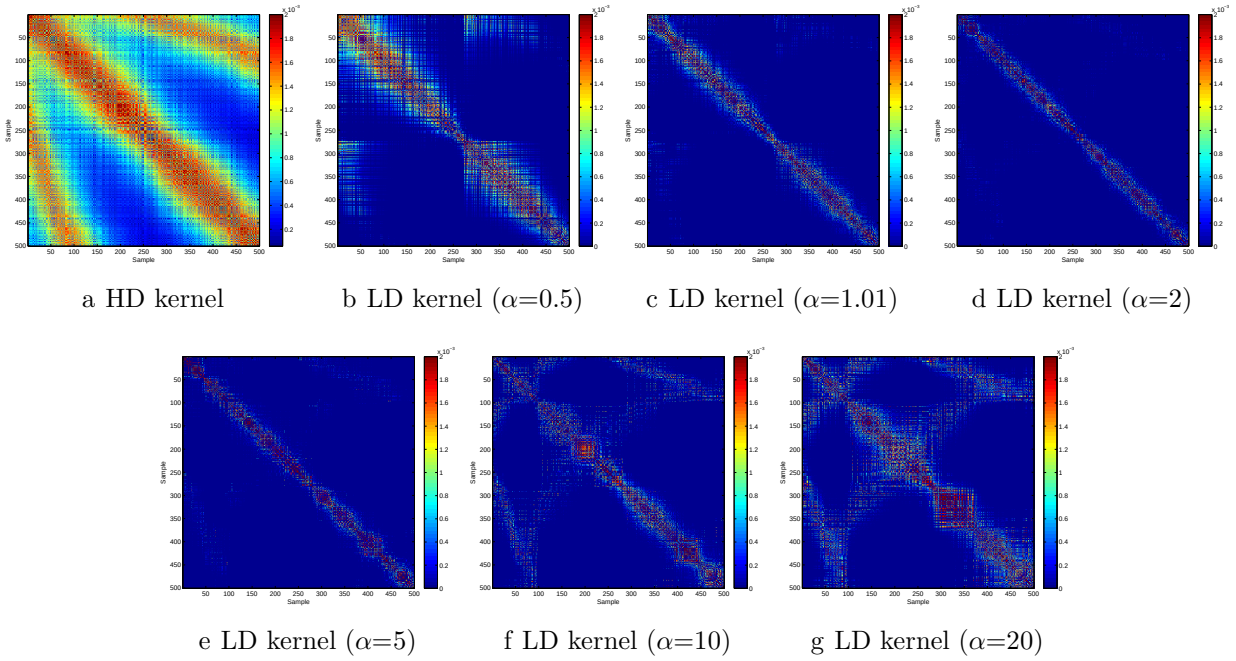
5.5. Results and discussion

First, the KEDR approach is tested over the Swiss-Roll 3D dataset. Our aim is to test the KEDR performance while varying the Renyi's α-entropy order. So, provided the input data matrix $\boldsymbol{X}$, a HD kernel is computed as in eq. (5-10) by fixing the kernel bandwidth value based on the KEIPV stratagy. For concrete testing, different KEDR mappings are computed for each $\alpha\in\{0.5, 1.01, 2, 5, 10, 20\}$. Note that $\alpha \neq 1$ is a KEDR constraint so that we choose a near value to emulate the Shanon's entropy case (α=1.01). figs. **5-3** and **5-4** show the computed HD kernels and the LD embeddings for the Swiss-Roll dataset according to each provided α value, respectively.

As seen in fig. **5-3**, the computed HD kernel matrix describes mainly the global structure of the data, where some high similarities are exhibited among far away points into the Swiss-Roll. Since the HD relationships are estimated in an RKHS spanning the most widely information force magnitudes among data points local data structures are lost. In addition, it can be quoted how KEDR aims to conserve global data properties for high α values. Nonetheless, due to KEDR is based on a geometric averaging formulation, computed kernels tends to accentuate low similarities relationships. Moreover, according to fig. **5-4**, attained LD representations for $\alpha\in\{0.5, 1.01, 2\}$ seem to preserve both the local and the global structures of the Swiss-Roll. In contrast, the higher the α-entropy value more intrusions are exhibited in the LD space due to the employed α-entropy value highlights strong similarities, that is, the RKHS is deformed by collapsing all samples. Then, KEDR will retain local and/or global data structures depending of the employed α value, which can be interpreted as an α-norm in probability space. Aforementioned statements can be corroborated by the rank-based quality assessment results shown in fig. 5.4g. Note how fixing α=1.01 allows finding an embedding that suitable preserves both small and global neighborhoods. Similarly, by setting α=0.5 and α=2, KEDR is able to find acceptable embeddings according to the employed assessment. On the other hand, for $\alpha > 2$, KEDR mappings only preserves large neighborhoods, which can be explained by the fact that the higher the entropy order, the more prone the algorithm is to find unimodal solutions [129].

Following, the T1KEDR approach is tested over the Swiss-Roll dataset with varying the trade-off parameter value γ in eq. (5-13). Again, the HD kernel is computed as based on the introduced KEIPV approach. For concrete testing, the entropy order is fixed as α=1.01 and different T1KEDR mappings are computed for each $\gamma\in\{0, 0.2, 0.4, 0.6, 0.8, 0.99\}$. In fig. **5-5** the obtained T1KEDR LD kernel matrices are presented for the Swiss-Roll dataset. It is remarkable how for high γ values ($\gamma \to 1$), the T1KEDR approach aims to minimize the joint entropy leading to RKHSs where samples are collapsed (see fig. **5-6**). In contrast, when a low γ value is used ($\gamma \to 0$), the T1KEDR approach approximates the KEDR solution, which aims to minimize the joint entropy while maximizing the LD marginal entropy, that is, T1KEDR favors sparse solutions by separating the samples in $\boldsymbol{Y}$.

Aforementioned behavior can be corroborated by the rank-based quality criterion. As



a HD kernel b LD kernel (α=0.5) c LD kernel (α=1.01) d LD kernel (α=2)

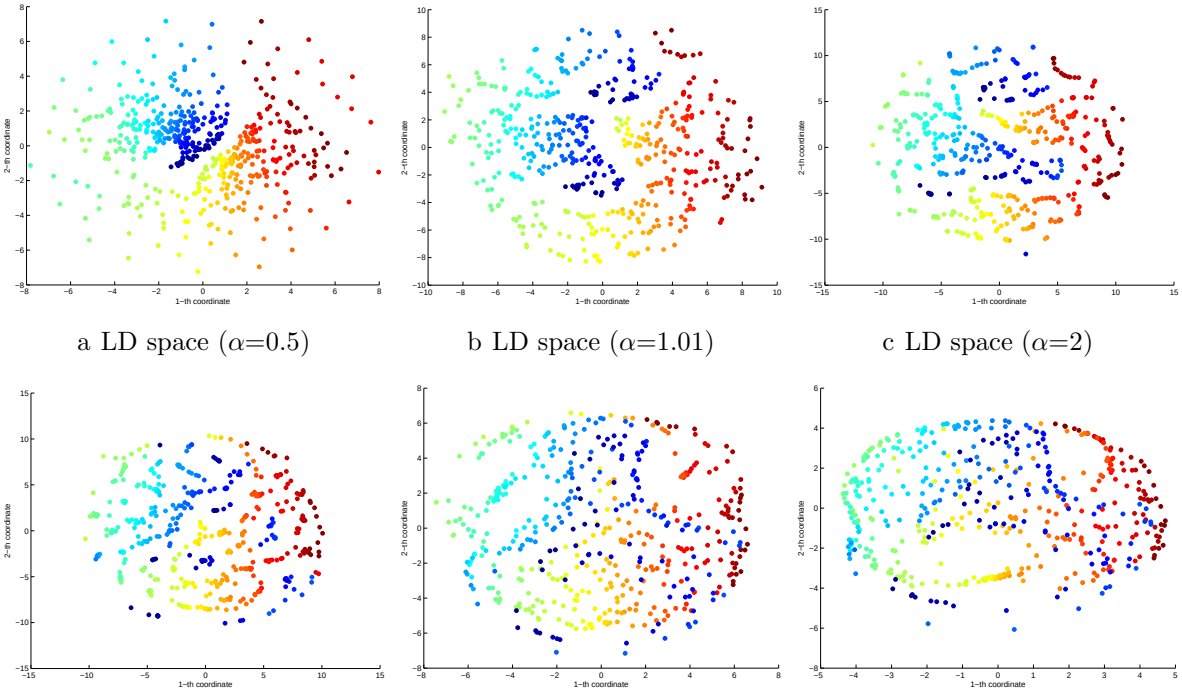
e LD kernel (α=5) f LD kernel (α=10) g LD kernel (α=20)

Figure 5-3.: HD and LD KEDR kernel matrices with varying the Renyi's α-entropy order value (Swiss-Roll dataset).

seen in fig. 5.6g the T1KEDR embedding with low γ values subserve the preservation of small neighborhoods. On the other hand, when $\gamma \to 1$ the computed embeddings exhibit overlapped mappings where the input data structure is poorly preserved.

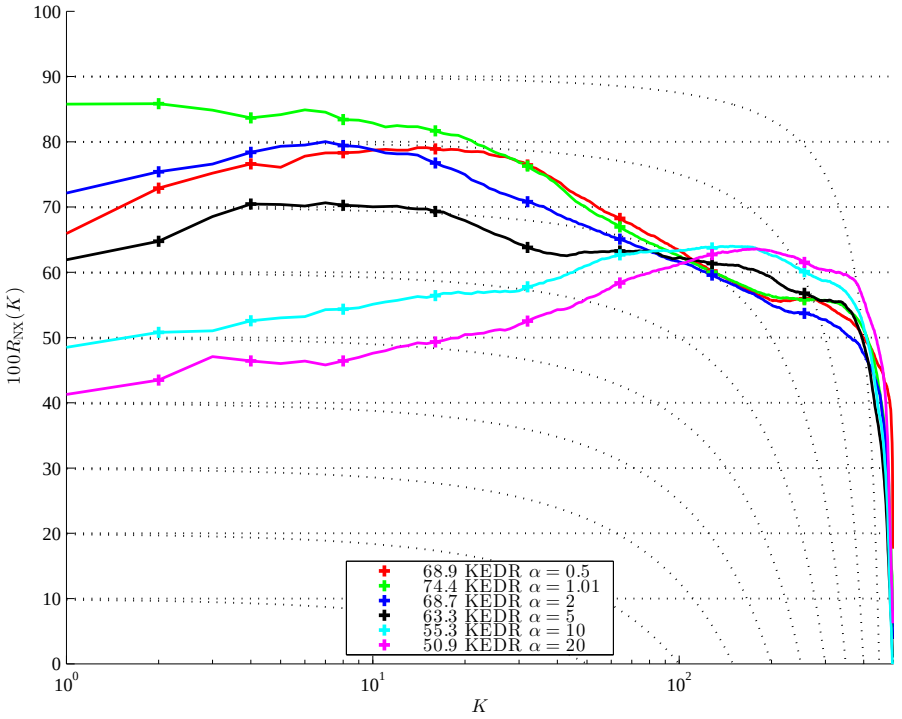
Then, the T2KEDR approach is tested over the Swiss-Roll data with varying the trade-off parameter value $\gamma \in \{0, 0.2, 0.4, 0.6, 0.8, 0.99\}$ in eq. (5-17). The HD kernel is computed as in eq. (5-10) by fixing the kernel bandwidth value based on the KEIPV. For concrete testing, the α-entropy order is fixed as α=1.01. As seen in fig. **5-7**, for $0 \leq \gamma \leq 0.6$ the introduced T2KEDR approach estimates LD spaces where neither the local nor the global data structures are well-preserved. Indeed, samples in closed neighborhoods are collapsed to a unique point in the LD space according to the block structures of the LD kernels (see figs. **5-7** and **5-8**). The latter can be explained by the noise that can encoded the initial guess about $\boldsymbol{Y}$. So, such a bias are propagated in T2KEDR due to the combined representation matrix $\boldsymbol{Z}$. In contrast, for $0.8 \leq \gamma < 1$, the T2KEDR algorithm aims to reveal the main Swiss-Roll structure. Regarding this, when $\gamma \to 1$ the T2KEDR approach avoids the influence of the joint entropy between the LD and the combined data relationships (see eq. (5-17)). Consequently, the noise of the initial guess about $\boldsymbol{Y}$ is diminished during the mapping. Aforementioned results can be corroborated by the rank-based quality criterion presented in fig. 5.8g.

In turn, all considered DR methods and all studied databases are tested. Figure fig. **5-9** presents the 2D embeddings of the Swiss-Roll dataset. A quick glance shows that PCA,



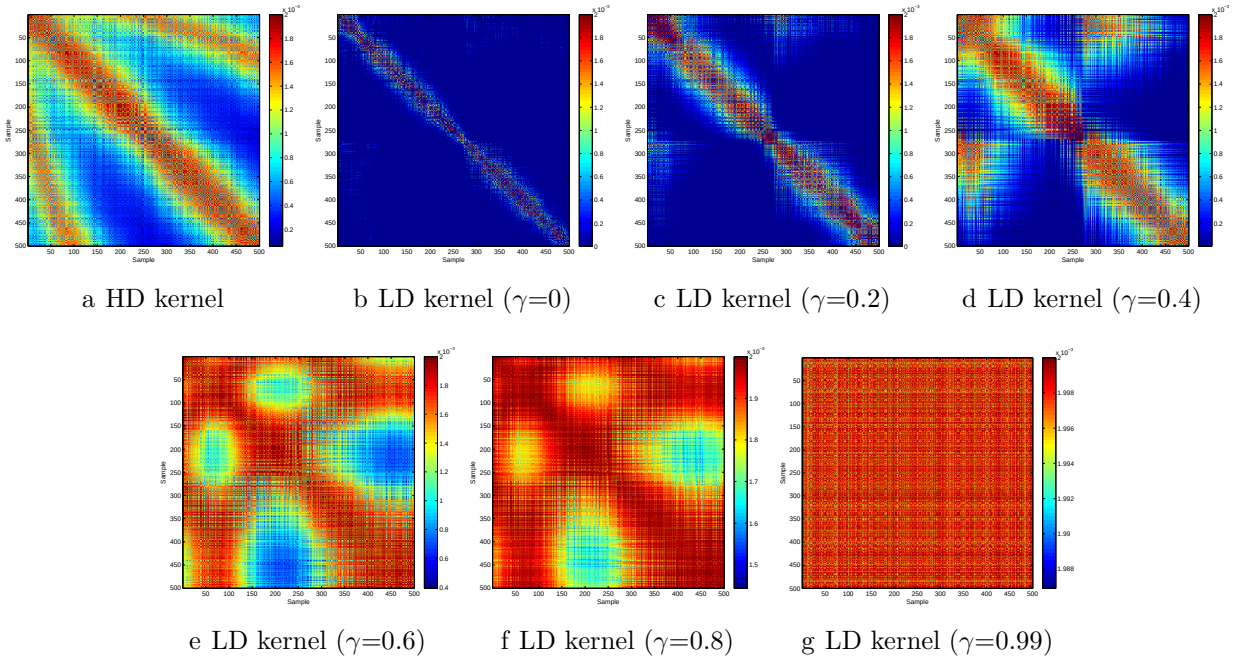
a LD space (α=0.5) b LD space (α=1.01) c LD space (α=2)

d LD space (α=5) e LD space (α=10) f LD space (α=20)



g Rank-based quality assessment

Figure 5-4.: KEDR Swiss-Roll 3D results with varying the Renyi's α-entropy order value.

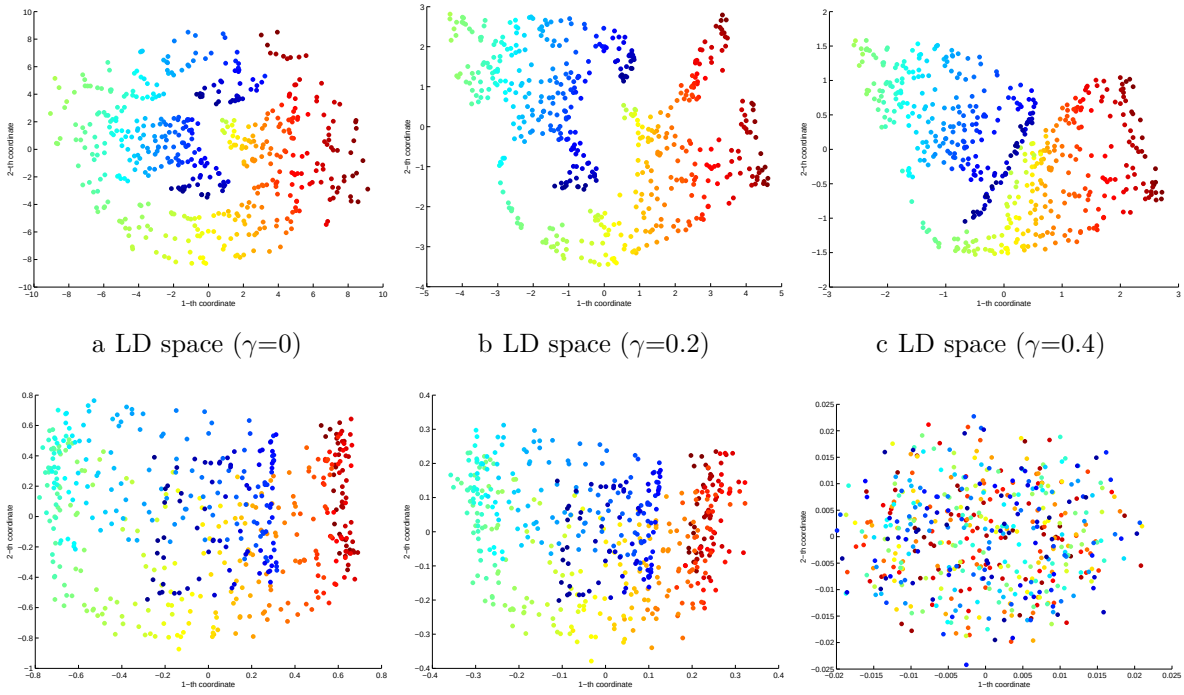


a HD kernel b LD kernel (γ=0) c LD kernel (γ=0.2) d LD kernel (γ=0.4)

e LD kernel (γ=0.6) f LD kernel (γ=0.8) g LD kernel (γ=0.99)

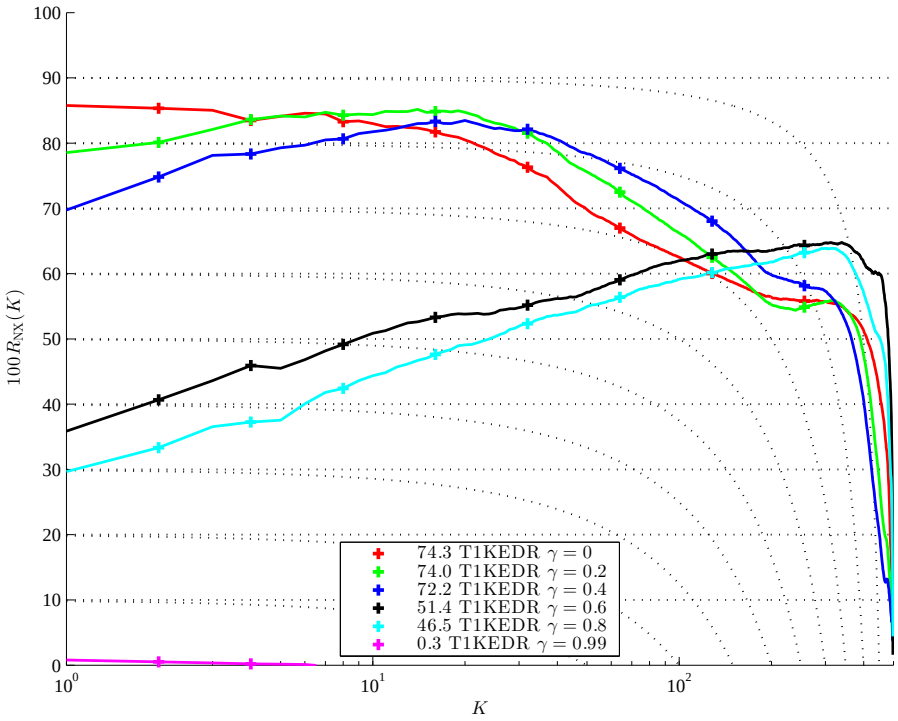
Figure 5-5.: HD and LD T1KEDR kernel matrices with varying the trade-off parameter value γ (Swiss-Roll dataset).

NMDS, and KPCA squash the Swiss-Roll onto a plane. Above behavior can be explained by the norm and the distance concentration drawbacks of the spectral methods. Consequently, those methods are not able to conserve small neighborhoods according to the employed rank-based measure (see fig. 5.9l). In turn, LEM, t-SNE, and JSE approaches aim to cut the Swiss-Roll and unfold it. Nonetheless, LEM embedding does not suitable conserve neither the local nor the global neighborhoods as seen in fig. 5.9l, which can be explained again by the norm concentration issue for the spectral methods. With regard to the t-SNE and the JSE results, it can be quoted how these approaches favor significantly the preservation of local neighborhoods in terms of the employed rank-based assessment. However, the global relationships among samples are not preserved, computing LD spaces where the main structure of the Swiss-Roll is lost (see fig. 5.9l). Indeed, t-SNE stretches the distances in LD because of its Student t similarity, highlighting mainly local data structures. Likewise, JSE stands out local similarities based on its mixture of divergences that favors sparse relationships matching. Since the rank-based assessment is computed in log scale not preserving small neighborhoods is stronger penalized than not preserving the global ones. Now, with respect to the SNE result it can be noted how it tries to conserve both the local and global neighborhoods, however, the attained projection exhibits overlapped samples which deteriorates the embedding quality. That is, as seen in fig. 5.9l, the SNE algorithm does not reveal properly both the local and the global data properties due to the lack of a suitable framework that allows to regularize the neighborhood preservation when unfolding the man-



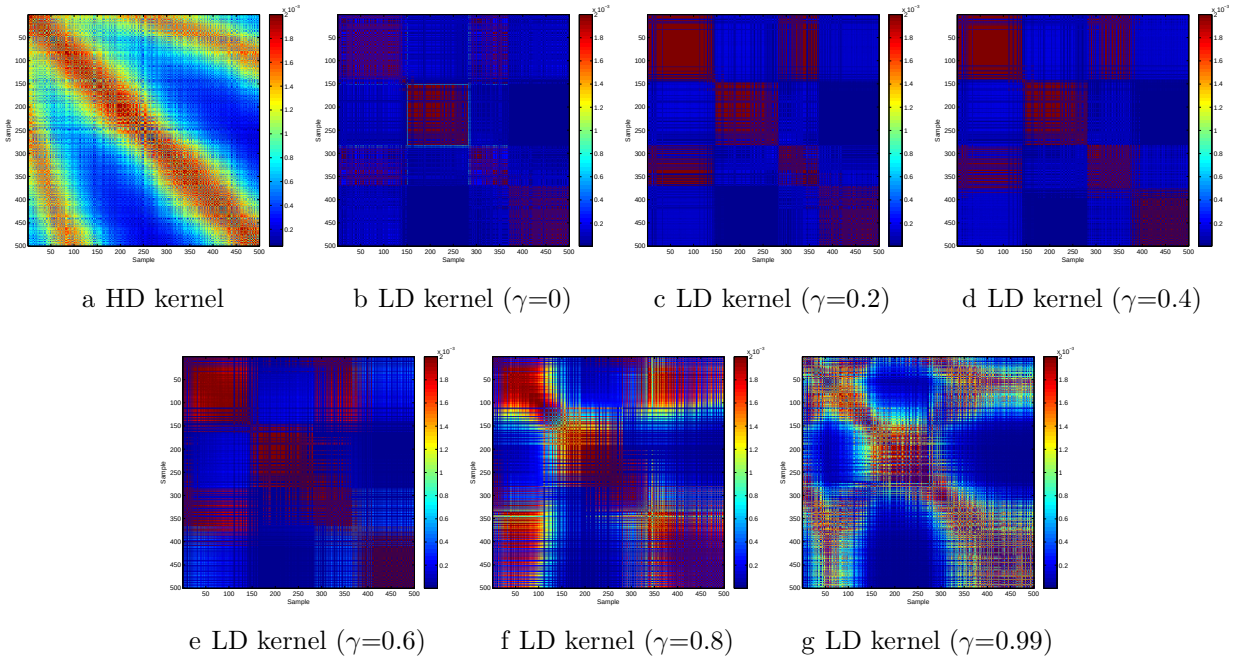
a LD space (γ=0) b LD space (γ=0.2) c LD space (γ=0.4)

d LD space (γ=0.6) e LD space (γ=0.8) f LD space (γ=0.99)



g Rank-based quality assessment

Figure 5-6.: T1KEDR Swiss-Roll results with varying the trade-off parameter value γ.



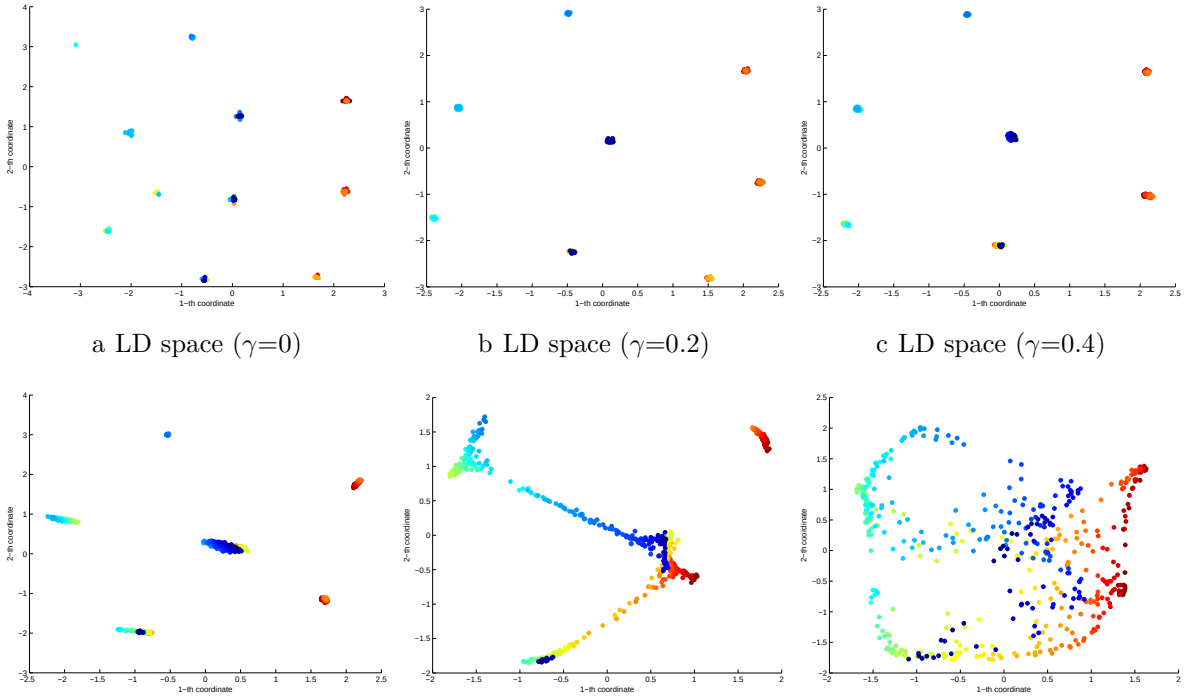
a HD kernel b LD kernel (γ=0) c LD kernel (γ=0.2) d LD kernel (γ=0.4)

e LD kernel (γ=0.6) f LD kernel (γ=0.8) g LD kernel (γ=0.99)

Figure 5-7.: HD and LD T2KEDR kernel matrices with varying the trade-off parameter value γ (Swiss-Roll dataset).

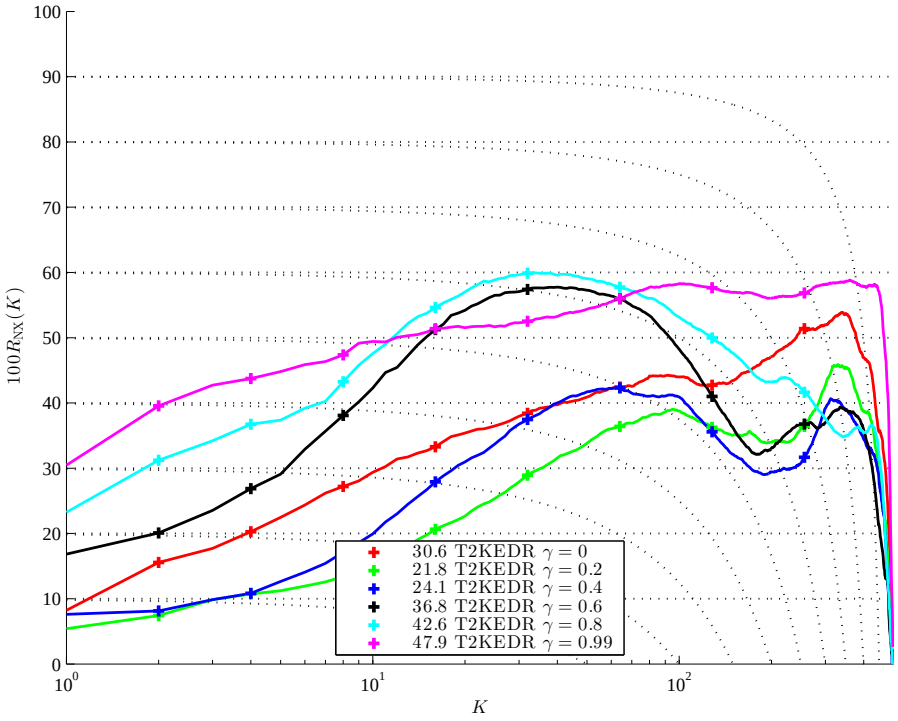
ifold. Regarding the KEDR, the T1KEDR, and the T2KEDR embeddings it is possible to notice how these approaches are able to unfold the manifold by preserving both the local and the global structures. Indeed, the KEDR and the T1KEDR algorithms achieve the highest rank-based quality values.

Now, with respect to the Olivetti and COIL-20 images results, it is important to note that they are much more difficult problems than the Swiss-Roll due to the very high-dimensionality of the vectorized images and the presence of clusters. Again, spectral-based embeddings, i.e., PCA, NMDS, KPCA, and LEM, are not able to reveal the main structures of the input data as seen in figs. 5.10a to 5.10c, and fig. 5.10d, and figs. 5.11a to 5.11c, and fig. 5.11d. In fact, cluttered LD embeddings are calculated in most of the cases where different clusters are overlapped each other. In turn, SNE-based algorithms obtain better embeddings than the spectral-based approaches. Thereby, the t-SNE and JSE results preserves the local data structures encoded in the different clusters, while the SNE and the NeRV embeddings are more prominent to overlap different clusters (see figs. 5.10e to 5.10g, and fig. 5.10h and figs. 5.11e to 5.11g, and fig. 5.11h). Finally, the introduced KEDR-based approaches achieve competitive results in terms of visual inspection and rank-based quality (see figs. 5.10i to 5.10k, and figs. 5.11i to 5.11k). In particular, KEDR and T1KEDR are able find LD spaces where both the local cluster structure and the relationships among different clusters are preserved. In this sense, the proposed T1KEDR can lead with the inter-cluster relationships by fixing a suitable trade-off parameter value that encoded a compromise be-



a LD space (γ=0) b LD space (γ=0.2) c LD space (γ=0.4)

d LD space (γ=0.6) e LD space (γ=0.8) f LD space (γ=0.99)



g Rank-based quality assessment

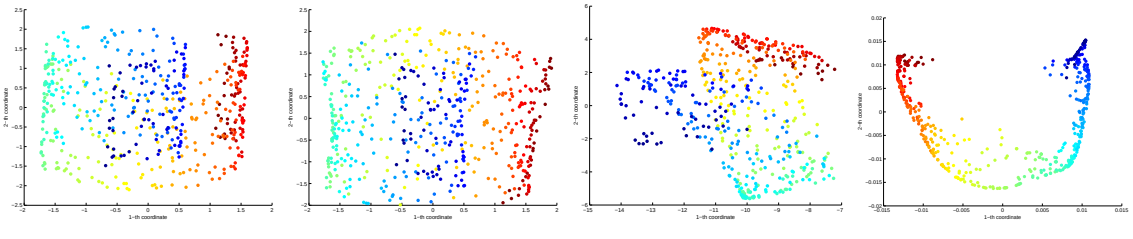
Figure 5-8.: T2KEDR Swiss-Roll results with varying the trade-off parameter value γ.

tween sparsity (local structures) and kernel-based similarity matching between HD and LD spaces. Overall, SNE-based and KEDR-based DR approaches are able to achieve competitive performance in terms of the rank-based criteria as seen in figs. 5.10l and 5.11l.

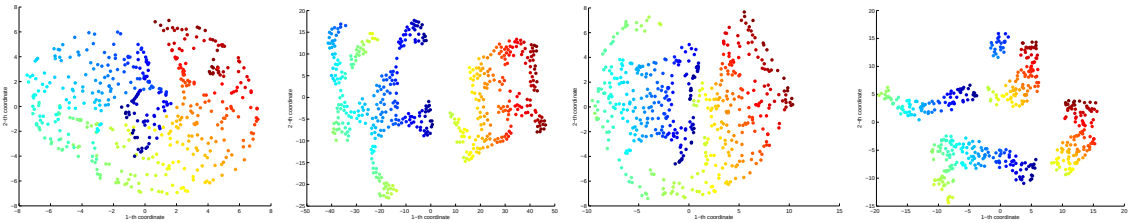
5.6. Summary

We introduce a kernel-based representation strategy that consider both the statistical distribution and the salient data structures form an ITL-based functional for Gramm matrices. Namely, our approach, termed KEDR, is based on a Gram matrix estimation of Renyi's α-entropy. The introduced strategy is a data-driven framework for ITL based on infinitely divisible matrices. In this sense, we employ estimators of entropy-like quantities for Gram matrices that can be computed by evaluating infinitely divisible kernels on pairs of samples to measure the DR mismatch. Our approach do not assume that the density of the data has been estimated, which can be advantageous to develop flexible DR approaches where even defining a density is not feasible. Furthermore, the proposed KEDR is extended by switching from Renyi's-based entropies to parameterized mixtures of divergences aiming to improve the preservation of both the local and the global data structures. KEDR is tested as a representation tool to support DR tasks. Thus, provided scheme provides a flexible alternative to deal with complex structures from data driven estimations of ITL for Gram matrices. Moreover, the main relations between the well-known SNE algorithm and the introduced KEDR are presented. Regarding this, our approach can be viewed as a generalized version of the SNE algorithm and variants from an ITL prospective in terms of the DR mismatch cost functional.

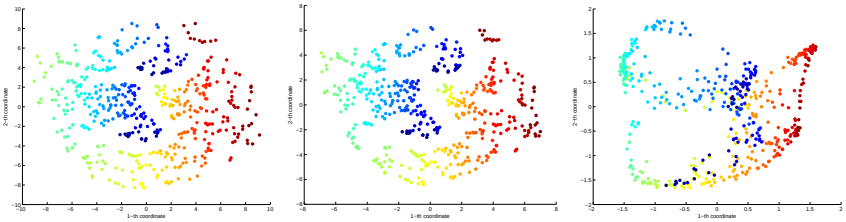
Our DR proposal is tested on both synthetic and real-world datasets. Several state-of-the-art algorithms are employed as baselines, including spectral methods and SNE-based techniques, and the DR performances are validated in terms of both visual inspection and neighborhood preservation (rank-based criteria). Overall, the introduced approach is competitive in comparison to the state-of-the-art methods, being able to encode HD data relationships by computing LD representations where the local and the global structures are preserved.



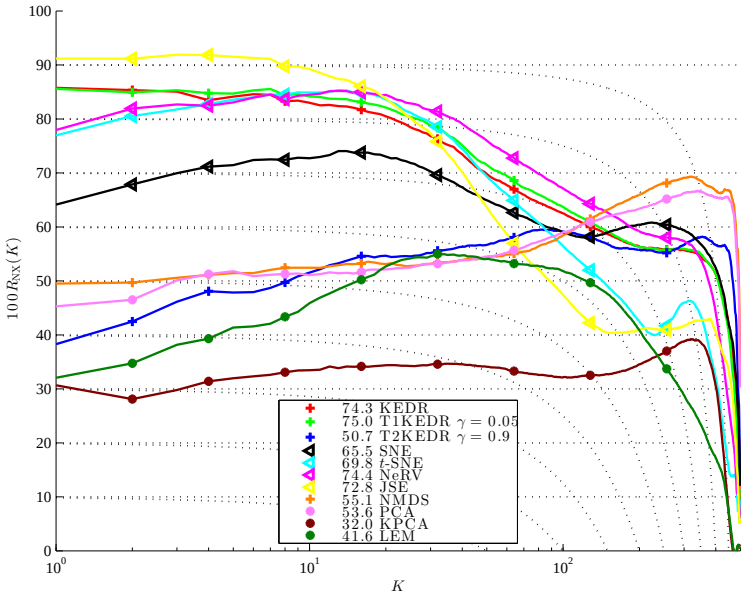
a PCA b NMDS c KPCA d LEM



e SNE f t-SNE g NeRV h JSE

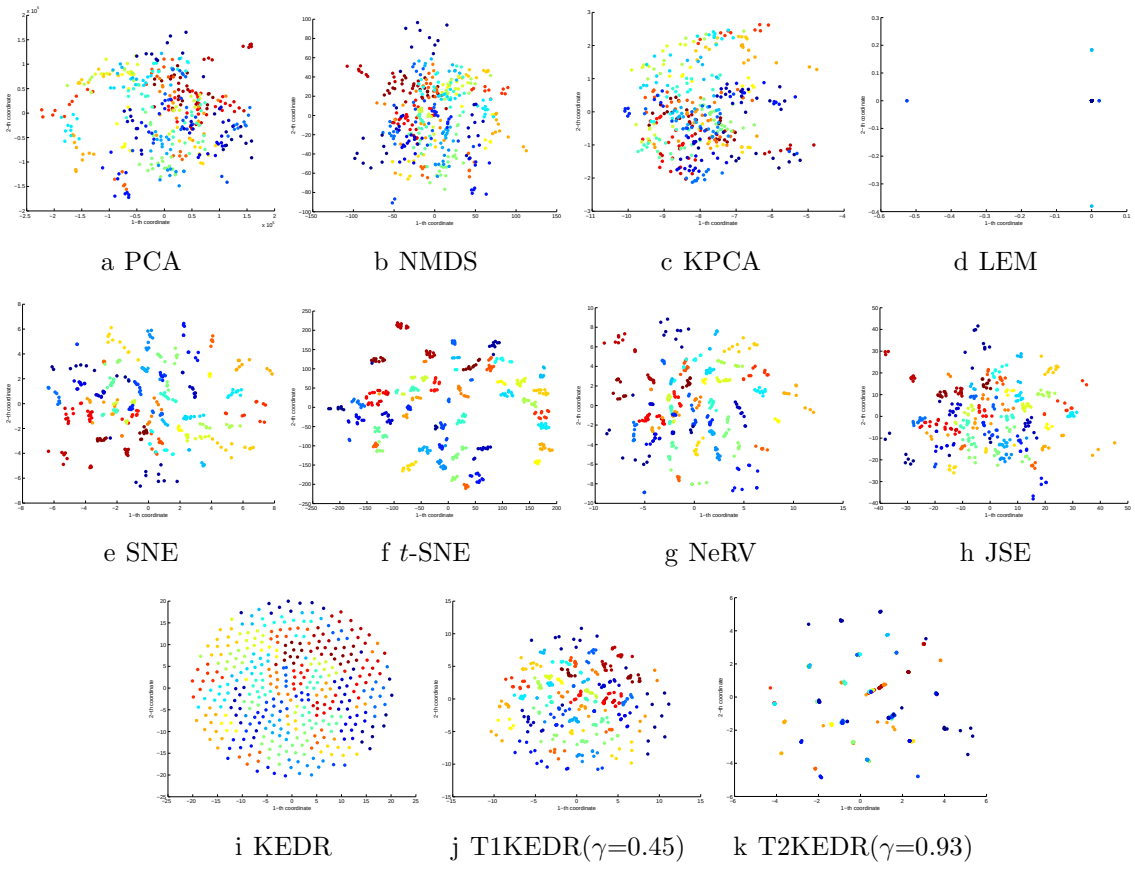


i KEDR j T1KEDR(γ=0.05) k T2KEDR(γ=0.9)



l Rank-based quality assessment

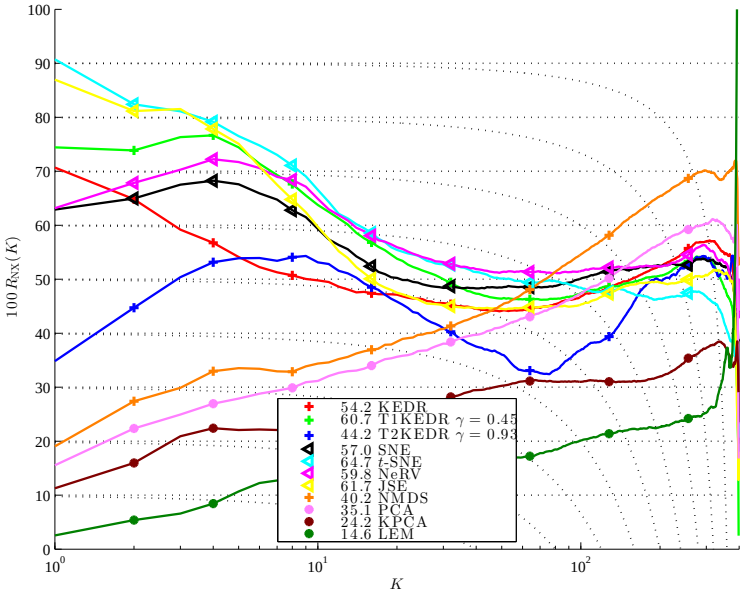
Figure 5-9.: Swiss-Roll embedding results (all methods).



a PCA b NMDS c KPCA d LEM

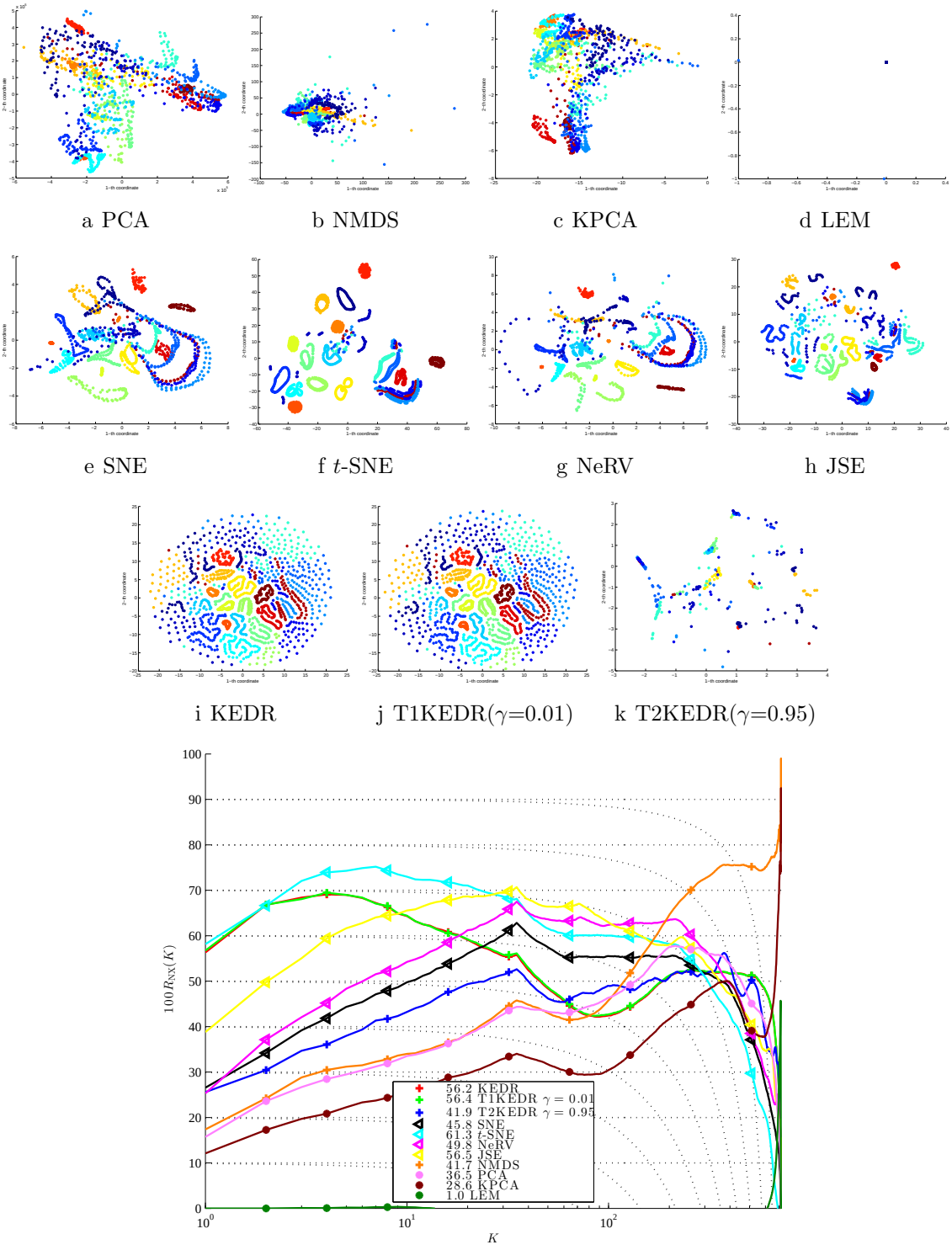
e SNE f t-SNE g NeRV h JSE

i KEDR j T1KEDR(γ=0.45) k T2KEDR(γ=0.93)



l Rank-based quality assessment

Figure 5-10.: Olivetti embedding results (all methods).



a PCA b NMDS c KPCA d LEM

e SNE f t-SNE g NeRV h JSE

i KEDR j T1KEDR(γ=0.01) k T2KEDR(γ=0.95)

l Rank-based quality assessment

Figure 5-11.: Coil-100 embedding results (all methods).

Part III.

Supervised kernel-based representation approaches

6. Kernel-based data representation incorporating prior user knowledge: a supervised relevance analysis strategy

The choice of relevant features is one of the key steps in the design of learning algorithms. In most of the cases, such a choice is typically left to the user and represents his prior knowledge. For this, a poor choice makes learning challenging while a better choice makes it more likely to be successful. Furthermore, the input feature space provides a huge number of features and a limited number of samples, being difficult to the user to choice a relevant representation that encodes the studied phenomenon. Hence, reduction dimension appears as an important stage to dealing with large dimensions.

Generally speaking, reduction dimension can be divided in two main strategies: feature selection and feature embedding [39]. Feature selection aims to determine the most compact set of relevant input features that encode the main information of the studied phenomenon so that further distinction of data patterns can be performed with suitable accuracy. Since the selected features are not transformed from the input space, the assessed economical representation keeps the meaningful feature sense and favors the interpretation of the underlying application. In practice, the majority of the feature selection methods assumes the interactions between features, usually, through an introduced distance. The relationship complexity ranges from the basic principal component analysis [164], Fisher Criterion [113], Mutual Information [163], to algorithms employing weighted distances adapted by learning [73]. On the other hand, heuristic search strategies have been increasingly used. Nonetheless, such approaches avoid conjectures about the feature interactions and evaluate sets of solutions simultaneously. Also, they are not prone to getting stuck in local minima [123].

Despite the extensive research on feature selection, the following issues remain for identification of relevant patterns: *i)* Most of the selected feature sets with the smallest size suffer from low accuracy, resulting in a high rate of false alarms and missed detections. This situation hinders a solid interpretation of the mechanisms underlying the problem [47]. *ii*) Their computational burden is a strong constraint due to the huge processing time and the necessity of parameter tuning (mainly in heuristic methods). In fact, there is a need for identifying the most discriminating features by finding a trade-off between system complexity and accuracy [15].

One of the approaches to overcoming the difficulties mentioned above is to have a low-

dimensional space where it is easier to gauge the relevant information content according to the experimental condition. So, the optimized representation of the inputs is useful for visualizing the similarity between input feature vectors to similar conditions for both unsupervised and supervised scenarios [30, 20]. For this exploratory analysis, linear and nonlinear strategies can be employed. The former approaches extract the relevant information from sample covariances, but regardless of including their unsupervised or supervised estimations they lead to unsatisfactory results [33]. Instead, the latter approaches aim to preserve the similarity structure between samples in a low-dimensional representation through more elaborate approaches of reduction dimension like embedding, kernel analysis, or manifold learning. Nonetheless, the adaptation of the nonlinear strategies to the complex data relationships is far from being easy, especially, when taking advantage of the supervised information. Moreover, a direct interpretability from a nonlinear-based mapping is not always possible.

In this study, we propose a supervised kernel-based approach of feature relevance analysis (termed supervised data representation based on kernel alignment (SKRA)) to enhance the automatic identification of relevant input patterns. SKRA incorporates two kernel functions to take advantage of the actual joint information associating the available labels to the corresponding input samples to a certain condition/label. Particularly, we employ the Centered Kernel Alignment (CKA) strategy to learn a linear projection encoding discriminative input features, capitalizing the nonlinear notion of similarity behind the studied kernels [32]. Also, an iterative gradient descent optimization is introduced to compute both the SRKA projection matrix and the required kernel free parameters. In this sense, the kernel free parameters are fixed based on the introduce KEIVP strategy. Therefore, SRKA can be carried out carried out either as feature selection or a feature embedding tool. As a result, we provide a feature relevance analysis strategy that allows enhancing the system performance while favoring the data interpretability. The proposed SRKA is validated on two well-known tasks: motor imagery discrimination and epileptic seizure detection. Attained results show that SRKA allows finding a relevant feature representation space by ensuring a suitable identification of brain activity patterns while favoring the physiological interpretation of the studied phenomenon.

6.1. Supervised data representation based on kernel alignment (SRKA)

Kernel functions are bivariate measures of similarity based on the inner product between samples embedded in a Hilbert space. For a given domain $\mathscr{X}$ containing the input feature estimation of a given machine learning task, a kernel $\kappa_X{:}\mathscr{X}\times\mathscr{X}\to\mathbb{R}$ is assumed to be a positive-definite function, which defines an implicit mapping $\phi_X{:}\mathscr{X}\to\mathscr{H}_X$ that embeds any element $x\in\mathscr{X}$ into the element $\phi_X(x)\in\mathscr{H}_X$ of some RKHS noted as $\mathscr{H}_X$.

Also, we set a positive definite kernel $\kappa_L{:}\mathscr{L}\times\mathscr{L}\mapsto$over a target space $\mathscr{L}$ related to the user

prior knowledge, e.g., the label space. Then, the RKHS $\mathscr{H}_L$ defines the implicit mapping $\phi_L{:}\mathscr{L}\mapsto\mathscr{H}_L$, which maps any element $l{\in}\mathscr{L}$ into the element $\phi_L(l){\in}\mathscr{H}_L$.

As a consequence, we apply two kernel-based functions sequentially to assess the joint information between the input feature space to a certain target and the corresponding labels, where each kernel reflects different notion of similarity. Therefore, we must still evaluate how well the kernel function, κ_X, aligns to the target kernel of targets, κ_L. To this end, we employ a kernel target alignment to measure the similarity between the couple of characterizing kernel functions. That is, we employ the inner product of both kernel functions to estimate the dependence between jointly sampled data [57]. Thus, the statistical alignment between κ_X and κ_L is computed from the expected value of their normalized inner product across all pairs of realizations, termed the Centered Kernel Alignment (CKA) [32]:

$$\rho_{CKA}\left(\kappa_X,\kappa_L\right)=\frac{\mathbb{E}_{xx',l,l'}\left\{\bar{\kappa}_X\left(x,x'\right)\bar{\kappa}_L\left(l,l'\right)\right\}}{\sqrt{\mathbb{E}_{xx'}\left\{\bar{\kappa}_X^2\left(x,x'\right)\right\}\mathbb{E}_{ll'}\left\{\bar{\kappa}_L^2\left(l,l'\right)\right\}}},\tag{6-1}$$

where the centered versions of $\kappa_X\left(x,x'\right)$ and $\kappa_L\left(l,l'\right)$ are estimated as follows, respectively:

$$\bar{\kappa}_X\left(x,x'\right)=\kappa_X\left(x,x'\right)-\mathbb{E}_{x'}\left\{\kappa_X\left(x,x'\right)\right\}-\mathbb{E}_{x}\left\{\kappa_X\left(x,x'\right)\right\}+\mathbb{E}_{xx'}\left\{\kappa_X\left(x,x'\right)\right\},\tag{6-2a}$$

$$\bar{\kappa}_L\left(l,l'\right)=\kappa_L\left(l,l'\right)-\mathbb{E}_{l'}\left\{\kappa_L\left(ll'\right)\right\}-\mathbb{E}_{l}\left\{\kappa_L\left(l,l'\right)\right\}+\mathbb{E}_{ll'}\left\{\kappa_L\left(l,l'\right)\right\}.\tag{6-2b}$$

Notation $\mathbb{E}_x\left\{\cdot\right\}$ stands for the expected value operator calculated over the random variable $x{\subset}\mathscr{X}$. Therefore, $\rho_{CKA}{\in}[0,1]$ is an estimate of the statistical dependence between $\mathscr{X}$ and $\mathscr{L}$ spaces. So, the larger the similar pairs between interspace variables, the higher the value of CKA.

In practice, we are given an input representation set $\boldsymbol{X}{\in}\mathbb{R}^{N\times P}$ ($\mathscr{X}{\subset}\mathbb{R}^P$) and a target sample vector, e.g., label, $\boldsymbol{l}{\in}\mathbb{Z}^N$ ($\mathscr{L}{\subset}\mathbb{Z}$), from which we extract the characterizing kernel matrices: $\boldsymbol{K}_{\boldsymbol{X}}{\in}\mathbb{R}^{N\times N}$ and $\boldsymbol{K}_{\boldsymbol{l}}{\in}\mathbb{R}^{N\times N}$, respectively. The former matrix holds elements:

$$k^{\boldsymbol{X}}_{nn'}=\kappa_X\left(\boldsymbol{x}_n,\boldsymbol{x}'_n\right)\tag{6-3}$$

with $\boldsymbol{x}_n,\boldsymbol{x}_{n'}{\in}\boldsymbol{X}$ and the latter matrix has elements:

$$k^{\boldsymbol{l}}_{nn'}=\kappa_L\left(l_n,l'_n\right),\tag{6-4}$$

with $l_n,l_{n'}{\in}\boldsymbol{l}$ ($n,n'{\in}[1,N]$). Hence, the empirical estimate of the CKA is computed in accordance to [19]:

$$\hat{\rho}_{CKA}\left(\bar{\boldsymbol{K}}_{\boldsymbol{X}},\bar{\boldsymbol{K}}_{\boldsymbol{l}}\right)=\frac{\langle\bar{\boldsymbol{K}}_{\boldsymbol{X}},\bar{\boldsymbol{K}}_{\boldsymbol{l}}\rangle_{\mathtt{F}}}{\sqrt{\langle\bar{\boldsymbol{K}}_{\boldsymbol{X}},\bar{\boldsymbol{K}}_{\boldsymbol{X}}\rangle_{\mathtt{F}}\langle\bar{\boldsymbol{K}}_{\boldsymbol{l}},\bar{\boldsymbol{K}}_{\boldsymbol{l}}\rangle_{\mathtt{F}}}},\tag{6-5}$$

where $\langle\cdot,\cdot\rangle_{\mathbf{F}}$ is the matrix-based Frobenius norm. Notation $\bar{\boldsymbol{K}}$ stands for the centered versions of the kernel matrix $\boldsymbol{K}$ calculated as $\bar{\boldsymbol{K}}{=}\tilde{\boldsymbol{I}}\boldsymbol{K}\tilde{\boldsymbol{I}}$. $\tilde{\boldsymbol{I}}{=}\boldsymbol{I}{-}\mathbf{1}^\top\mathbf{1}/N$ is the empirical centering matrix, $\boldsymbol{I}{\in}\mathbb{R}^{N\times N}$ is the identity matrix, and $\mathbf{1}{\in}\mathbb{R}^N$ is the all-ones vector.

Further, we rely on the Mahalanobis distance to carry out the pairwise comparison between samples $\boldsymbol{x}_n$ and $\boldsymbol{x}_{n'}$ by fixing κ_X as a Gaussian kernel. Namely, the distance function in $\mathscr{X}$ is fixed as:

$$\mathrm{d}_A^2(\boldsymbol{x}_n,\boldsymbol{x}_{n'}) = (\boldsymbol{x}_n-\boldsymbol{x}_{n'})\,\boldsymbol{A}\boldsymbol{A}^\top\,(\boldsymbol{x}_n-\boldsymbol{x}_{n'})^\top \tag{6-6}$$

where matrix $\boldsymbol{A}{\in}\mathbb{R}^{P\times M}$ holds the linear projection:

$$\boldsymbol{y}_n = \boldsymbol{x}_n\boldsymbol{A}, \tag{6-7}$$

with $\boldsymbol{y}_n{\in}\mathbb{R}^M$, $M{\leq}P$, and $\boldsymbol{A}\boldsymbol{A}^\top$ is the corresponding inverse covariance matrix of the introduced Mahalanobis distance in the input feature space.

To compute the projection matrix $\boldsymbol{A}$, the formulation of the CKA-based function in eq. (6-5) can be integrated into the following kernel-based learner:

$$\hat{\boldsymbol{A}} = \arg\max_{\boldsymbol{A}} \log\left(\hat{\rho}_{CKA}\left(\bar{\boldsymbol{K}}_{\boldsymbol{X}},\bar{\boldsymbol{K}}_{\boldsymbol{l}};\boldsymbol{A}\right)\right), \tag{6-8}$$

where the logarithm function is used for mathematical convenience. fig. **6-1** describes the introduced SRKA from an RKHS perspective.

Gradient descend-based optimization of SRKA. The explicit objective function of the empirical CKA in eq. (6-5) yields [19]:

$$\hat{\rho}_{CKA}\left(\boldsymbol{K}_{\boldsymbol{X}},\boldsymbol{K}_{\boldsymbol{l}}\right) = \log\left(\mathrm{tr}\left(\boldsymbol{K}_{\boldsymbol{X}}(\boldsymbol{A},\sigma)\tilde{\boldsymbol{I}}\boldsymbol{K}_{\boldsymbol{l}}\tilde{\boldsymbol{I}}\right)\right) - \tfrac{1}{2}\log\left(\mathrm{tr}\left(\boldsymbol{K}_{\boldsymbol{X}}(\boldsymbol{A},\sigma)\tilde{\boldsymbol{I}}\boldsymbol{K}_{\boldsymbol{X}}(\boldsymbol{A},\sigma)\tilde{\boldsymbol{I}}\right)\right) + \rho_0, \tag{6-9}$$

where $\rho_0{\in}\mathbb{R}$ is a constant that we assume does not depend on $\boldsymbol{A}$.

Consequently, the optimizing approach in eq. (6-8), besides learning the optimal projection matrix $\hat{\boldsymbol{A}}$, also demands tuning of the Gaussian kernel bandwidth σ_X. To deal with the joint parameter estimation, we propose to optimize iteratively one variable at a time while the other variable is fixed. Moreover, we employ the gradient descent approach to solve iteratively the optimizing KRA task.

In terms of the maximizing parameter $\boldsymbol{A}$ and fixed σ_X, the gradient function of the objective function in eq. (6-9) results in the form:

$$\nabla_{\boldsymbol{A}}\left(\hat{\rho}_{CKA}\left(\boldsymbol{K}_{\boldsymbol{X}},\boldsymbol{K}_{\boldsymbol{l}}\right)\right) = -4\boldsymbol{X}^\top\left(\left(\boldsymbol{G}\circ\boldsymbol{K}_{\boldsymbol{X}}(\boldsymbol{A},\sigma)\right) - \mathrm{diag}\left(\mathbf{1}^\top\left(\boldsymbol{G}\circ\boldsymbol{K}_{\boldsymbol{X}}(\boldsymbol{A},\sigma)\right)\right)\right)\boldsymbol{X}\boldsymbol{A}, \tag{6-10}$$

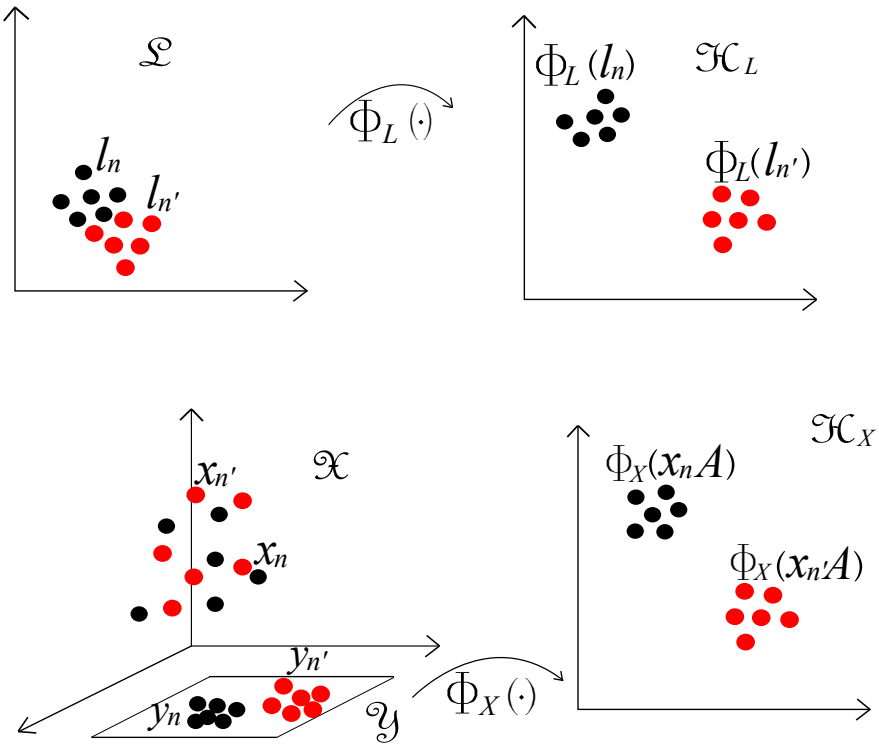


Figure 6-1.: Diagrama of the proposed SRKA approach.

where notations diag$(\cdot)$ and $\circ$ denote the diagonal operator and the Hadamard product, respectively. $\boldsymbol{G}\in\mathbb{R}^{N\times N}$ is the gradient of the objective function with respect to $\boldsymbol{K_X}(\boldsymbol{A},\sigma)$, calculated as follows:

$$\boldsymbol{G}=\nabla_{\boldsymbol{K_X}(\boldsymbol{A},\sigma)}\left(\hat{\rho}_{CKA}\left(\boldsymbol{K_X},\boldsymbol{K_l}\right)\right)=\frac{\tilde{\boldsymbol{I}}\boldsymbol{K_l}\tilde{\boldsymbol{I}}}{\operatorname{tr}\left(\boldsymbol{K_X}(\boldsymbol{A},\sigma)\tilde{\boldsymbol{I}}\boldsymbol{K_l}\tilde{\boldsymbol{I}}\right)}-\frac{\tilde{\boldsymbol{I}}\boldsymbol{K_X}(\boldsymbol{A},\sigma)\tilde{\boldsymbol{I}}}{\operatorname{tr}\left(\boldsymbol{K_X}(\boldsymbol{A},\sigma)\tilde{\boldsymbol{I}}\boldsymbol{K_X}(\boldsymbol{A},\sigma)\tilde{\boldsymbol{I}}\right)}. \tag{6-11}$$

For updating the estimation of $\boldsymbol{A}$, we use the standard stochastic gradient descent update rule, provided the initial guess $\boldsymbol{A}^o$, as follows:

$$\boldsymbol{A}^{t+1}=\boldsymbol{A}^t-\mu_A^t\nabla_{\boldsymbol{A}^t}\left(\hat{\rho}_{CKA}\left(\boldsymbol{K_X},\boldsymbol{K_l}\right)\right) \tag{6-12}$$

where $\mu_A^t\in\mathbb{R}^+$ is the step size of the learning rule. $\boldsymbol{A}^t$ and σ^t are the samples we use at the time step t.

Since the kernel bandwidth allows scaling all pairwise distances on the projected space $\boldsymbol{Y}^t=\boldsymbol{X}\boldsymbol{A}^t$, we estimate σ^t through the introduced he introduced *Kernel Function Estimation from Information Potential Variability*-(KEIPV) in Chapter 2. Thus, we maximize the

overall variability of the so termed information potential of all samples over $\boldsymbol{Y}^t$ with respect to the kernel bandwidth parameter so that all information force magnitudes spread more widely.

6.2. SRKA as feature selection/embedding approach

Once the projection matrix $\hat{\boldsymbol{A}}$ is estimated, we can estimate the relevant feature matrix $\boldsymbol{Y}\in\mathbb{R}^{N\times M}$ holding row vectors $\boldsymbol{y}_n$ to encode the linear combination of discriminative input features according to the prior knowledge considered in $\boldsymbol{K}_{\boldsymbol{l}}$.

In turn, we introduce a feature relevance vector index, noted as $\boldsymbol{\varrho}\in\mathbb{R}^P$, that is devote to measuring the contribution of each input feature for building the projection matrix $\hat{\boldsymbol{A}}$ as:

$$\varrho_p = \sum_{m=1}^{M} |a_{pm}|; \forall p\in P, \tag{6-13}$$

with $a_{pm}\in\boldsymbol{A}$. The main assumption behind the introduced relevance index is that the largest values of ϱ_p should point out to better input attributes since they exhibit higher overall dependencies to the estimated metric based on the CKA principle. As a result, the calculated relevance vector $\boldsymbol{\varrho}$ can be employed to rank the original features.

In addition, aiming to estimate a representation space encoding discriminant input patterns, we compute the matrix $\boldsymbol{X}_S\in\mathbb{R}^{N\times M_S}$ ($M_S\leq P$) holding the features in $\boldsymbol{X}$ satisfying the following condition:

$$\varrho_p \geq \mathbb{E}_p\left\{\varrho_p\right\}. \tag{6-14}$$

Lastly, the embedding matrix $\boldsymbol{Y}_S\in\mathbb{R}^{N\times M_E}$ is calculated as:

$$\boldsymbol{Y}_S = \boldsymbol{X}_S\boldsymbol{A}_E, \tag{6-15}$$

where $\boldsymbol{A}_E\in\mathbb{R}^{M_S\times M_E}$ is a rotation matrix which is computed from $\boldsymbol{X}_S$ using eq. (6-8), where $M_S\leq M_E$.

6.3. Experimental set-up

The validation of the proposed SRKA approach as a suitable tool to support data discrimination patterns includes the following main stages: *i)* Feature estimation from the preprocessed datasets, *ii)* Relevance analysis of the estimated feature set, and *iii)* Discrimination learning among different labels. fig. **6-2** summarizes the main scheme of the proposed data discrimination approach based on SRKA.

Figure 6-2.: Diagrama of the proposed SRKA approach as a tool to support classification tasks.

The validating experiments are carried out on two databases reflecting different brain activity tasks:

Motor Imagery Database (MIDB) [146]. Dataset 1 used in the BCI competition IV 2008. This electroencephalogram (EEG) collection is widely used in motor imagery (MI) tasks and holds seven subjects with EEG signals recorded from 59 channels. All recordings are submitted firstly to a bandpass filter with bandwidth ranging from 0.05 to 200 Hz, and then to a 10-order low-pass Chebyshev II filter with stop-band ripple of 50 dB down and stop-band edge frequency of 49 Hz. All recordings are further digitized at 1000 Hz and down-sampled to supply the sampling frequency at 100 Hz. The whole session is performed without feedback and 100 repetitions are recorded for each MI class per person. The section of interest is 4 s during when the subject is instructed to perform the MI task indicated by a pointing arrow on a screen. These periods lasting 2 s are interleaved with a blank screen and a fixation cross in the screen center. As the preprocessing stage, a 5-order band-pass Butterworth filter is implemented with bandwidth ranging from 8 to 30 Hz. Then, we carry out a data-driven supervised decomposition of the EEG multi-channel data based on the Common Spatial Patterns (CSP) algorithm. Both strategies are applied aiming to extract adaptively components carrying MI information [62].

"Klinik für Epileptology" database (KEDB) [9]. This dataset is widely used in the automated detection of epileptic seizures and contains five subsets noted as `A`, `B`, `C`, `D`, and `E`. Each subset is composed of 100 single channel EEG segments of 23.6 s duration. The subsets `A` and `B` are acquired from five healthy subjects with eyes opened and closed, respectively. All signals from subsets `C, D` and `E` come from five epileptic subjects. Subsets `C` and `D` include seizures-free interictal signals measured on the epileptic zone and on the hemisphere opposite to the hippocampal formation of the brain. Set `E` contains epileptic signals recorded from each aforementioned location during an ictal seizure. Subsets `C, D` and `E` were recorded intracranially. Besides, all provided EEG signals in KEDB were digitized at 173.61 Hz and 12 - bit resolution. To retain relevant EEG information related to the studied normal and epileptic conditions, all signals were filtered through a low-pass filter with a 40 Hz cutoff frequency. For the validation purpose, this data is tested on three problems, according to the medical interest [151]: *Bi-class* (2C), normal (`A`-type) and seizure (`E`-type) labeled

recordings are distinguished; *Three-class* (3C), closely represents real medical applications, including three categories: normal (A-type EEG segments), seizure-free interictal (D-type EEG segments), and seizure (E-type EEG segments); *Five-class* (5C), all five classes are investigated: normal (Types A and B), interictal (Types C and D) and seizure (Type E).

Feature set estimated from MIDB . Let $\{\boldsymbol{\Psi}_n \in \mathbb{R}^{C_h \times T}\}$ $(n \in [1, N])$ be a set of N EEG raw data trials for a given subject of MIDB, where C_h and $T \in \mathbb{N}$ are the number of EEG channels and the amount of time samples, respectively. Depending on the used principle of extraction, the following short-time features are computed for each EEG trial $\boldsymbol{\Psi}_n$ to carry out discrimination of the MI paradigm [5]:

- *Spectral parameters*: For each row channel vector $\boldsymbol{\psi}_n^c \in \boldsymbol{\Psi}_n$ $(c_h \in [1, C_h])$, the vector $\boldsymbol{r} \in \mathbb{R}^{R_s}$ is computed using the Power Spectral Density (PSD), where $R_s = \lfloor F_s/2 \rfloor$ is the number of frequency bins and $F_s \in \mathbb{R}^+$ is the sampling frequency. Then, $\boldsymbol{\psi}_n^c$ is split into $T_J \in \mathbb{N}$ overlapped segments of length $J \in \mathbb{N}$ and a piecewise stationary assumption is imposed by means of a smooth time weighting window $\boldsymbol{w} \in \mathbb{R}^L$ [31]. Hence, the windowed segments $\boldsymbol{v}^{t_J} \in \mathbb{R}^J$ $(t_J \in [1, T_J])$ are extracted. The modified periodogram vector $\boldsymbol{u} = \{u_f \in \mathbb{R}^+ : f \in [1, R_s]\}$, $\boldsymbol{u} \in \mathbb{R}^{R_s}$ is computed by using the Discrete Fourier Transform and each PSD element is further calculated as:

$$r_f = \frac{u_f}{T_J \nu}, \tag{6-16}$$

  where $\nu = \mathbb{E}_j \{|w_j|^2 : \forall j \in [1, J]\}$.

- *Hjorth parameters*: Three time-domain based parameters are estimated from each windowed segment $\boldsymbol{v}^{t_J}$.

  – Activity: $\boldsymbol{\varsigma}_{\boldsymbol{v}}^2 \in \mathbb{R}^{T_J}$, that holds elements:

$$\varsigma_{t_J}^2 = \text{var}\left\{\boldsymbol{v}^{t_J}\right\}. \tag{6-17}$$

  – Mobility: $\boldsymbol{\lambda}_{\boldsymbol{v}} \in \mathbb{R}^{T_J}$, that measures the signal mean frequency as:

$$\lambda_{t_J} = \sqrt{\frac{\text{var}\left\{\partial \boldsymbol{v}^{t_J}\right\}}{\text{var}\left\{\boldsymbol{v}^{t_J}\right\}}}, \tag{6-18}$$

  being $\partial \boldsymbol{v}$ the derivative of $\boldsymbol{v}$.

– Complexity: $\boldsymbol{\vartheta_v}\in\mathbb{R}^{T_J}$, measuring frequency variations as the deviation of the signal from the sine shape as:

$$\vartheta_{t_J} = \frac{\partial \iota_{t_J}}{\iota_{t_J}}, \tag{6-19}$$

where $\partial\iota_{t_J}$ is the derivative of ι_{t_J} [124]. For concrete testing, the segment length value L needed during calculation of the PSD and the Hjorth parameters is adjusted as $L{>}F_r/F_s$, where $F_s{=}100\,Hz$ and $F_r{=}8\,Hz$ [148].

- *Time-Frequency parameters.* The Continuous Wavelet Transform (CWT) vector $\boldsymbol{\varsigma}^g\in\mathbb{C}^T$ is extracted from channel $\boldsymbol{\psi}_n^c$ at scale $g\in\mathbb{R}$ as:

$$\varsigma_t^g = \sum\nolimits_{\tau=1}^{T} \psi_{n\tau}^{c_h}\gamma^*\left((\tau{-}t)\zeta_t/g\right), \tag{6-20}$$

where $\gamma(\cdot)$ is the mother wavelet function, $\zeta_t\in\mathbb{R}$ is a time spacing, and $(^*)$ denotes the complex conjugate. Both procedures of Wavelet scaling g and translating through the localized time index t are used to model amplitude time variations. Also, the Discrete Wavelet Transform (DWT) is considered by computing the detail vector $\boldsymbol{d}^j\in\mathbb{C}$ at level j as follows:

$$d_t^j{=}\sum\nolimits_{k\in\mathbb{Z}} \wp_{jk}\psi_{jk}(t), \tag{6-21}$$

where $\wp_{jk}{=}\sum_{t\in T}\psi_{nt}^{c_h}h_{jk}(t)$; $\wp_{jk}\in\mathbb{C}$, and being $h_{jk}(t)\in\mathbb{C}$ the impulse response of a given wavelet filter. The DWT of $\boldsymbol{\psi}_n^{c_h}$ is computed for a given mother wavelet $\varpi\,(\cdot)$ as:

$$\psi_{nt}^c = \sum\nolimits_{j\in\mathbb{Z}}\sum\nolimits_{k\in\mathbb{Z}} \wp_{jk}\varpi_{jk}(t). \tag{6-22}$$

For computation of the WT-based feature subset, the Morlet wavelet is employed because its wave shape and EEG signals. Thus, we derive the short-time instantaneous CWT amplitudes using a couple of Morlet wavelets; one centered at $10\,Hz$ and another at $22\,Hz$, aiming to extract the α_B and β_B rhythms during the MI task, respectively [90]. Likewise, we employ for the DWT the Symlet wavelet (Sym-7) that is closely associated with the electrical brain activity and proved to be appropriate in similar applications [23]. For the tested MIDB EEG data, we compute the detail coefficient vector as to include the α and β rhythms, resulting in the second and third levels of decomposition.

Once we calculate all the above short-time parameters, several of their statistical measures are further considered to extract the input feature matrix $\boldsymbol{X}\in\mathbb{R}^{N\times P}$. Namely, the mean, the variance, and the maximum values are estimated. Consequently, the row vector $\boldsymbol{x}_n\in\mathbb{R}^P$ ($P{=}C{\times}Q$) concatenates all features extracted from each n-th MI trial per channel, being $Q{\in}N$ the number of provided features. Thus, 27 features are obtained for each EEG channel. So, the size of the concatenated feature vector is $P{=}59{\times}27$, and the number of samples is $N{=}200$.

Feature set estimated from KEDB. The rhythms carrying out clinical and physiological interest fall primarily within the following four frequency sub-bands: Delta denoted as δ_b with frequencies $f{<}4\,Hz$, Theta (θ_B, $f{\in}[4,8]$) Hz, Alpha (α_B, $f{\in}[8,13]\,Hz$), and Beta rhythms (β_B, $f{\in}[14,30]\,Hz$). Then, we select the linear filter bank for representation of EEG signals because they may more accurately refined to each rhythm frequency bandwidth. Therefore, we use five cepstral coefficients associated with δ_B, θ_B, α_B, and β_B rhythms, extracted as dynamic features as in [41]. As a result, instead of a widely used scalar-valued parameter set extracted from the EEG signal, neural activities relating to epileptic seizures are detected by using a vector set of short-time rhythms.

We carry out the validation of the proposed SRKA method for two scenarios of training: *i*) SRKA as a feature selection tool to provide better understanding of the salient aspects of the input feature set, facilitating the physiological interpretation task. *ii*) SRKA as a feature embedding tool that generates, by feature transformation, new composites of the input feature set with the goal of improving overall data discrimination performance.

In the former scenario, the relevance vector $\boldsymbol{\varrho}$ ranks the original feature set of $\boldsymbol{X}$. We calculate the performed accuracy curve of the brain activity classification through the 10-fold cross-validation scheme, adding one by one the features ranked by the amplitude of $\boldsymbol{\varrho}$. For the purpose of comparison in terms of physiological interpretation, the proposed SRKA is contrasted with a baseline variance-based relevance analysis (termed VRA) that ranks the input short-time features grounded on a variability criterion. Namely, VRA computes a relevance vector based on a linear transformation of the input representation space. Thus, VRA estimates the covariance among input features and the projection matrix maximizing the embedded space variability is fixed to computed such a linear transformation [35]. Therefore, provided a set of features $\Xi{=}\{\boldsymbol{\xi}_p : p{\in}[1,P]\}$, where $\boldsymbol{\xi}_p{\in}\mathbb{R}^N$ corresponds to each column of the input data matrix $\boldsymbol{X}$, the relevance of $\boldsymbol{\xi}_p$ can be measured by computing the following variability vector $\boldsymbol{\rho}{\in}\mathbb{R}^P$ [111]:

$$\boldsymbol{\rho} = \mathbb{E}_p\left\{|\chi_p \boldsymbol{w}_p| : \forall p{\in}M \leq P\right\}, \tag{6-23}$$

where $\chi_p{\in}\mathbb{R}^+$ and $\boldsymbol{w}_p{\in}\mathbb{R}^P$ are respectively the eigenvalues and eigenvectors of the covariance matrix estimated as $\boldsymbol{X}^\top\boldsymbol{X}/P$. The main assumption behind the relevance measure introduced in eq. (6-23) is that the largest values of λ_p should point out to the better input

attributes since they exhibit higher overall correlations to the estimated principal components.

In the latter scenario of training, we aim to estimate a feature representation space that encodes discriminant patterns through the embedding matrix $\boldsymbol{Y}_S$. Furthermore, the other aspect of training to consider is the tuning of the SKRA free parameters. To this end, we calculate the number of embedded dimensions M_S and M_E as to maintain 95% of the variance explained. Besides, the number of nearest neighbors of the applied k-nearest neighbor classifier is fixed as the one reaching the best accuracy within the following testing range $\{1, 3, 5, 7, 9, 11\}$.

6.4. Results and discussion

By the above-described validation stage, firstly we examine the carried out relevance analysis as a feature selection tool. So, fig. **6-3** shows the obtained relevance planes averaged over all subjects for both compared algorithms, SRKA and VRA. While the vertical axis holds the number assigned to each of 27 feature estimation principles, horizontal axis stands for the cardinal number of each one of the 59 channels labeled with the international 10-20 electrode location montage.

As seen in fig. 6.3a showing the performed relevance by the VRA algorithm, there are three spatially distinguished channel groups of relevance, meaning that the performed relevance analysis allows differing the contribution of every single electrode position. To quantify the difference, the top plot of the relevance plane in fig. 6.3a displays the marginal relevance per channel that is averaged over all used features. As noted, the first labeled 13 channels that are placed over the association cortex have the strongest influence with the lowest dispersion. Then, the positions with labels ranging from 34 to 59, which collect the neural activity in the anterior parts of the anterior parietal cortex, supply lower relevance values. Lastly, the electrode positions 14 to 33 produce the lowest relevance having even the highest variance. This electrode set is positioned on the precentral gyrus (14-28) and the dorsal lateral premotor area (29-34). Although the distinguished groups remain the above, the SRKA algorithm performs distinct relevance values for each channel as seen in the top plot of fig. 6.3b. The marginal profile shows that the positions from 14 to 32 now become the most important succeeded by the group 33 to 59. The channels 1-13 perform the worst in contrast to the VRA approach.

With regard to the principle of feature extraction, the studied features do not cluster so distinctly for either selection algorithm, though most of the characteristics behave differently depending on the electrode position of measurement (see right-side plots of figs. 6.3a and 6.3b). However, the chosen statistical measure used for characterizing all short-time parameters plays a significant role.

For the sake of visual representation, we rearrange each plane so that the relevance estimates are now valued ranked in decreasing order on the channel and extraction principle

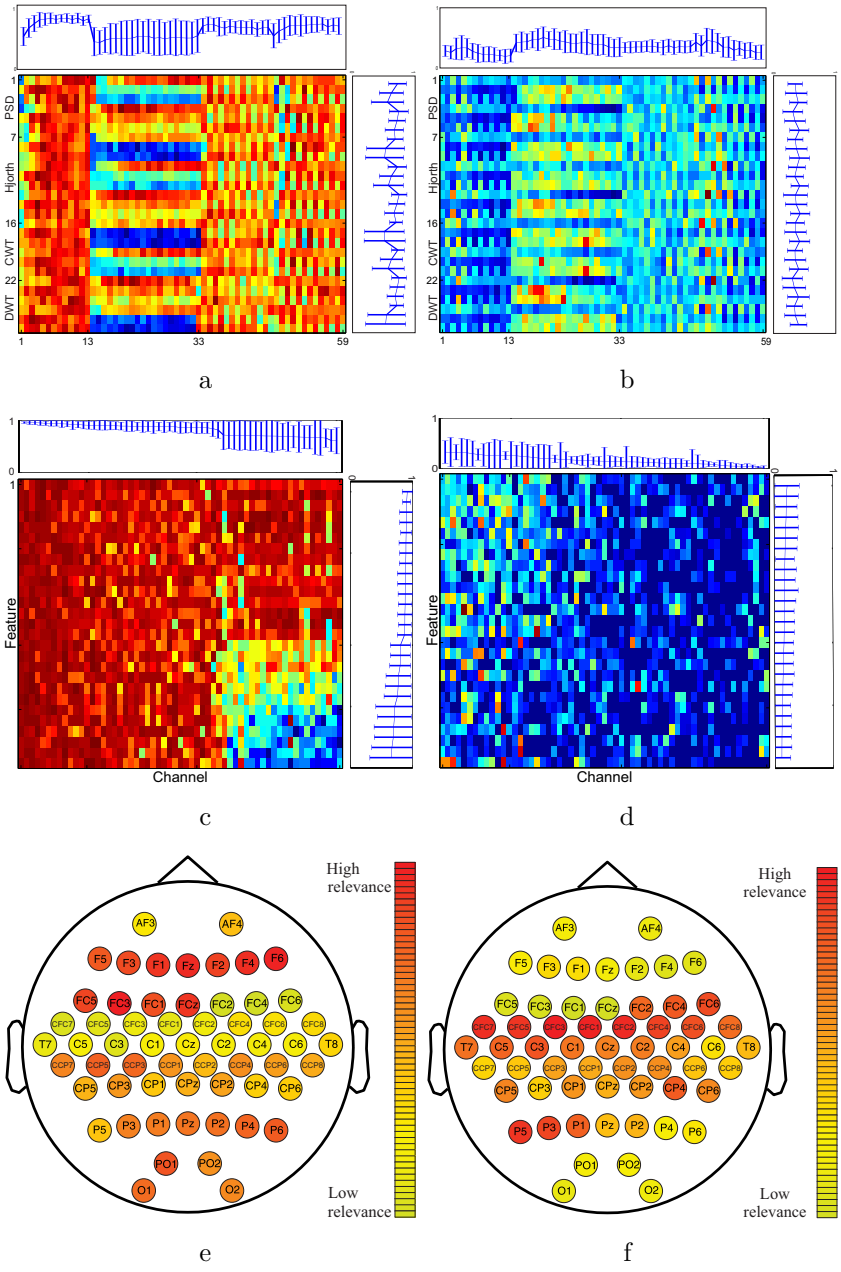
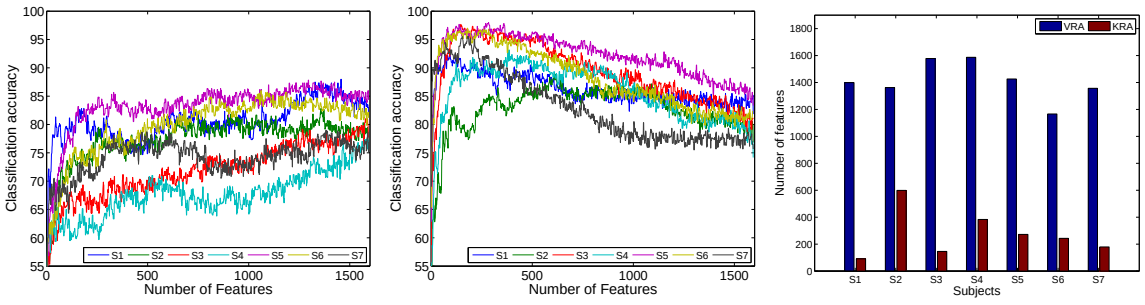


Figure 6-3.: Performed MIDB relevance analysis: VRA - left column, SRKA - right column. Top row: computed planes of relevance averaged over all subjects. Plot on the top shows the marginal relevance per channel. right-side plot: averaged marginal relevance for all considered features. middle row: computed planes in decreasing relevance. Bottom row shows the feature relevance channel distribution.

axes. So, the features that do not contribute to 95% of the variance explained are zero-valued. Inspection of figs. 6.3c and 6.3d reinforces the finding that either relevance estimator associates the input training set in a different way. Overall, the baseline VRA algorithm produces higher values of the relevance marginal (see top and right-side plots of each plane) in comparison to the proposed SRKA, suggesting that the latter approach encodes the whole brain activity task into a lower number of features. This advantage of SRKA can be explained by the following two facts: *i*) the use of the MI label information to reveal features, which must be salient in terms of the studied paradigm. Thus, the brain activity patterns are better localized. *ii*) Representation through enhanced RKHS allows dealing with complex neighboring data dependencies, rejecting more efficiently redundant features. In contrast, VRA mainly explains the relevance in terms of its energy-based cost functional that emphasizes the brain regions with strong activity, which are activated during the time the stimuli goes. Yet, this assumption does not necessarily hold for MI paradigms.

To further explore the physiological interpretation of the carried out feature selection, all computed relevance values are arranged in the 10-20 channel montage as displayed in the bottom row of fig. **6-3**. It is worth noting that we will describe the MI brain activity performed by a hypothetical medium person due to the estimated relevance planes are averaged over all subjects. So, VRA produces the highest contribution of relevance for the middle frontal gyrus represented by channels F5, F3, F4, and F6 as shown in fig. 6.3e. But, the middle frontal gyrus should not be related to any imagery stimulation [58]. Rather, this brain area activates as a response to body movements, e.g., powerful EEG artifacts. In other words, the presence of EEG channels with high-energy disturbances may mislead the VRA estimator, identifying wrongly MI patterns.

On the other hand, SRKA assigns the bigger values of relevance to the EEG channels placed over two brain areas that are commonly related to MI tasks. Namely (see fig. 6.3f), the posterior superior parietal cortex (P3, P1, PZ, and P2), and the left precentral sulcus at the level of the middle frontal gyrus (CFC5, C3, CFC3, C1, and CFC1). Furthermore, the middle frontal gyrus has the lowest contribution, weakening the influence of movement artifacts.



a VRA b SRKA c DR

Figure 6-4.: Selected training set for MI discrimination.

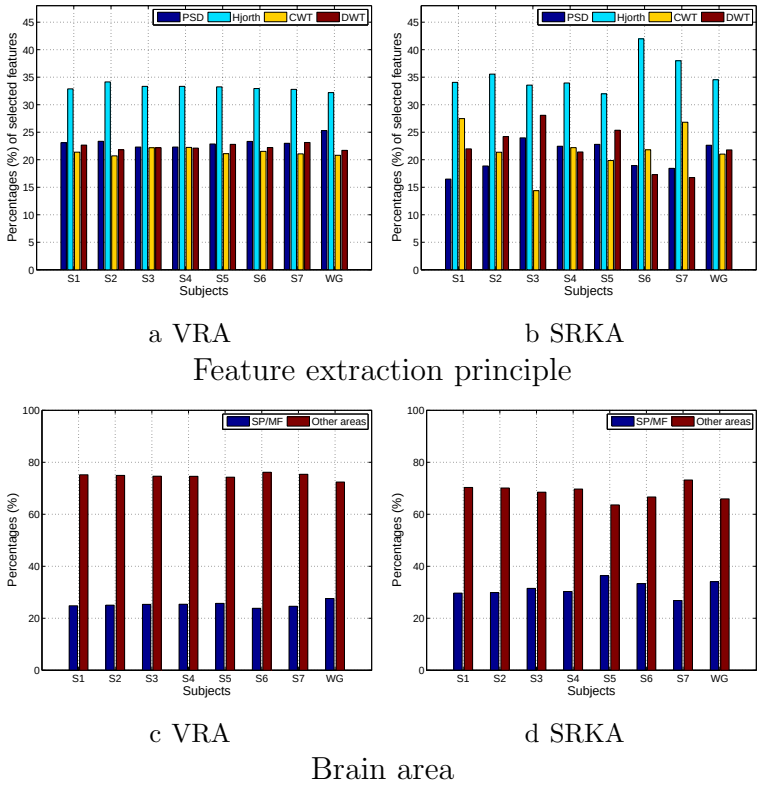


Figure 6-5.: Contribution of the selected feature set to the MI discrimination performance.

The following training scenario of validation is the discrimination of the contemplated MI tasks. In this case, we assume the selected training set as the one containing the minimum amount of features to reach the maximum classification accuracy. To this end, the k-nearest neighbor classifier is fed by adding one by one the relevant features ranked in decreasing order so that we have the plots shown in fig. **6-4**, where the classification accuracy is performed through the 10-fold cross-validation scheme.

As seen in fig. 6.4a, VRA performs an accuracy close to $\sim$85.16$\pm$3.88% and definitely falls behind the SRKA algorithm that reaches $\sim$95.71$\pm$3.01% (see fig. 6.4b) averaged for all subjects. Hence, fig. 6.4c display the number of selected training features estimated for each subject by the VRA and SRKA algorithms.

Along with the MI discrimination performance, another important aspect to explain is the number of selected training features from the whole input set (1593). As displayed in fig. 6.4c, VRA chooses about 1410 features, but SRKA does only 275 features. Consequently, a dimension reduction is close to one and 5.8, respectively, averaged for all subjects. Therefore, SRKA is as much as five times more efficient than the contrasted baseline estimator, regarding the reduction dimension processing.

A detailed analysis of both selected features set gives the following findings:

– The Hjorth principle of extraction clearly supplies the features with the highest rele-

vance, contributing the most to the discrimination of MI classes regardless of the used estimator of relevance (see figs. 6.5a and 6.5b). The remaining spectral characteristics have a comparable contribution though some differences apply for SRKA.

– As seen in figs. 6.5c and 6.5d displaying the proportion of features encoding MI information (the *superior parietal* plus *middle frontal gyrus*), SRKA produces a higher number of salient features.

– As one of the major challenges in BCI research, it is worth mentioning the inter-subject variability with respect to spatial patterns and spectrotemporal characteristics of brain signals [17]. In the contemplated MI task, some subjects might not focus their gaze in the proper direction, and thus the EEG recordings will not be reliable for interpretation. From the comparison plots of relevance in fig. **6-5**, it follows that KRA better adapts the BCI system for each particular subject, at least, in terms of revealing the most discriminating features.

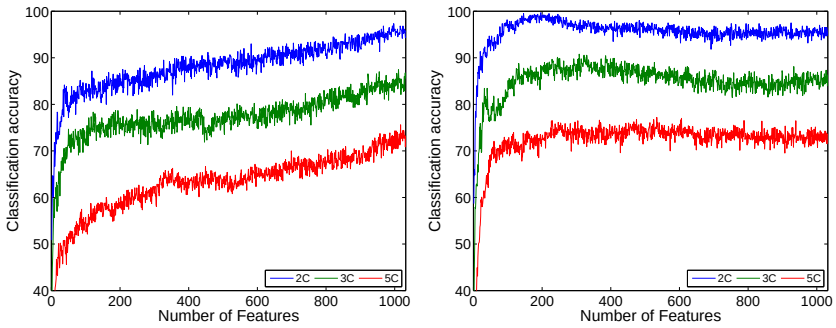
As shown in table **6-1** for all subjects, we compare the suggested Kernel-based relevance analysis for feature selection and feature embedding (noted as SRKA$*$) in terms of the classifier accuracy achieved for the contemplated MI task. In the former training scenario, SRKA reaches an averaged accuracy 95.71 ± 03.01 and outperforms the contrasted baseline VRA that produces 92.86 ± 03.77. For the sake of comparison of the latter scenario, we include the accuracy estimated by the approach submitted in [168] that selects an extracted spatiotemporal feature set, from which a non-linear regression for predicting the time-series of class labels is applied. Also, the work in [62] that uses an adaptive frequency band selection of the spatial preprocessed features that feed an SVM classifier. Lastly, we consider the approach in [65] involving common space-time-frequency patterns to design certain time-windows adopted for the MI task, as well as, the LDA classifier. All referred approaches of training underperform the proposed SRKA method.

With respect to the KEDB results, in the beginning, we test both, VRA and SRKA, approaches as a feature selection tool of the spectral coefficients extracted from the physiological rhythms (δ_B, θ_B, α_B, and β_B). Since the KEDB dataset only has one-channel EEG recordings, the physiological interpretation of the selected feature set only covers the influence of the physiological waveforms on the three available challenges of epileptic seizure detection. The selected feature set is calculated as in fig. **6-4** for which the accuracy of the k-nearest neighbor classifier is also performed through the 10-fold cross-validation scheme.

As seen in figs. 6.6a and 6.6b, either comparative approach of feature selection attains the highest accuracy (100%) for the bi-class task. Further, SRKA betters the baseline VRA for the tasks of three-classes (99.67 versus 90.78%, respectively) and five-classes (89 versus 77.2%). Regarding the number of selected training features, once again the SRKA approach outperforms VRA in all tasks (see fig. 6.7a). It is useful to note that the VRA classification accuracy increases as the number of features grows, requiring the whole input feature set

Table **6-1**.: Performed classification accuracy for MI discrimination (average ± standard deviation [%]). Notation (-) stands for Not provided. Note that the accuracy of SRKA and VRA is estimated as the highest value performed in fig. 6.4b for each tested subject.

Subject	VRA [5]	SRKA	He [62]	Zhang [168]	Higashi [65]	SRKA*
# 1	91.50 ± 05.29	94.16±05.30	67.70 ± 02.20	77.20 ± 00.03	92.30 ± 02.50	**96.00 ± 03.94**
# 2	96.50 ± 03.37	90.16±05.88	70.70 ± 01.20	70.80 ± 00.02	90.60 ± 7.20	**96.50 ± 06.25**
# 3	91.50 ± 04.74	98.50±08.57	83.90 ± 01.30	-	-	**98.00 ± 04.83**
# 4	87.00 ± 06.32	94.50±07.01	93.00 ± 01.20	-	-	**100 ± 00.00**
# 5	91.50 ± 07.47	98.50±04.60	93.20 ±01.20	-	-	**100 ± 00.00**
# 6	98.50 ± 02.42	97.66±04.82	-	76.80 ± 00.03	93.30 ± 03.60	**100 ± 00.00**
# 7	93.50 ± 07.09	96.50±03.45	-	80.00 ± 00.03	94.10 ± 04.10	**96.00 ± 02.10**
Average	92.86 ± 03.77	95.71±03.01	81.70 ± 12.06	76.20 ± 03.87	92.58 ± 01.51	**98.07 ± 01.92**



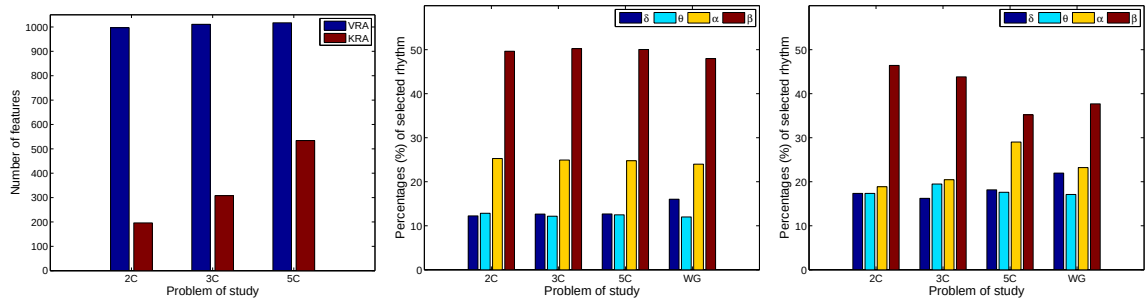
a VRA b SRKA

Figure 6-6.: Performed accuracy for epileptic seizure detection

to reach the maximum performance. Meanwhile, the addition of more features drops the performance once the SRKA approach gets the highest accuracy, indicating that the inclusion of other features may be redundant. As a result, the dimension reduction is two, three, and five times bigger than the one obtained by VRA for the 2C,3C, and 5C tasks, respectively. This aspect can be of benefit for reliable on-line monitoring of traces of interictal/ictal states of epilepsy since the demanded time-window of EEG analysis may be remarkably shortened.

fig. **6-7** shows the normalized relevance values that are estimated for each rhythm. By VRA (see fig. 6.7b), the selected features make the α_B and β_B waveforms the most relevant for all considered tasks. At the same time, low-frequency rhythms (δ_B, θ_B) exhibit modest values of relevance. Although SRKA infers a similar contribution of the rhythms, the relationship between the high to low-frequency rhythms decreases as the number of classes increases as shown in fig. 6.7c. This result indicates on the energy redistribution happening as the complexity of the task increases and has been also explained in similar works [40].

For the sake of comparison, the SRKA approach is tested against some recent approaches using KEDB. Although this comparison may not be entirely fair due to different details on



a Dimension Reduction b VRA c KRA

Figure 6-7.: Relevant rhytms in terms of MI discrimination performance

the testing procedures [81, 165], it seems to be the best possible option. table **6-2** shows the best classification accuracy for each considered problem. The obtained results indicate high classification ability in the epileptic seizure detection. For the first (*Bi-class*) and second (*Three-class*) classification problems almost all benchmark approaches present high classification accuracy, i.e., from 99.5% to 100% for *Bi-class* problem, and from 90.78% to 100% for *Three-class*. Concerning the third classification problem (*Five-class*), the discrimination performance varies from 77.2% to 99.2%. In this case, authors in [152] perform the best classification accuracy.

6.5. Summary

We discuss a novel supervised kernel-based representation approach for feature relevance analysis to enhance the automatic identification of relevant patterns. To this end, the proposed relevance analysis, termed SRKA, incorporates two kernel functions to take advantage of the available joint information associating the input features to a certain condition with the corresponding labels. Then, a kernel alignment functional learns all relevant patterns from the input space. Validation of the proposed approach is carried out in two scenarios of training: feature selection and feature embedding. In particular, two tasks of brain activity identification are studied: motor imagery discrimination and epileptic seizure detection.

From the obtained results of validation of real EEG data, the following observations deserve close attention:

The need for handling a couple of kernels encoding different notions of similarity encourages the use of the well-known kernel alignment to unify both tasks into a single optimization framework. However, selection of the distances implementing each aligned kernel as well as the same alignment mostly determines the affectivity of the kernel-based approach for a given application. In the particular case of brain activity identification, we rely on the Mahalanobis distance to carry out the pairwise comparison between samples based on the Gaussian kernel. Thus, a linear projection is further learned from the employed CKA-based

Table **6-2**.: Accomplished classification results for KEDB

Problem	*Authors*	*Features/Classifier*	*Accuracy*
2-class	[142]	*t-f* analysis/RNN	99.6
	[48]	WT/PNN	99.99
	[120]	PCA FFT/AIRS	100
	[71]	CC+PSD/vot. rule	100
	[99]	TFR-2DPCA/*k-nn*	100
	[151]	*t-f* analysis/ANN	100
	[41]	short-time /*k-nn*	99.5
	Proposal	SRKA	**100**
	Proposal	SRKA*	**100**
3-class	[51]	PCA-RBF/ANN	96.60
	[106]	EV/MLP NN	97.50
	[145]	PSD+CLZ/SVMA	98.72
	[99]	TFR-2DPCA/*k-nn*	98.80
	[151]	*t-f* analysis/ANN	100
	[41]	short-time /*k-nn*	100
	Proposal	SRKA	**90.78**
	Proposal	SRKA*	**99.67**
5-class	[152]	(WT+eig)/SVM	**99.20**
	[99]	TFR-2DPCA/*k-nn*	94.40
	[151]	*t-f* analysis/ANN	89.00
	[41]	short-time /*k-nn*	95.78
	Proposal	SRKA	77.20
	Proposal	SRKA*	89

functional as an alternative to highlight the salient input features, taking advantage of the nonlinear notion of similarity behind the studied kernels. For the sake of simplicity, the iterative gradient descent optimization is employed to calculate the projection matrix and the Gaussian kernel free parameter.

For the purpose of implementing SRKA as a feature selection tool, we introduce a feature relevance vector index devoted to measuring the contribution of each input feature for building the projecting CKA matrix. By ranking this contribution, we assess the selected feature set that satisfies a given stopping criteria (here, we fix the proportion of explained variance). Thus, the feature selection by SRKA demands small feature sets with the benefit of providing better interpretation of the space brain activity distribution and the principle of employed feature extraction. Besides, the SRKA-based ranking separates redundant features, which usually tend to drop the system accuracy. As another advantage, SRKA adapts the relevance analysis to the inter-subject variability that remains one of the most challenging issues of training for BCI systems. Therefore, SRKA as a feature selection tool can reach a suitable classification accuracy with a remarkable dimension reduction factor, providing better physiological interpretation of the brain activity patterns.

For the another training scenario of feature embedding, we use the relevance index vector to estimate the representation space that optimizes a trade-off between separability and no redundancy of the available neural patterns. As a result, our proposal outperforms those compared approaches that carry out the multivariate feature selection and/or embedding. Indeed, the SRKA-based embedding spaces handle the brain activity complexity to support further classification stages in terms of system accuracy and reliability.

7. Relevant data representation from different kernel spaces: a generalized cross-correntropy measure

Nowadays, there is a need to process a large amount of information, high-dimensional-(HD) data, for supporting automatic learning procedures, such as classification, prediction, and dimensionality reduction. Regarding this, natural processes of interest for engineering are composed of two basic characteristics: statistical distribution of amplitudes and data domain-varying behavior, for instance, the time structure. On one hand, there are widely used measures that quantify the time structure like the autocorrelation function. On the other hand, there are a number of methods that are solely based on the statistical distribution, ignoring the data domain-varying behavior [130]. Then, a single measure that includes both of these important characteristics could greatly enhance the performance of machine learning systems when dealing with complex data relations.

Keeping in mind aforementioned needs, authors in [93] proposed a new measure of similarity termed cross-correntropy as a localized similarity based on ITL. Therefore, cross-correntropy contains higher order moments of the pdf but is much simpler to estimate directly from samples and bypasses the need for conventional moment expansions. In addition, cross-correntropy are closed related to kernel methods and RKHS from a covariance function perspective [10]. Thereby, cross-correntropy is a generalized correlation function in terms of inner products of vectors in a kernel feature space. Since inner products are a measure of similarity, this function in effect measures the pairwise interaction of the feature vectors. Besides, from an ITL point of view, this measure quantifies the shape and size of the group of points in feature space, which gives the information of the statistical distribution in the input space. In addition, cross-correntropy is directly related to Renyi's quadratic entropy estimate of data using Parzen windowing. Therefore, cross-correntropy allows extracting more information from the data structure for the adaptation process in machine learning applications, yielding solutions that are more accurate than traditional mean square error methods, specially, for non-Gaussian process [60, 59, 61, 96].

With an abundance of tools based on kernel methods and ITL, cross-correntropy appears as a flexible alternative to find pair-wise sample relations taking advantage of relevant local dependencies. Nonetheless, there are some limitations regarding this measure: *i)* There is still a lack regarding the required free parameters, namely, the kernel parameters. *ii)* It is

not clear how to reveal relevant input features from correntropy-based measures aiming to support the data interpretability, *iii)* such a localized measure employs a unique RKHS to encode data dependencies, which can be not appropriate when dealing with different data structures/dynamics, e.g., nonstationary processes, and *iv)* the user prior knowledge can not be directly incorporated when computing the data similarities.

In this work, a new generalized similarity measure, termed Generalized cross-Correntropy-(GCC), is introduced to reveal relevant dependencies among HD samples. Our approach is a data-driven kernel-based measure that includes both the distribution and the structure of the input process by representing the samples using different RKHSs. In addition, a relevant feature ranking criteria is described based on GCC to highlight each input feature contribution into the studied process. Furthermore, an adaptive learning constraint is imposed in GCC to incorporate the domain-varying behavior of the HD data, when available. The proposed GCC is an extension of the well-known cross-Correntropy measure based on Hilbert space embeddings. So, it is shown how GCC can be interpreted from kernel methods as well as from an information theoretic points of view. To test the capability of the proposed approach, two main learning tasks are studied in our experiments: HD data clustering and multi-channel time-series prediction. Attained results demonstrate that GCC supports the performance of further learning stages, e.g., clustering and prediction, in terms of both system accuracy and data interpretability.

7.1. The cross-correntropy measure

The cross-correntropy is a generalized similarity measure between two arbitrary scalar random variables X and Y, with domain $\mathscr{X}$, defined by [130]:

$$\xi\left(X,Y\right)=\mathbb{E}_{xy}\left\{\kappa_{\xi}\left(x,y\right)\right\}, \tag{7-1}$$

where $x{\in}X$, $y{\in}Y$ and

$$\kappa_{\xi}\left(x,y\right)=\langle\psi(x),\psi(y)\rangle_{\mathscr{C}} \tag{7-2}$$

is a symmetric positive definite kernel, commonly assumed as Gaussian, associated to the nonlinear mapping $\psi:\mathscr{X}\rightarrow\mathscr{C}$ to the RKHS $\mathscr{C}$ [132].

In practice, the joint pdf $p_{XY}(x,y)$ is unknown and only a set of finite data $\{x_n{\in}X, y_n{\in}Y : n{\in}[1,N]\}$ is available, leading to the sample estimator of cross-correntropy as follows:

$$\hat{\xi}\left(X,Y;\{\sigma_n\}\right)=\frac{1}{N}\sum_{n=1}^{N}\kappa_G\left(x_n-y_n;\sigma_n\right), \tag{7-3}$$

where $\sigma_n \in \mathbb{R}^+$ is the Gaussian kernel bandwidth for the n-th pair of samples (commonly $\sigma_1 = \sigma_2 = \ldots = \sigma_N = \sigma$) and κ_G is the well-known Gaussian kernel defined as:

$$\kappa_G\left(x_n - y_n; \sigma_n\right) = \exp\left(\frac{-\|x_n - y_n\|_2^2}{2\sigma_n^2}\right). \tag{7-4}$$

If X and Y are very close to each other, their cross-correntropy value yields the 2-norm distance, while it asymptotically evolves to the 1-norm distance when the variables tend to get apart. Furthermore, cross-correntropy falls to the zero-norm as given variables become very far apart. Among the cross-correntropy properties, the following are the most important [93]:

- Bounded positive definiteness, that is, $0 \leq \xi\left(X, Y\right) \leq 1$ when using a normalized kernel, reaching its maximum at $X = Y$.
- The existence of all even moments estimated from the cross-correntropy difference when using the Gaussian kernel:

  $$\hat{\xi}\left(X, Y; \sigma\right) = \frac{1}{\sqrt{2\pi}\sigma} \sum_{q=0}^{\infty} \frac{(-1)^q}{2^q q!} \mathbb{E}_{xy}\left\{\left(x - y\right)^{2q} / \sigma^{2q}\right\},$$

  since the high-order values decay when σ increases, the second order moment dominates.
- Given the joint pdf $p_{XY}\left(x, y\right)$ of an i.i.d. data sample $\{x_n \in X, y_n \in Y\}$, the value $\hat{\xi}\left(X, Y\right)$ tends to the Parzen estimation of the pdf $\hat{p}_E$, at $e = 0$ $(e \in E)$, where $E = X - Y$ is termed the *error*.
- Since the cross-correntropy depends on the kernel bandwidth σ, it is strictly concave within the range $E \in [-\sigma, \sigma]$.

Moreover, based on the Hilbert space embeddings framework [143, 160], the cross-correntropy functional described in eq. (7-1) can be analyzed as a mean operator in $\mathscr{C}$. Thus, from the expectation operator definition, eq. (7-1) can be rewritten as follows:

$$\xi\left(X, Y\right) = \iint \kappa_\xi\left(x, y\right) p_{XY}(x, y) dx dy. \tag{7-5}$$

In addition, from the Representer theorem [131]:

$$f(x) = \langle f, \psi(x) \rangle_{\mathscr{C}}, \tag{7-6}$$

where $f \in \mathscr{C}$, we can define in $\mathscr{C}$ the mean operator of the joint space as:

$$\langle \mu_{xy}, f \rangle_{\mathscr{C}} = \mathbb{E}_{xy} \{ f(x) \} . \tag{7-7}$$

So, we can clearly see that:

$$\xi (X, Y) = \langle \mu_{xy}, \kappa_{\xi} (x, y) \rangle_{\mathscr{C}}. \tag{7-8}$$

This type of operation is normally not done when we are interested in input data, but it can be important when operating with kernel functions. Then, from eqs. (7-7) and (7-8), we can compute the value of the mean operator μ_{xy} at any point in the domain $\mathscr{X}$ as:

$$\mu_{xy}(y) = \mathbb{E}_{xy} \{ \kappa_{\xi} (X, Y{=}y) \} . \tag{7-9}$$

7.2. Generalized cross-correntropy measure (GCC)

Note that the cross-correntropy in eq. (7-1) measures the similarity between X and Y by mapping both of them to the same RKHS $\mathscr{C}$. However, it would be interesting to define a cross-correntropy-based function that allows to relate the random variables X and Y when each of them is mapped to a different RKHS. In this sense, let us denote X and Y as two random variables with domain $\mathscr{X}$ and $\mathscr{Y}$, respectively. Likewise, let $\varphi : \mathscr{X} \to \mathscr{F}$ and $\phi : \mathscr{Y} \to \mathscr{G}$ be two nonlinear mapping functions associated to the positive definite kernels κ_X and κ_Y, respectively, such that:

$$k^X_{xx'} = \langle \varphi(x), \varphi(x') \rangle_{\mathscr{F}} \tag{7-10}$$

$$k^Y_{yy'} = \langle \phi(y), \phi(y') \rangle_{\mathscr{G}}. \tag{7-11}$$

Here, a Generalized cross-correntropy-(GCC) measure between the random variables X and Y is introduced as follows:

$$\Xi(X, Y) = \mathbb{E}_{xy,x'y'} \left\{ \kappa_{\Psi} \left(k^X_{xx'}, k^Y_{yy'} \right) \right\} , \tag{7-12}$$

where κ_{Ψ} is a positive definite kernel associated to the nonlinear mapping $\Psi : (\mathscr{F} \times \mathscr{F}) \times (\mathscr{G} \times \mathscr{G}) \to \mathscr{A}$, ($\mathscr{A}$ is a RKHS). Furthermore, given a set of finite data $\{x_n \in X, y_n \in Y : n \in [1, N]\}$ and employing the well-known Gaussian kernel, a sample estimator of the GCC yields:

$$\hat{\Xi} (X, Y; \{\sigma_{nn'}\}) = \frac{1}{N^2} \sum_{n=1}^{N} \sum_{n'=1}^{N} \kappa_G \left(k^X_{nn'} - k^Y_{nn'}; \sigma_{nn'} \right), \tag{7-13}$$

where

$$k^X_{nn'} = \kappa_G\left(x_n - x'_n; \sigma_X\right) \tag{7-14}$$
$$k^Y_{nn'} = \kappa_G\left(y_n - y'_n; \sigma_Y\right), \tag{7-15}$$

and $\sigma_X, \sigma_Y, \sigma_{nn'} \in \mathbb{R}^+$.

Therefore, the localized similarity described in eq. (7-13) allows revealing the main relations between X and Y taking advantage of different RKHS-based representations. Indeed, the higher the GCC value, the higher the similarity between the mappings of the random variables in $\mathscr{F}$ and $\mathscr{G}$.

As aforementioned, the cross-correntropy measure can be interpreted as a mean operator in $\mathscr{C}$ (see eq. (7-8)). So, it would be interesting to find a similar interpretation for the introduced GCC. Regarding this, given $f \in \mathscr{F}$, $g \in \mathscr{G}$ and based on the Representer theorem, the expectation of $f(x)g(y)$ can be expressed as an inner product:

$$\mathbb{E}_{xy}\left\{f(x)g(y)\right\} = \mathbb{E}_{xy}\left\{\langle f, \varphi(x)\rangle_{\mathscr{F}}\langle g, \phi(y)\rangle_{\mathscr{G}}\right\} = \mathbb{E}_{xy}\left\{\langle f \otimes g, \varphi(x) \otimes \phi(y)\rangle_{\mathscr{F}\otimes\mathscr{G}}\right\}, \tag{7-16}$$

where $\otimes$ stands for the tensor product operator and $\mathscr{F} \otimes \mathscr{G}$ is also a RKHS. Moreover, eq. (7-16) can be rewritten as follows:

$$\mathbb{E}_{xy}\left\{f(X)g(Y)\right\} = \langle f \otimes g, C_{XY}\rangle_{\mathscr{F}\otimes\mathscr{G}}, \tag{7-17}$$

where the term

$$C_{XY} = \mathbb{E}_{xy}\left\{\varphi(x) \otimes \phi(y)\right\} \tag{7-18}$$

is the uncentered cross-covariance operator [46, 140]. Since the uncentered cross-covariance operator is only determined by the joint probability distribution $p_{XY}(x, y)$ on $\mathscr{X} \times \mathscr{Y}$, it can be treated as the joint distribution embedding in $\mathscr{F} \otimes \mathscr{G}$.

Moreover, the norm of C_{XY} can be defined as:

$$\|C_{XY}\|^2_{\mathscr{F}\otimes\mathscr{G}} = \mathbb{E}_{xy,x'y'}\left\{\langle \varphi(x) \otimes \phi(y), \varphi(x') \otimes \phi(y')\rangle_{\mathscr{F}\otimes\mathscr{G}}\right\}, \tag{7-19}$$

based on tensor product properties:

$$\|C_{XY}\|^2_{\mathscr{F}\otimes\mathscr{G}} = \mathbb{E}_{xy,x'y'}\left\{\langle \varphi(x), \varphi(x')\rangle_{\mathscr{F}}\langle \phi(y), \phi(y')\rangle_{\mathscr{G}}\right\}. \tag{7-20}$$

As seen in eqs. (7-12) and (7-20) the introduced GCC is very close to the cross-covariance norm. In fact, when the Gaussian kernel is avoided to computed the relations between x, x'

and y, y' in eq. (7-12), and the basic tensor product is applied, the GCC measure yields to the cross-covariance norm described in eq. (7-20). Thereby:

$$\kappa_{\Psi}\left(\langle\varphi(x),\varphi(x')\rangle_{\mathscr{F}},\langle\phi(y),\phi(y')\rangle_{\mathscr{G}}\right)=\langle\varphi(x),\varphi(x')\rangle_{\mathscr{F}}\langle\phi(y),\phi(y')\rangle_{\mathscr{G}}. \tag{7-21}$$

fig. **7-1** describes the introduced GCC from different RKHSs and its relation with the cross-covariance embedding norm. As seen, the introduced GCC is a localized enhancement of the cross-covariance embedding norm based on the cross-correntropy foundations.

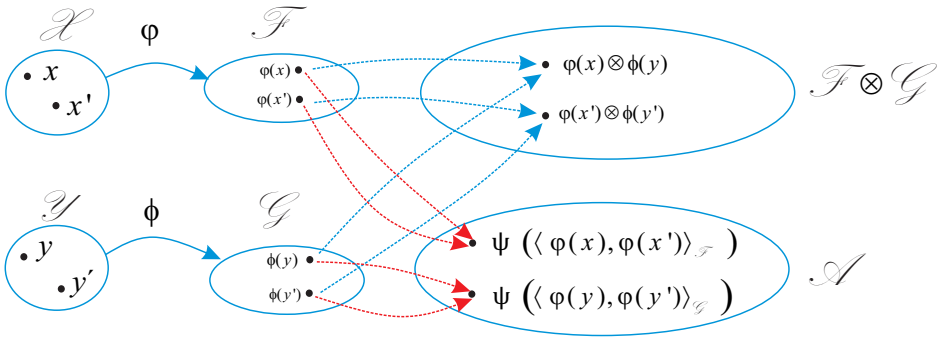


Figure 7-1.: Connection between GCC and cross-covariance operator.

7.3. Relevant data representation based on GCC

Let $\boldsymbol{X}\in\mathbb{R}^{N\times P}$ be a High-Dimensional-(HD) input matrix holding N subject observations and P features. Aiming to capture hidden dependencies among features from each provided subject vector $\boldsymbol{x}_n\in\mathbb{R}^P$ in $\boldsymbol{X}$ $(n\in[1,N])$, the embedding matrix $\boldsymbol{Z}_n\in\mathbb{R}^{L\times J}$ is calculated using an embedding function as follows:

$$\boldsymbol{Z}_n=\vartheta\left(\boldsymbol{x}_n\right), \tag{7-22}$$

where $\vartheta:\mathbb{R}^P\to\mathbb{R}^{L\times J}$ and $L\times J\geq P$.

To find a relevant data representation, each l-th row vector $\boldsymbol{z}_n^l\in\mathbb{R}^J$, $l\in[1,L]$, can be seen as the l-th embedded feature containing the prior user knowledge about all feature relationships. Moreover, to encode the main dynamics in each embedding representation matrix $\boldsymbol{Z}_n$, let us assume that each $\boldsymbol{z}_n^l\in\mathscr{Z}_n$ is a sample datum of the random variable $\mathscr{Z}_n\subset\mathbb{R}^J$. Besides, let $\varphi_n:\mathscr{Z}_n\to\mathscr{F}_n$ be a nonlinear mapping function to the RKHS $\mathscr{F}_n$, which is associated to the positive definite kernel $\kappa_{\mathscr{F}_n}$. In this way, the similarity between $\boldsymbol{Z}_n$ and $\boldsymbol{Z}_{n'}$ $(n,n'\in[1,N])$ is computed based on the introduced GCC measure as follows:

$$\Xi(\boldsymbol{Z}_n,\boldsymbol{Z}_{n'})=\mathbb{E}_{ll'}\left\{\kappa_{\Psi}\left(k_{ll'}^{\mathscr{F}_n}-k_{ll'}^{\mathscr{F}_{n'}}\right)\right\}, \tag{7-23}$$

where $l, l' \in [1, L]$ and

$$k_{ll'}^{\mathscr{F}_n} = \left\langle \varphi_n\left(\boldsymbol{z}_n^l\right), \varphi_n\left(\boldsymbol{z}_n^{l'}\right) \right\rangle_{\mathscr{F}_n} \tag{7-24}$$

$$k_{ll'}^{\mathscr{F}_{n'}} = \left\langle \varphi_{n'}\left(\boldsymbol{z}_{n'}^l\right), \varphi_{n'}\left(\boldsymbol{z}_{n'}^{l'}\right) \right\rangle_{\mathscr{F}_{n'}}. \tag{7-25}$$

Furthermore, employing the Gaussian kernel and the sample estimator to calculate GCC yields:

$$\hat{\Xi}\left(\boldsymbol{Z}_n, \boldsymbol{Z}_{n'}; \{\sigma_{ll'}\}\right) = \sum_{l=1}^{L}\sum_{l'=1}^{L} \kappa_G\left(s_n^{ll'} - s_{n'}^{ll'}; \sigma_{ll'}\right), \tag{7-26}$$

where

$$s_n^{ll'} = \kappa_G\left(\boldsymbol{z}_n^l - \boldsymbol{z}_{n'}^{l'}; \sigma_{Z_n}\right), \tag{7-27}$$

and $\sigma_{ll'}, \sigma_{Z_n} \in \mathbb{R}^+$. Regarding this, the GCC-based pair-wise representation in eq. (7-26) allows extracting salient information from the HD subject set.

In most of the cases, there is a need to rank the input feature set in terms of the contribution of each provided feature in the studied process [159, 15]. To this end, each feature contribution is measured in terms of its stochastic variability extracted from the pair-wise dependencies in each RKHS $\mathscr{F}_n$. Therefore, we carry out a variability-based relevance analysis based on a subspace framework that searches for a projection maximally bearing input information while preserving only those data that contribute most to the GCC measure. Specifically, let $\boldsymbol{S} \in \mathbb{R}^{N \times M}$, be a feature representation matrix, with $M = L(L-1)/2$, holding row vectors:

$$\boldsymbol{s}_n = \{s_n^{ll'}\ l, l' \in 1, L; l < l'\}. \tag{7-28}$$

The stochastic feature set $\boldsymbol{S}$ can be written as a linear combination of $M' < M$ independent basis functions where the minimum mean squared-based error is assumed as the evaluation measure of the subspace-based linear transformation. Thus, a set of orthogonal vectors is estimated so that the M' resulting components can approximate each input feature of $\boldsymbol{S}$, in such a way, that the pair-wise embedded data information is maximally preserved.

Hence, provided the set of features $\{\boldsymbol{\varsigma}_m : m \in [1, M]\}$, where $\boldsymbol{\varsigma}_m \in \mathbb{R}^N$ corresponds to each column of the matrix $\boldsymbol{S}$, the relevance of every $\boldsymbol{\varsigma}_m$ can be measured by computing the following variability vector $\boldsymbol{\rho} \in \mathbb{R}^M$ [5, 111]:

$$\boldsymbol{\rho} = \mathbb{E}_m\left\{|\boldsymbol{v}_m \boldsymbol{\lambda}_m|\right\}, \tag{7-29}$$

where $\boldsymbol{v}_m \in \mathbb{R}^+$ and $\boldsymbol{\lambda}_m \in \mathbb{R}^M$ are respectively the eigenvectors and eigenvalues of the relevant feature covariance matrix estimated as: $\boldsymbol{S}^\top \boldsymbol{S}/M$.

The main assumption behind the relevance measure expressed in eq. (7-29) is that the largest values of ρ_m should point out to the better input attributes since they exhibit higher overall correlations to the estimated principal components. The M' value is fixed as the number of dimensions needed to preserve some percentage of the input data variability. As a result, the calculated relevance vector $\boldsymbol{\rho}$ is employed to rank the pair-wise embedded features dependencies in $\boldsymbol{S}$.

Consequently, we estimate the embedded feature relevance vector $\boldsymbol{\rho}^* \in \mathbb{R}^L$ from $\boldsymbol{\rho}$ as the contribution of each embedded feature over all provided pair-wise feature variability encoded in $\boldsymbol{\rho}$ as follows:

$$\rho_l^* = \mathbb{E}_{l'} \left\{ r_{ll'} : \forall l' \in [1, L] \right\}, \rho_l^* \subset \boldsymbol{\rho} \tag{7-30}$$

where $r_{ll'} = \begin{cases} \rho_m, & \forall m = (l-1)L + l' \\ 0, & l = l'. \end{cases}$.

7.4. Dynamical enhancement of GCC

In practice, the similarity between all feature set may vary throughout a particular domain, e.g., when dealing with spatial and/or time-varying data. In order to include the domain-varying behavior, we can constraint the estimation of the GCC between $\boldsymbol{Z}_n$ and $\boldsymbol{Z}_{n'}$ by using an adaptive learning scheme, resulting in the so-called Dynamical GCC-(DGCC) measure. In other words, we undertake the problem of function estimation upon the online learning framework [78].

So, let $\{(\boldsymbol{Z}_n, d_n); n \in [1, N]\}$ be a sequence of input-output pairs where $d_n \in \mathbb{R}$ is a given output signal that is related to the interactions immersed in $\boldsymbol{Z}_n$, the main goal of adaptive learning systems is to compute the continuous mapping $d_n = \upsilon(\boldsymbol{Z}_n)$ based on the risk minimization analysis, $\upsilon : \mathbb{R}^{L \times J} \to \mathbb{R}$.

Grounded in the introduced GCC approach, the all possible hypothesis $\upsilon \in \mathscr{A}$, where $\mathscr{A}$ is a RKHS equipped with the inner product $\langle \cdot, \cdot \rangle_{\mathscr{A}}$. Consequently, based on the Representer theorem the following properties take place:

- The reproducing property: $\upsilon(\boldsymbol{Z}_n) = \langle \upsilon, \kappa_\Psi(\boldsymbol{Z}_n, \cdot) \rangle_{\mathscr{A}}$.
- $\mathscr{A}$ is the closure of the span of all $\kappa_\Psi(\boldsymbol{Z}_n, \cdot)$, that is, all $\upsilon \in \mathscr{A}$ are linear combinations of kernel functions.

From the aforementioned properties, the problem of estimating the hypothesis online can be viewed as an adaptive filtering task that is a sequential estimator of υ, such that its

domain-variant version υ_n is updated on the basis of the last estimate υ_{n-1} and the current input-output pair $\{\boldsymbol{Z}_n, d_n\}$. In order to deal with the nonlinear relationships between the embedding samples $\boldsymbol{Z}_n$, we may use the kernel adaptive filtering approach (termed Kernel Least Mean Square –KLMS) that aims to exploit the kernel mapping from an input space to a RKHS as follows [92]:

$$\begin{cases} \upsilon_1 = 0 \\ e_n = d_n - \upsilon_{n-1}(\boldsymbol{Z}_n) \\ \upsilon_n = \upsilon_{n-1} + \eta e_n \Xi(\boldsymbol{Z}_n, \cdot) \end{cases} \tag{7-31}$$

where $0 < \eta < 1$ is the filter step size and

$$e_n{=}d_n - \upsilon_{n-1}(\boldsymbol{Z}_n) \tag{7-32}$$

is termed the error. In addition, the continuous mapping is estimated based on the introduced GCC measure described in eq. (7-23) as:

$$\hat{d}_n = \upsilon_{n-1}(\boldsymbol{Z}_n){=}\sum\nolimits_{j=1}^{n-1} \omega_j \hat{\Xi}(\boldsymbol{Z}_n, \boldsymbol{Z}_j; \{\sigma_{ll'}\}), \tag{7-33}$$

being $\omega_j{=}\eta e_j$ the j-th element of the vector $\boldsymbol{\omega}{\in}\mathbb{R}^{T_n}$, where T_n is the filter size at the n-th instant.

Nonetheless, KLMS in eq. (7-31) uses all learned observations to estimate the output of a new input, resulting in a complex function that may lead to over-fitting (not mentioning its high computational load). To cope with this issue, the quantized version of KLMS (termed QKLMS) is considered to get an adequate trade-off between system complexity and accuracy performance [171]. The QKLMS approach aims to discover the main model structure by computing the Euclidean distance on the original input space between a given sample and the codebook. As an alternative, we propose to estimate the similarity between a new sample and the system model taking advantage of the RKHS.

So, assuming that $\boldsymbol{C}_{n-1}{=}\{\boldsymbol{Z}_j : j{\in}[1, T_n]\}$ is the QKLMS system codebook at the $n-1$ iteration with network size T_n, the quantization value $\Upsilon{\in}\mathbb{R}^+$ for a new embeded sample $\boldsymbol{Z}_n$ is estimated in the form:

$$\Upsilon(\boldsymbol{Z}_n, \boldsymbol{C}_{t-1}) = \max_j \hat{\Xi}(\boldsymbol{Z}_n, \boldsymbol{Z}_j; \{\sigma_{ll'}\}), \tag{7-34}$$

for all $\boldsymbol{Z}_j{\in}\boldsymbol{C}_{n-1}$, termed the j-th codeword in $\boldsymbol{C}_{n-1}$.

Then, each $\boldsymbol{Z}_n$ is either merged or not into $\boldsymbol{C}_{n-1}$ in dependence on the yielded comparison to the threshold parameter $\Upsilon(,) \gtrless \gamma{\in}\mathbb{R}^+$. Algorithm 1 develops the proposed DGCC.

Algorithm 1 Dynamical enhancement of GCC

Input: $\boldsymbol{Z}_n \in \mathbb{R}^{L \times J}, d_n \in \mathbb{R}$, $0 < \eta < 1$, $\{\sigma_{ll'} \in \mathbb{R}^+\}$, $\forall l, l' \in [1, L]$, $\gamma \geq 0$

Output: $\hat{d}_n \in \mathbb{R}, \boldsymbol{C}_n, \boldsymbol{\omega}_n \in \mathbb{R}^{T_n}$

; $\boldsymbol{C}_1 = \{\boldsymbol{Z}_1\}, \omega_1 = \{\eta d_1\}$ **while** $\{\boldsymbol{Z}_n, d_n\}$ $(n > 1)$ available **do**

- $\hat{d}_n = \sum_{j=1}^{T_n} \omega_{n-1}^j \hat{\Xi}(\boldsymbol{Z}_n, \boldsymbol{Z}_j; \{\sigma_{ll'}\})$, $\forall \boldsymbol{Z}_j \in \boldsymbol{C}_{n-1}$
- $e_n = d_n - \hat{d}_n$
- $j^* = \arg\max_j \hat{\Xi}(\boldsymbol{Z}_n, \boldsymbol{Z}_j; \{\sigma_{ll'}\})$
- **if** $\Upsilon(\boldsymbol{Z}_n, \boldsymbol{Z}_j) > \gamma$ **then**
  - $\boldsymbol{C}_n = \boldsymbol{C}_{n-1}$; $\omega_{n-1}^{j^*} = \omega_{n-1}^{j^*} + \eta e_n$ $\boldsymbol{\omega}_n = \boldsymbol{\omega}_{n-1}$
- **else**
  - $\boldsymbol{C}_n = \{\boldsymbol{C}_{n-1}, \boldsymbol{Z}_n\}$; $\boldsymbol{\omega}_n = \{\boldsymbol{\omega}_{n-1}, \eta e_n\}$

return

7.5. Experimental set-up

In order to test the capability of the proposed GCC and DGCC approaches for dealing with HD data, two main learning tasks are studied: data clustering and multi-channel time-series prediction. Regarding this, the GCC is employed to extract relevant data representations according to either studied task by fixing the embedding function $\vartheta(\cdot)$ based on the prior user's knowledge. For all provided Gaussian kernel functions, the kernel bandwidth value is fixed based on the introduced KEIPV. fig. 7-2 displays the outline of the proposed GCC-based HD data analysis approach.

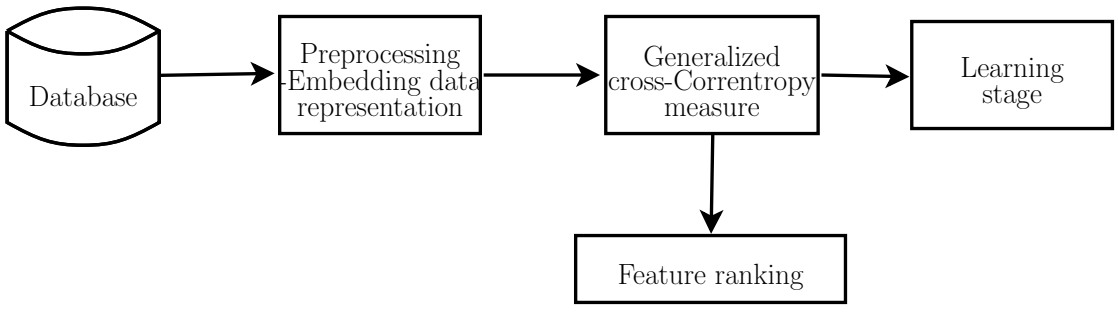


Figure 7-2.: GCC-based HD data analysis main scheme.

For concrete testing, the following three real-world datasets are studied:

- **IXI.** The `IXI` dataset is a brain imaging study holding Magnetic Resonance Images (MRI) from 575 normal subjects that age between 20 and 80 years. Subjects are provided with T1, T2, PD, DTI and angiogram volumes. The image sequences were acquired with three different scanners (Philips 1.5T, Philips 3Tm, and a GE 3T), anonymised and converted to NIFTI format. Additionally, basic demographic information for each subject is included (age, gender, ethnicity, among others). The whole

dataset is freely available online. Here, only the T1 sequences of N=314 subjects (acquired with the GE 3T scanner) were taken into account. T1 sequences are composed of $256 \times 256 \times 150$-sized volumes with a voxel size of $0.9375 \times 0.9375 \times 1.2mm$. Thus, the considered subset is composed of 141 male and 175 female subjects.

– **BCI-FTT.** The Brain Computer Interface (BCI) Foot Tracking Task (FTT) dataset from the *Brain Science Institute* is tested. In this case, some monkeys were trained to reach for food offered by an experimenter using the hand contralateral to the implanted hemisphere while the monkey movement was recorded by an optical motion acquisition system. So, BCI-FTT holds Electrocorticography (ECoG) signals from two Japanese macaques (conditionally designated as *Monkey A* and *Monkey K*). Both monkeys were chronically implanted with electrodes in the subdural space. Particularly, 32 electrodes were implanted in the right hemisphere of *Monkey A*, and 64 electrodes were used in the left hemisphere of *Monkey K*. Brain signals were recorded at $1\,kHz$ while motion signals were captured at $120\,Hz$ [25].

– **MoCap.** The Motion Capture Database (MoCap) was recorded in a laboratory at Carnegie Melon University that contains 12 Vicon infrared MX-40 cameras recording at 120 *Hz* with 4 megapixel-resolution images. The cameras are placed around a rectangular area of $3m \times 8m$ in the center of the room. Subjects wear a black jumpsuit with 38 markers taped on while the infra-red cameras see them. The images that the various cameras pick up are triangulated to get a $3D$ data representation. The subjects are asked to perform several human motion activities, which are captured by the MoCap system. Then, a video in *Biovision Hierarchy* format (BVH) for each motion activity by a given subject is recorded.

7.6. Results and discussion

IXI clustering results. We test the proposed GCC as a suitable feature extraction approach to support clustering tasks on 3D MRIs. In this regard, two preprocessing steps are performed over all the IXI dataset images. Initially, each image is registered to the MNI305 template by an affine transform so that the whole dataset is referenced to the Talairach space [42]. Due to the registering, each volume is re-sampled to $197 \times 233 \times 189$ size. Then, an intensity normalization procedure is performed by scaling each voxel value so that the mean intensity of the white matter is fixed to be 110 [44]. Both preprocessing steps, normalization and registering, are performed with the Freesurfer image analysis suite that is freely available online[1]

[1] `http://surfer.nmr.mgh.harvard.edu/`

We explore two similarity-based image representation techniques to find patient patterns from MRIs. The first one is a baseline where each image voxel is used as feature [1], while the second one uses the proposed GCC to encode MRI pairwise relationships. As regards the voxel-wise approach, given the input matrix $\boldsymbol{X}\in\mathbb{R}^{N\times P}$ (N=314, P=8675289) holding row vectors $\boldsymbol{x}_n\in\mathbb{R}^P$ after column-based concatenation of each MRI volume, we calculate each element of the voxel-wise similarity matrix $\boldsymbol{K}^o\in\mathbb{R}^{N\times N}$ as follows: $k^o_{nn'}=\kappa_G\left(\boldsymbol{x}_n-\boldsymbol{x}'_n;\sigma_v\right)$, with $n,n'\in[1,N]$. In turn, regarding the GCC-based approach, from each input vector $\boldsymbol{x}_n$ the embedding data representation matrix $\boldsymbol{Z}^v_n\in\mathbb{R}^{L_v\times J_v}$ with row vectors $\boldsymbol{z}^v_n\in\mathbb{R}^{J_v}$ is computed according to the considered MRI axis view, namely: Axial, Sagittal, and Coronal, respectively, noted as $v\in\{a,s,c\}$ and with $L_v\in\{197,233,189\}$ and $J_v\in\{44037,37233,45901\}$. So, $\boldsymbol{z}^{l_v}_n$ is the vector concatenation of the $l_v\in[1,L_v]$ slice in the n-th MRI according to the v-th axis view. Further, we impose smooth variations between adjacent MRI slices to encode each Inter-Slice Similarity (ISS) along the v-th axis as: $\kappa_G(\boldsymbol{z}^l_n-\boldsymbol{z}^{l'}_n;\sigma_{Z_n})$. Thus, the matrix $\boldsymbol{S}^v\in\mathbb{R}^{N\times M_v}$ with row vectors computed as in eq. (7-28), where $M_v\in\{19306,27028,17766\}$. Moreover, a relevant feature ranking is carried. Particularly, the relevance of each MRI slice is computed as in eq. (7-29) for each provided axis view.

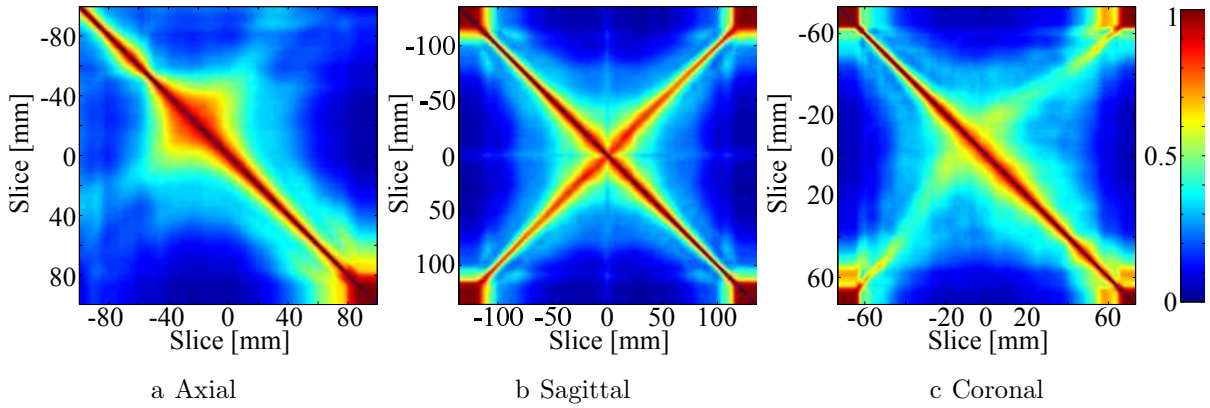
In this way, three relevant ranking vectors $\hat{\rho}^v\in\mathbb{R}^{L_v}$ are computed as in eq. (7-30). figs. 7.5a to 7.3c show a concrete MRI example illustrating the embedded feature relationships for all the three views. As seen in figs. 7.5a to 7.5b displaying their corresponding estimated ISS, the red corner patches keep the MRI edges with no content, i.e., the background. Moreover, regarding the Sagittal view (see fig. 7.5b) it exhibits symmetry with respect to the anti-diagonal, being clear that such representation is able to keep the head sagittal symmetry. Although the latent phenomenon is the same for all ISSs, each of them are providing a different view of the patient brain structures.

Besides, figs. 7.4a to 7.4c exhibit the attained relevant feature ranking for the studied MRI data. Therefore, due to the ISS-based kernel shapes and the relevant vectors varies accordingly to the brain structure distribution, we infer that proposed approach suitably characterizes head shapes. Aside, note that the introduced ISS allows obtaining a feature representation space that resides in a lower dimension in comparison with the original voxel-wise representation ($M_v\ll P$).

Since the proposed ISS allows representing high-dimensional image information along every axis, a Marginal Image Similarity (MIS) matrix $\boldsymbol{K}^v\in\mathbb{R}^{314\times 314}$ can be estimated for each axis v based on the proposed GCC by assuming that each ISS encodes a different RKHS (see eq. (7-26)). Afterward, to explore joint image similarity matrix $\boldsymbol{K}\in\mathbb{R}^{N\times N}$ through all axes, we put forward a joint affinity measure of MIS matrices by introducing the following convex combination of MIS-based kernels:

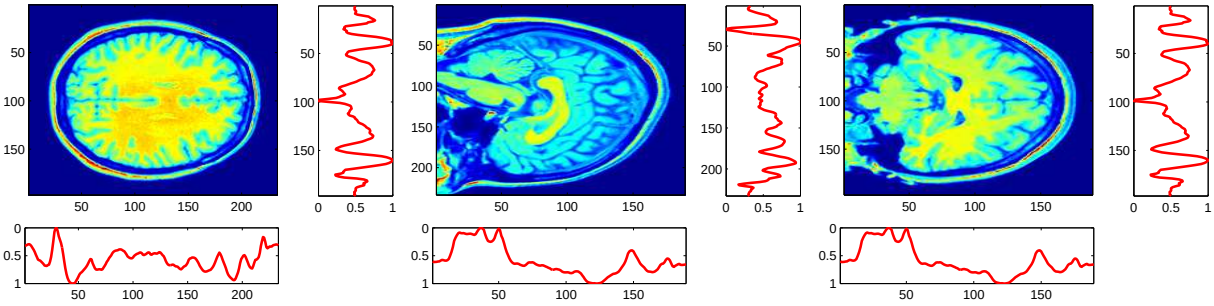
$$k_{nn'}=\mathbb{E}_v\left\{k^v_{nn'}\right\}v. \tag{7-35}$$

figs. 7.5a and 7.5b present the attained values of MRI similarity by using both the voxel-



a Axial b Sagittal c Coronal

Figure 7-3.: Embedded relationships of the MRI feature set - ISS.



a Axial view b Sagittal view c Coronal view

Figure 7-4.: Ranking by relevance of the MRI feature set. –Normalized relevant vectors.

wise approach and the introduced GCC-based strategy, respectively. The MRIs are sorted by gender and age *IXI* categories. Particularly, four classes are considered: *Adult Male (AM)*, *Senior Male (SM)*, *Adult Female (AF)*, and *Senior Female (SF)*. The *Adult* label corresponds to age values between 20 and 50 years, and the *Senior* label to age values higher than 50 years. As seen, the gender and age slots are highlighted for the GCC-based approach supplying evidence about some possible MRI patterns. Furthermore, to identify visually the MRI clusters a 3D low-dimensional space is computed from each similarity matrix based on the well-known Kernel Principal Component Analysis (KPCA) algorithm [132], as seen in figs. 7.5c and 7.5d. Regarding the age as a demographic category, by visual inspection it can be seen that the proposed methodology can unfold the age and the gender better than the baseline decomposition results. Additionally, a larger dispersion is shown in older subjects than on younger ones. This finding can be due to a larger head shape dispersion on older humans, which is according to anatomical head knowledge. In fact, it is known that brain anatomy is steady in middle age humans, while change (gray matter volume diminishes) faster on older humans [1].

For verifying the above-mentioned statements, a k-nearest neighbor classifier is trained

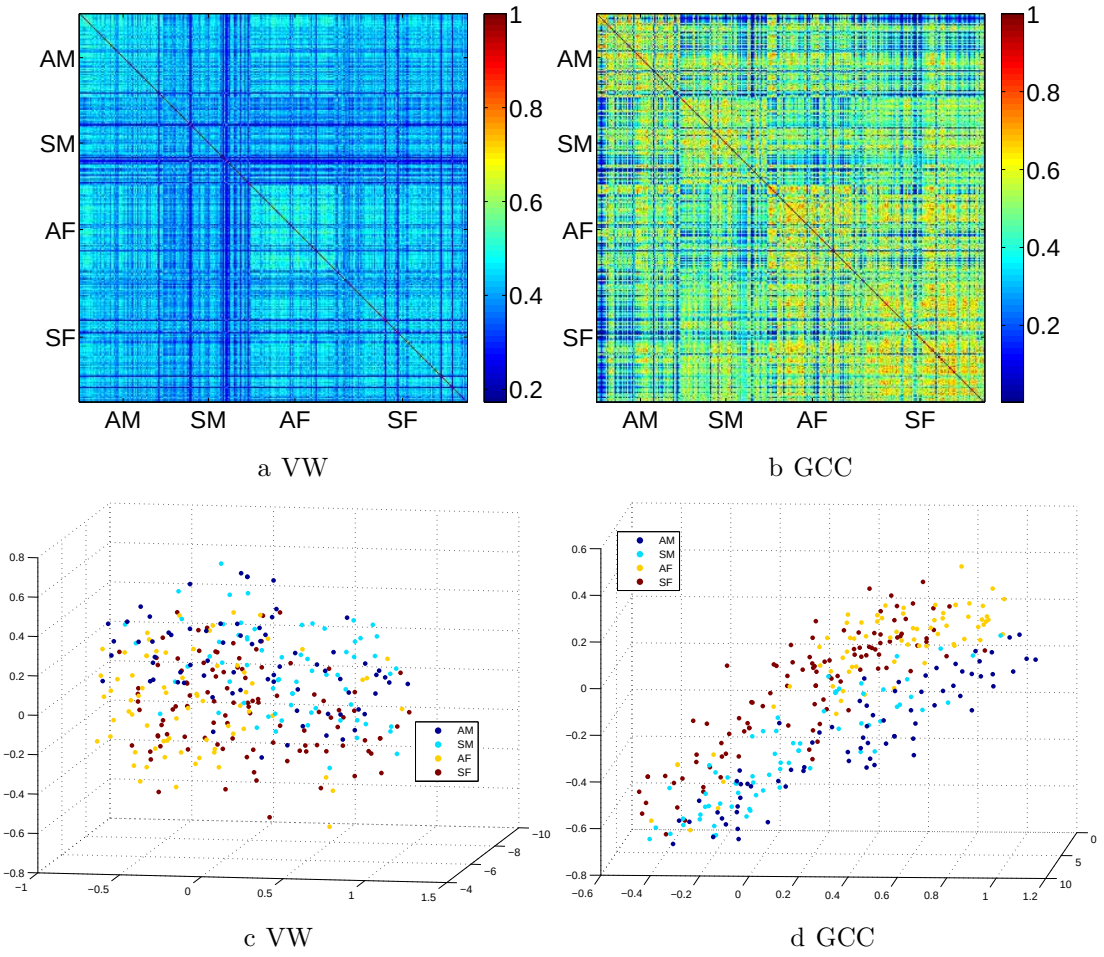


Figure 7-5.: MRIs visual inspection results. **1st row** computed similarities. **2nd row** KPCA-based embeddings.

using each similarity-based induced distance. Tables **7-1** to **7-2** present the attained confusion matrices for a leave-one-out validation scheme. It can be seen that our proposal attains higher classification accuracy than the baseline method. Indeed, presented GCC allows to identify similar brain structures by analyzing the joint axis view relationships. Regarding voxel-wise approach, it can obtain an acceptable gender discrimination. However, it is not able to distinguish age categories. So, complex brain structures, e.g., those related to patient age variations, can not be properly encoded by the HD MRI voxel representation.

BCI-FTT clustering results. We test the proposed GCC approach as a tool to support neural decoding systems from ECoG signals. So, the considered neural decoding system based on GCC can be summarized in following three main stages: *i)* preprocessing of brain activity signals aiming to avoid the influence of artifacts and to highlight frequency bands of interest, *ii)* GCC-based feature extraction stage facilitating the analysis of neural states, and

Table **7-1**.: Voxel-wise-based MRIs discrimination confusion matrix [%]

Category	*Adult Male*	*Senior Male*	*Adult Female*	*Senior Female*
Adult Male	83.33	50.74	06.94	04.85
Senior Male	04.17	28.36	0	05.82
Adult Female	08.33	04.48	83.33	28.15
Senior Female	04.17	16.42	09.72	61.16

Gender accuracy 88.54[%]
Gender-Age accuracy 64.33[%]

iii) the learning stage allowing to find neural patterns that are related to a given stimulation, e.g., the intention of the movement.

For the concrete testing, the BCI-FTT dataset is considered, and the system performance is validated in terms of data interpretability. With respect to the preprocessing stage, all ECoG signals are filtered using a low-pass filter with cutoff frequency of $250\,Hz$ and statistically normalized using a common average reference (CAR). Besides, the motion channels are interpolated to get every ECoG signal lasting the same length. We explicitly consider the experiment carried out with *Monkey A* on November 27th, 2008 that is set up with 56 channels. Namely, 32 ECoG signals plus 8 XYZ-dimensional motion channels, e.g., 24 motion signals. Thereby, the time sequence ranges from $400\,s$ to $500\,s$ is studied. Then, to decode the intention of movement from the brain activity a sliding window of $0.5\,s$ with 90 % of overlapping is applied on each ECoG channel.

Afterward, we explore two different feature representation approaches: The former uses each ECoG segment as input sample after applying the sliding window. So, we obtain the input matrix $\boldsymbol{X}\in\mathbb{R}^{N\times P}$ (N=2084, P=3200) holding row vectors $\boldsymbol{x}_n\in\mathbb{R}^{P}$ after concatenation of the channel segments. Then, the Independent Channel Representation (ICR) similarity matrix $\boldsymbol{K}^{o}\in\mathbb{R}^{N\times N}$ is calculated as: $k^{o}_{nn'}=\kappa_G\left(\boldsymbol{x}_n-\boldsymbol{x}_{n'};\sigma\right)$. The latter approach employs the introduced GCC strategy to find relevant dependencies among channels for supporting the neural decoding task. Thus, from each input vector $\boldsymbol{x}_n$ the embedding data representation matrix $\boldsymbol{Z}_n\in\mathbb{R}^{L\times J}$ (L=32, J=125) with row vectors $\boldsymbol{z}_n\in\mathbb{R}^{J}$ is computed. Here, the vector $\boldsymbol{z}_n$ is a given ECoG channel segment from the provided sliding window. In addition, inter-channel dependencies are encoded into the n-th time window as described in eq. (7-28) to calculate the relevant feature matrix $\boldsymbol{S}\in\mathbb{R}^{N\times M}$ (M=496).

figs. 7.6a and 7.6b present two relevant feature vectors as inter-channel dependencies for both movement and resting monkey conditions, respectively. As seen, few channels are

Table **7-2**.: GCC-based MRIs discrimination confusion matrix [%]

Category	*Adult Male*	*Senior Male*	*Adult Female*	*Senior Female*
Adult Male	70.83	17.91	04.17	05.82
Senior Male	16.67	64.18	0	02.91
Adult Female	05.56	0	69.44	09.71
Senior Female	06.94	17.91	26.39	81.55

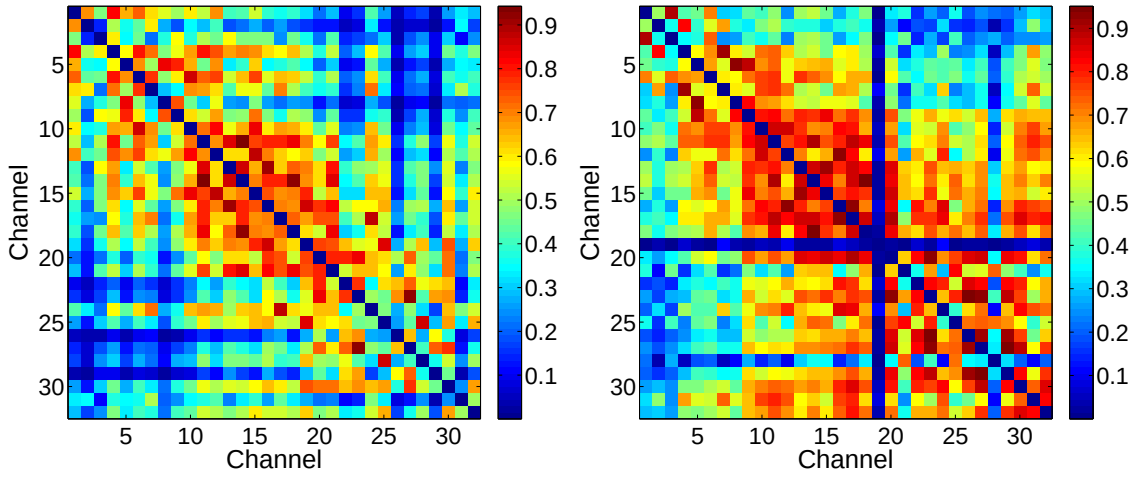
Gender accuracy 91.08[%]
Gender-Age accuracy 72.61[%]

influenced by the stimulus for the movement condition in comparison with the resting one. Thus, a particular region of the brain is activated when the monkey focuses on the FTT. The latter remark can be supported by the estimated relevant channel ranking. In this case, the relevant vector $\hat{\boldsymbol{\rho}} \in \mathbb{R}^L$ (L=32) is quantized in four states based on a the well-known *kmeans* algorithm: high relevance, moderate relevance, low relevance, and no relevant (see fig. **7-7**).

Furthermore, to test the proposed neural decoding system in terms of data interpretability, we try, from extracted ECoG features, to infer the already known activity, e.g., monkey motion. To this end, the a GCC-based similarity matrix $\boldsymbol{K} \in \mathbb{R}^{N \times N}$ is computed based on Eq. (7-26), and a spectral clustering algorithm is carried out [7]. For the sake of comparison, the spectral clustering algorithm is also applied to the ICR-similarity matrix. Although there are some motion activities that may be stated from the provided data collection, we only consider for the sake of simplicity the visual interpretability of two motion activities: *hand-to-food* and *hand-to-mouth*, thereby, we fix the number of clusters as two.

figs. 7.8a and 7.8b present the attained ECoG similarity matrix by using both the ICR and the GCC approaches, respectively. In addition, figs. 7.8c and 7.8d both plots show the 3D movement coordinates of the monkey left hand along the considered time sequence with the labels obtained from using spectral clustering on the ICR and GCC-based similarity matrices. As seen, there are mainly two stages: *peaks* and *steady state*; each one relating to strong and smooth changes, respectively. Therefore, we can infer that one cluster is associated to the *hand-to-food* activity while another to the *hand-to-mouth*. Besides, proposed GCC measure is able to identify better the *hand-to-food* cluster in comparison to the ICR benchmark. Hence, the GCC can be validated visually by the estimated clusters in the FTT database.

BCI-FTT prediction results. To test the proposed dynamical enhancement of GCC, termed DGCC, on neural decoding tasks, we predict the monkey movement position based on the



a Monkey in movement at 415 s

b Monkey in resting at 420 s

Figure 7-6.: Examples of FTT inter-channel dependencies

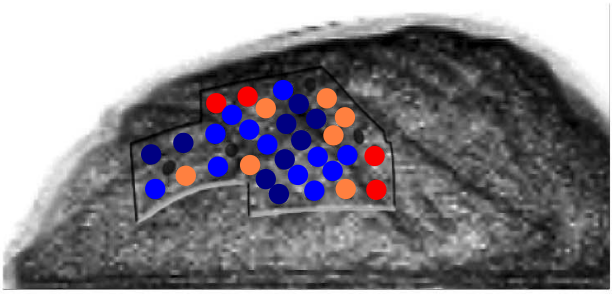
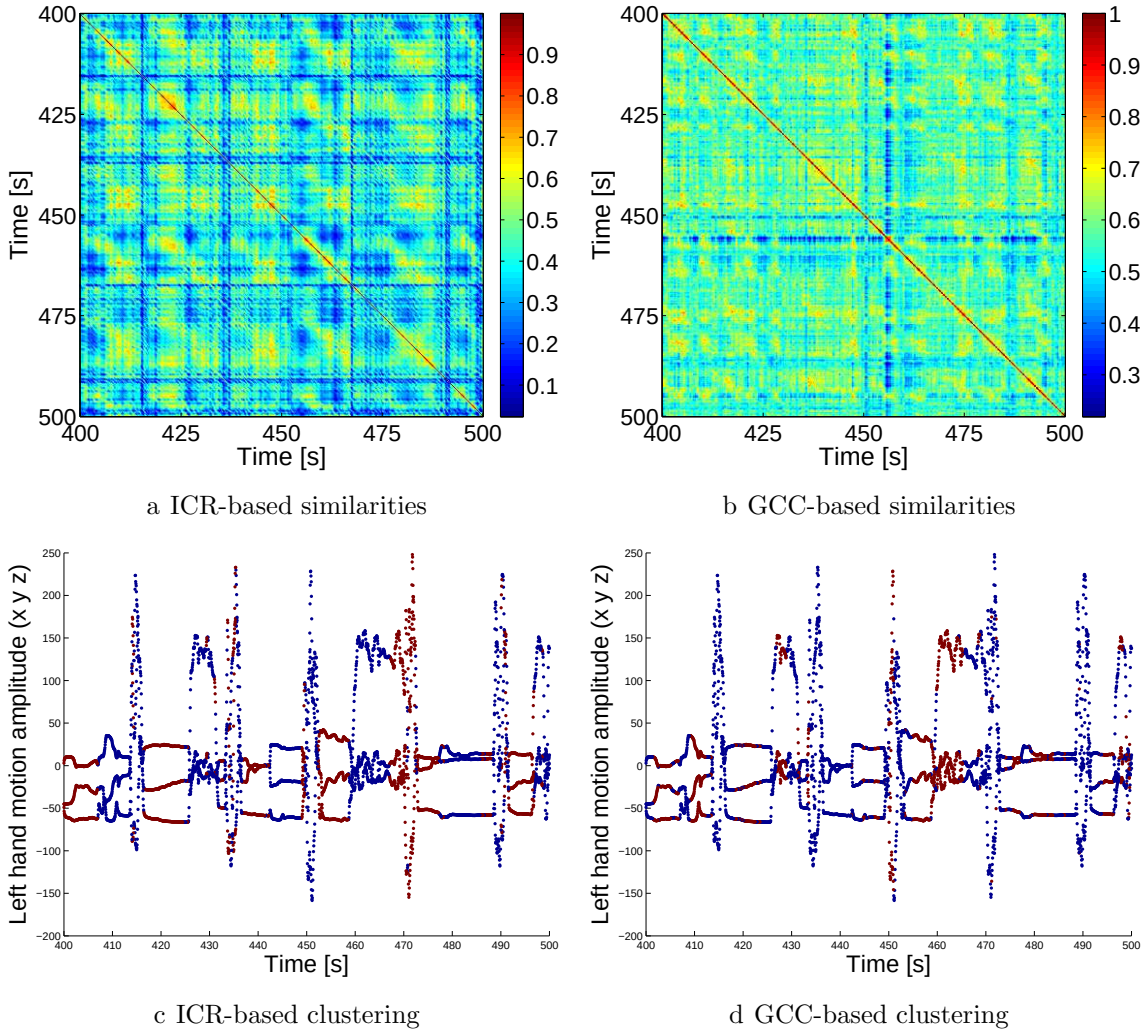


Figure 7-7.: FTT relevant channels based on GCC. •high relevance, •moderate relevance, •low relevance, •no relevant.

baseline experiments that are widely carried out for the FTT database [25]. The ECoG signals are low-pass filtered with the 500 Hz cutoff frequency. All recordings holding motion marker locations are down-sampled to 20 Hz. Moreover, both ECoG and motion channels are centered at CAR. Again, to predict the intention of movement a sliding window is applied to the ECoG recordings as in the BCI-FTT clustering experiments. Provided the input data, we try to predict each movement trajectory accomplished by the primate. based on the proposed DGCC.

For concrete testing, the filter step-size η and the quantization size γ values are empirically adjusted as 0.81 and 0.6, respectively. For the sake of comparison, the kernel adaptive filtering approach proposed in [171] (termed Quantized Kernel Least Mean Square- QKLMS) is also tested by using the ICR-based similarities. Moreover, the step and the quantization size values are fixed as in DGCC. To perform prediction accuracy, we calculate the relative error between the observed and the predicted trajectories of 30 randomly selected ECoG



a ICR-based similarities

b GCC-based similarities

c ICR-based clustering

d GCC-based clustering

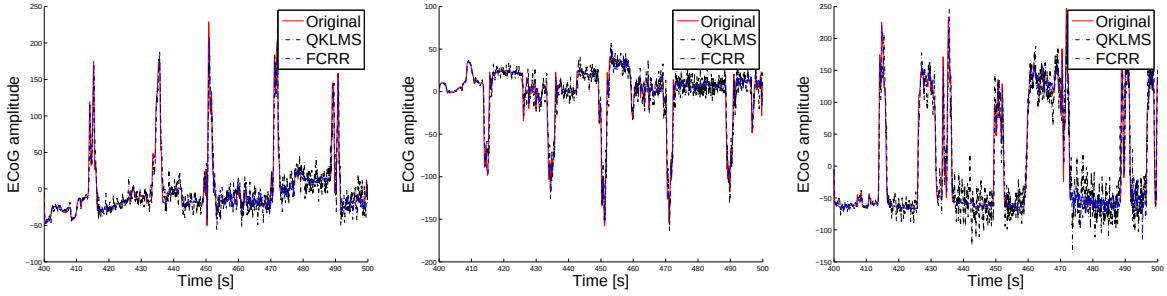
Figure 7-8.: FTT clustering results. •*hand-to-food.* •*hand-to-mouth.*

segments of 100 s each. Particularly, the last 40% of the signals are considered to measure the prediction accuracy.

figs. 7.9a to 7.9c show the predicted motion channels for the Left Hand (LHND) coordinates. As seen the proposed DGCC can reveal relevant ECoG dynamics properly for predicting the output signal. In contrast, the QKLMS is not able to estimate the monkey movement accurately and noisy predictions are obtained. Now, figs. 7.10a and 7.10b present the statistical analysis regarding the relative error of the studied ECoG segments for both the QKLMS and the DGCC algorithms, respectively. In this case, the analysis is carried out for each monkey joint, namely: Left Shoulder (LSHO), Left Elbow (LELB), Left Wrist (LWRI), Right Shoulder (RSHO), Right Elbow (RELB), Right Wrist (RWRI), Right Hand (RHND), and LHND. Particularly for each box, the central mark is the median. The edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data

points not considered outliers, and the outliers are plotted individually.

The results show how our algorithm improves the prediction task for all monkey joints in comparison to the baseline. This fact can be attributed to the proposed DGCC, since, relevant relationships among cortical regions are encoded properly along time, thereby, DGCC allows to suitable predict motion trajectories by adapting a filtering model. In addition, DGCC achieves lower filter complexity in comparison to the QKLMS strategy, in terms of filter size (see figs. 7.10c and 7.10d). Thus, our approach allows revealing the main multi-channel time series dynamics using a compact adaptive learning model.

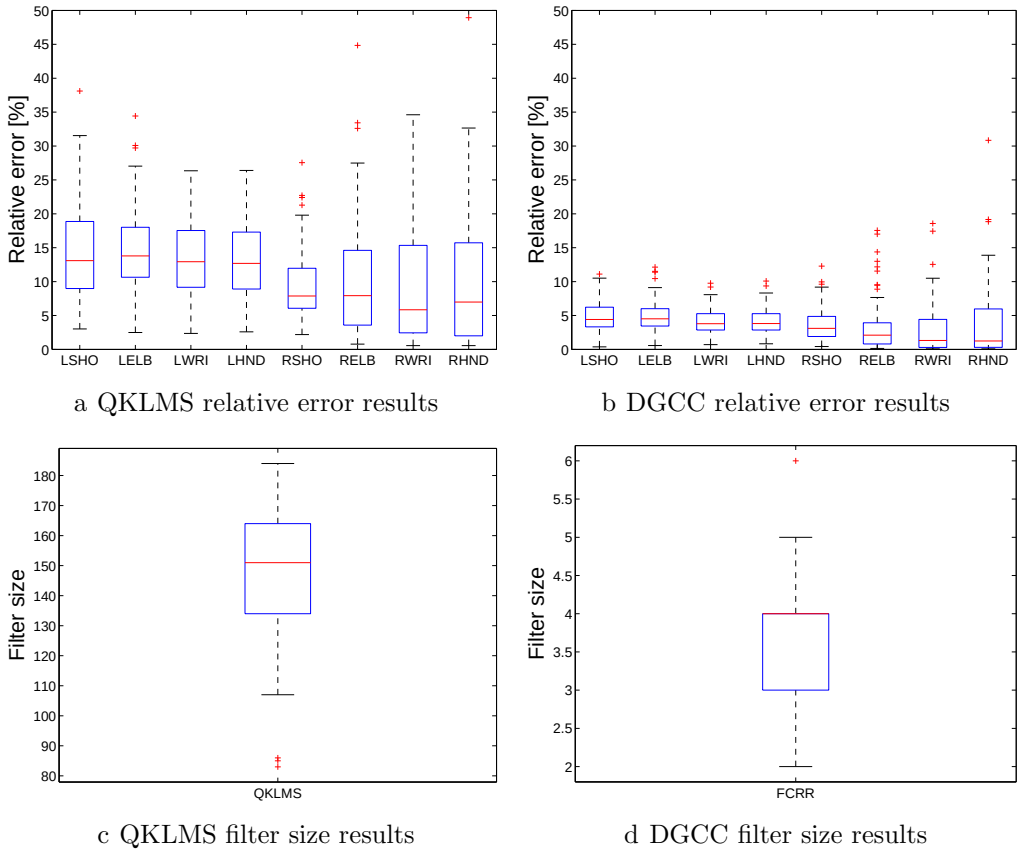


a x coordinate b y coordinate c z coordinate

Figure 7-9.: FTT stimuli prediction (LHND).

MoCap prediction results. In order to test the capability of the proposed approach to finding the main dynamics of multi-channel time-series, a 3D human pose task is studied in the MoCap dataset. For the concrete testing, the following activities are studied: walking (subject 05 video 01), jumping (subject 02 video 04), basketball (subject 06 video 15), and dancing (subject 05 video 11). Here, the XYZ angles between each joint and the Hips are considered to model the 3D subject skeleton. Then, we obtain 2D data by projecting from the 3D MoCap input format into 2D. Provided a MoCap video, an input multi-channel matrix $\boldsymbol{X}\in\mathbb{R}^{N\times P}$ is obtained by applying a sliding window of size $0.5\,s$ with 90% of overlapping, where P=38 $\times$ 2 $\times$ 60=4560 corresponds to the synthesized 2D angle coordinates in the n-th window and N represents the number of estimated windows. In addition, the third coordinate is inferred based on the proposed DGCC approach. Regarding this, from each window $\boldsymbol{x}_n$ the embedding data representation matrix $\boldsymbol{Z}_n\in\mathbb{R}^{L\times J}$ (L=76, J=60) is calculated. Afterward, inter-channel dependencies are encoded as described in eq. (7-28) to estimate the relevant feature matrix $\boldsymbol{S}\in\mathbb{R}^{N\times M}$ (M=2850). Besides, the relevant channel ranking is also computed. Here, a relevance vector $\hat{\boldsymbol{\rho}}\in\mathbb{R}^{38}$ is computed by averaging the 2D coordinates. Again, the *kmeans* algorithm is applied over $\hat{\boldsymbol{\rho}}$ to highlight four categories: high relevance, moderate relevance, low relevance, and no relevant.

Regarding the DGCC free parameters, the filter step size η and the quantization size γ values are empirically adjusted as 0.81 and 0.9, respectively. For the sake of comparison,



a QKLMS relative error results

b DGCC relative error results

c QKLMS filter size results

d DGCC filter size results

Figure 7-10.: FTT stimuli prediction results.

the QKLMS filter is also tested by using the ICR-based similarities as described in the BCI-FTT experiments. Moreover, the step and the quantization size values are fixed as in DGCC. To perform prediction accuracy, we calculate the relative error between the observed and the predicted 3D coordinate in the last 40% of the signals. For evaluating the system robustness against different testing noise conditions, the input data $\boldsymbol{X}$ is also corrupted with additive white Gaussian noise to get different Signal to Noise Ratio conditions - SNR, namely: SNR=$\{2, 5, 10\}[dB]$.

In figs. 7.11a to 7.11l some visual results are shown for the studied videos (free of noise conditions). As seen in figs. 7.11b, 7.11e, 7.11h and 7.11k the different relationships among channels for each activity are highlighted. Overall, there are some channels which share high similarity according to the given human pose, encoding relevant joint dependencies of the studied movement. Thus, the relevance of each joint varies according to the human dynamics (see figs. 7.12a to 7.12d). Likewise, temporal relationships are highlighted when analyzing the GCC-based matrices (see figs. 7.11c, 7.11f, 7.11i and 7.11l).

In this work, after visual inspection of fig. 7.11c, one can notice how the cyclic pattern of the walking movement is inferred by the proposed approach. Similar behavior is observed in the jumping video, as seen in fig. 7.11f that shows the temporal relationships among relevant

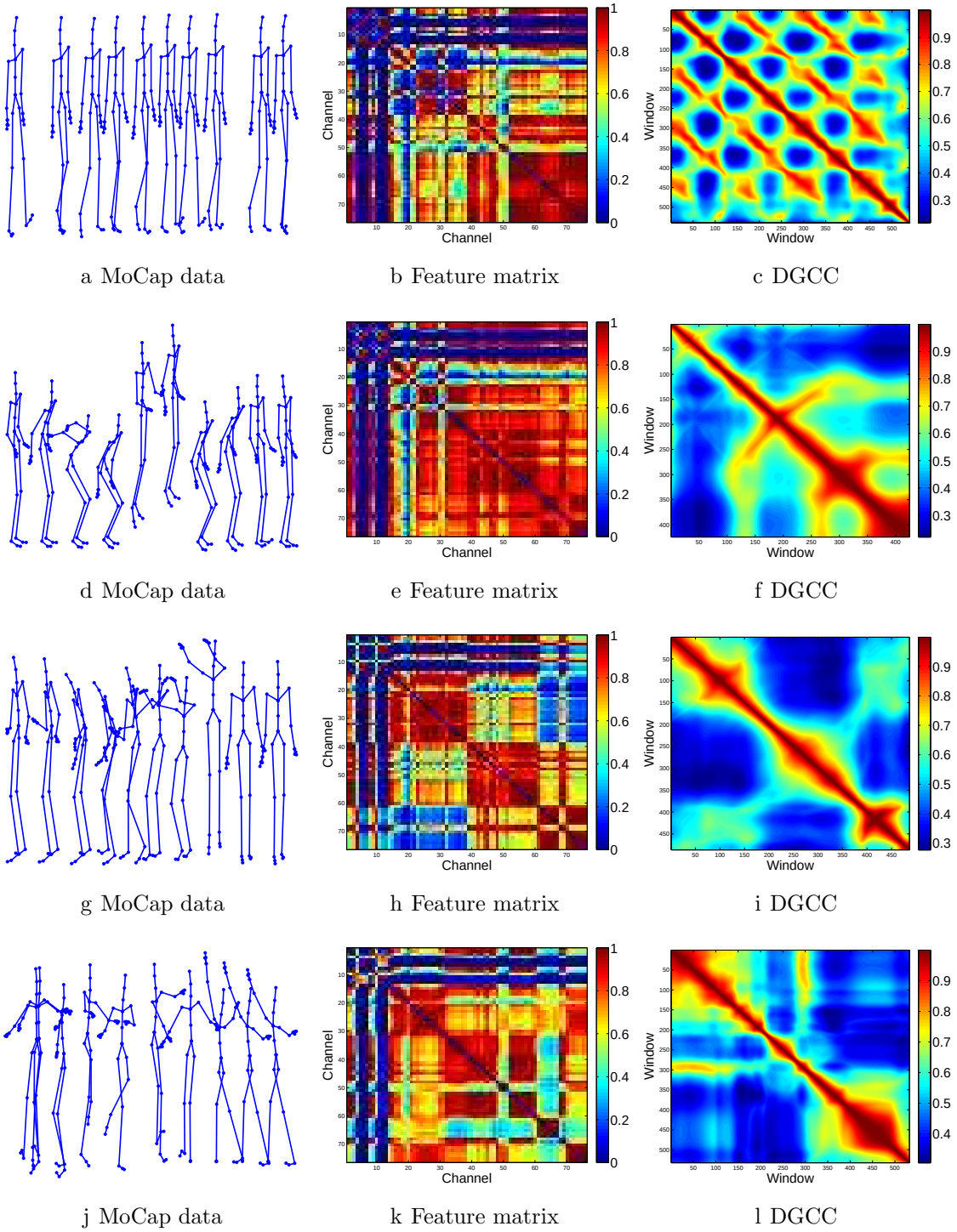
channel interactions, e.g., the relevant embedded features. Here, the DGCC discovers two jumping cycles and one static state at the end of the video since the subject jumps twice and then just stands for a while. Regarding more complex activities, in fig. 7.11i tree main functional assemblies can be seen for the basketball video. In this case, after visual inspection of the GCC-based matrix no cyclic connections (circular shapes) are obtained. In turn, attained results reveal similar behavior for the dancing video.

On the other hand, in most of the cases, proposed DGCC gets better performance than the baseline QKLMS in terms of obtained relative error results and final filter size as shown in figs. **7-13** and **7-14**. Particularly, in fig. **7-13** note that, on each box, the central mark is the median, the edges of the box are the $25th$ and $75th$ percentiles, the whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually. Particularly, the studied approaches attain an acceptable performance on the walking video in terms of the relative error, since this is a smooth activity, not requiring a complex function to infer it (see figs. 7.13a and 7.13e). Yet, DGCC estimates a suitable learning function supplying the lowest number of samples, as seen in fig. 7.14a. Similar results are obtained again for jumping, where DGCC outperforms the QKLMS algorithm in terms of relative error and final filter size as seen in figs. 7.13b, 7.13f and 7.14b.

In addition, the DGCC outperforms again the baseline algorithm for basket and dancing videos, as seen in Figs. figs. 7.13c to 7.13c, figs. 7.13g to 7.13h, and figs. 7.14c to 7.14d. This advantage can be explained based on the proposed relevant representation, which highlights the joint relationships as an assembly, before applying the recursive kernel-based filter. Thus, the temporal structure and the statistical dependencies among channels are suitable discovered. Furthermore, it is important to note that due to the employed similarity measure in DGCC, the system performance is notoriously better in comparison to the QKLMS approach for low SNR conditions.

7.7. Summary

A generalized cross-correntropy measure is introduced to analyze HD samples in learning systems. The proposed GCC approach is a data-driven measure that reveals relevant dependencies among HD samples, which includes both the distribution and the structure of the input process by representing the samples using different RKHSs. Besides, a relevant feature ranking criteria is proposed on GCC to highlight the contribution of each HD feature into the studied process. Furthermore, aiming to incorporate the user prior knowledge regarding the domain-varying behavior of the input data, an adaptive learning constraint is imposed in GCC, yielding the DGCC framework. Also, the required GCC free parameters, e.g., the kernel parameters, are fixed based on the introduced KEIVP. The proposed GCC is an extension of the well-known cross-Correntropy measure based on Hilbert space embeddings. Regarding this, the GCC measure is interpreted from kernel methods as well as from an ITL points of view. Our approach is tested as a data representation tool to analyze HD samples.



a MoCap data b Feature matrix c DGCC

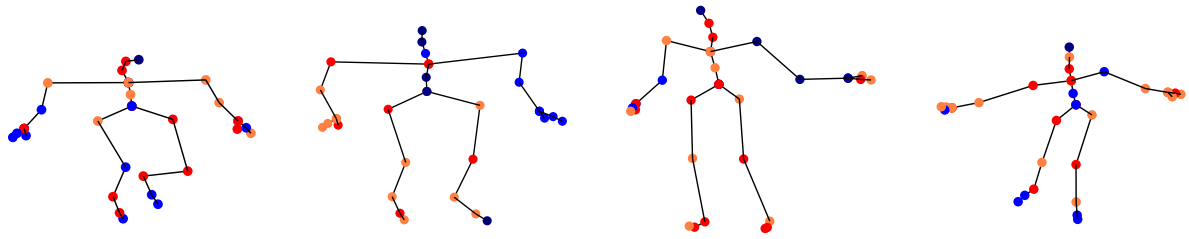
d MoCap data e Feature matrix f DGCC

g MoCap data h Feature matrix i DGCC

j MoCap data k Feature matrix l DGCC

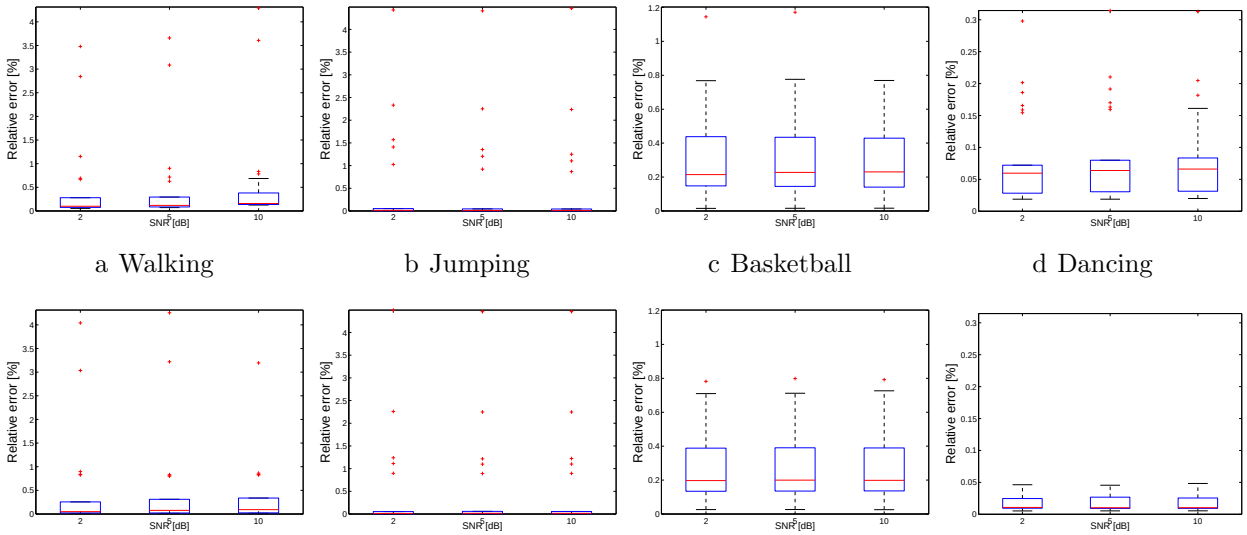
Figure 7-11.: MoCap data analysis based on CRR. **1st row**: walking. **2nd row**: jumping. **3th row**: basketball. **4th row**: dancing .

Particularly, two learning tasks are studied: data clustering and multi-channel time-series prediction. Overall, the proposed GCC allows encoding HD data structures and highlights



a Walking b Jumping c Basketball d Dancing

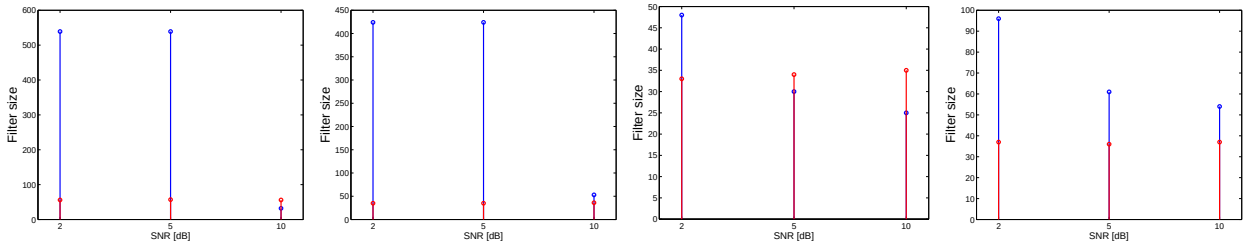
Figure 7-12.: MoCap relevant joints based on GCC. ●high relevance, ●moderate relevance, ●low relevance, ●no relevant.



a Walking b Jumping c Basketball d Dancing

e Walking f Jumping g Basketball h Dancing

Figure 7-13.: MoCap data prediction results. **1st row**: QKLMS. **2nd row**: DGCC.



a Walking b Jumping c Basketball d Dancing

Figure 7-14.: MoCap data prediction results - Final filter size. ●QKLMS. ●DGCC.

relevant feature dependencies to facilitate further learning stages.

With respect to the clustering task, GCC is applied as an image representation approach

to support 3D MRI analysis. Thus, GCC allows encoding smooth MRI inter-slice variations, which can be related to the brain structure distribution. Then, joint MRI similarities are calculated for enhancing both data interpretability and separability using patient demographic information. Taking into account the attained results over a well-known MRI dataset, the proposed approach proved to find the natural inherent distributions of MRIs, namely, age and gender categories. In addition, proposed methodology improves data separability in comparison to the state of the art algorithms based on Voxel-wise MRI representation. So, our proposal is suitable to support MRI clustering and similarity measurement tasks.

Likewise, GCC is tested as a feature extraction stage in a neural decoding task from ECoG recordings of macaques. Then, GCC is employed to infer relevant features from multi-channel time-series. Indeed, each ECoG channel is embedded by using a sliding window to extract relevant channel dependencies. Further, we encode the motion trajectory of the macaques in an FTT. The obtained results in terms of data interpretability show how GCC can infer the main neural state relationships, which are related to motion cyclic patterns: *hand-to-food* and *hand-to-mouth* activities.

Now, regarding the prediction tasks, the introduced DGCC strategy is used to predict the macaque motion trajectories in the FTT database. Again, each ECoG channel is embedded by using a sliding window to extract relevant channel dependencies by using the proposed strategy. Attained prediction results show that our approach outperforms a baseline algorithm in terms of the relative error between the original and the predicted time-series. Moreover, in most of the cases, the model complexity, i.e., the filter size, is lower than the benchmark one. Finally, the DGCC approach is also tested on a well-known MoCap database for tracking 3D human joints, from which some videos are used to track human activities. According to the attained results, our framework provides an adequate alternative for finding functional relevant dependencies as an assembly into multi-channel time-series. In this sense, the system accuracy improves in comparison with the baseline algorithms, which does not consider directly relevant channel dependencies. It is important to note that even when our approach can predict a given output time-series, it is also useful to interpret and to analyze complex relationships into multi-channel time-series visually.

Part IV.

Final remarks

8. Conclusions and future work

8.1. Conclusions

This dissertation highlighted several specific methodologies within a kernel-based representation framework. In this sense, five kernel strategies were proposed to learn automatically relevant data relations in machine learning systems. The introduced approaches naturally lead to data-dependent processing tuned to the particular samples constraints and to the considered learning scenario, including supervised and unsupervised tasks. Besides, the introduced kernel-based relevant data representation framework is tested in some clustering and classification tasks related to biosignal processing and image analysis applications. Overall, attained results demonstrated that proposed approaches allow summarizing and capturing the main input patterns, favoring the learning performance in terms of task accuracy and data interpretability in comparison to state-of-the-art methods. Following, the main concluding remarks regarding each provided representation strategy are described:

- A kernel-based approach that allows revealing the main input sample relations was presented. In fact, the introduced strategy, named KAGP, is able to include both the data statistical distribution and its relevant structures by imposing graph-based sparse constraints. So, KAGP reveals salient data structures to favor further learning stages. Namely, KAGP is a kernel-based graph pruning strategy within a spectral clustering framework which allows enhancing both the local and global data consistencies for a given input similarity matrix. For such a purpose, KAGP learns a kernel matrix based on a CKA-based function to measure the similarity between two kernel matrices, enhancing their local and global consistencies. In this way, our approach takes advantage of an initial guess of the relationships between points to identify relevant connections by encoding then using a compactly supported kernel function. Besides, a regularization-based criterion based on information-theory is introduced as to reach a trade-off between the local and the global consistency preservation during the graph pruning process. For the sake of comparison, KAGP is tested as relevant data representation strategy to support clustering tasks. Attained results showed that KAGP can handle complex data structures, yielding better clustering performance in comparison to the baselines. Moreover, the KAGP promotes learning performance less sensitive to outliers, noisy data, and overlapped groups.

- A new kernel function estimation based on ITL was presented to highlight relevant

pair-wise sample similarities. Our approach, termed KEIPV, estimates a RKHS that spans the widest information force magnitudes among data points. For such a purpose, KEIPV relates different kernel functions with the intrinsic information potential variations into a Parzen-based pdf estimation. Thereby, KEIPV find a RKHS that maximizes the overall information potential variability with respect to the global kernel parameter. As a case of interest, an updating rule for estimating the Gaussian kernel bandwidth parameter is proposed as a function of the forces induced by the distances among samples. The introduced KEIPV was tested on two classical machine learning tasks: classification and clustering, using both synthetic and real-world data. Obtained results show that KEIPV is able to find a suitable RKHS by imposing a statistical regularity constraint, the information protectional variability, which favors data groups separability.

- An entropy-like functional on positive definite matrices based on Renyi's definition was adapted to develop a kernel-based approach that considers the statistical distribution and the salient data structures from information theory-based constraints. Regarding this, the proposed approach, named KEDR, highlights relevant input pattern from Gramm matrices using ITL-based functionals. The introduced KEDR is a data-driven framework for ITL based on infinitely divisible kernel functions. For this, KEDR is a applied as a representation tool that estimates relevant mismatches between high-dimensional and low-dimensional spaces. Furthermore, according to our experiments on both synthetic and real-world datasets, KEDR-based performances are comparable with DR benchmarks in terms of both visual inspection and neighborhood preservation (rank-based criteria).

- A supervised kernel-based representation is introduced as a metric learning approach in a RKHS. In this way, proposed methodology, called SKRA, takes advantage of both the input samples statistical distribution and the user-prior knowledge regarding the studied learning task to highlight the relevant data regularities. So, proposed SKRA incorporates the introduced KEIPV strategy to impose the input data statistical distributions constraints as well as a centered-kernel alignment functional to adapt a linear mapping in a RKHS. As a result, relevant sample dependencies are highlighted by weighting the input features that mostly encode the supervised learning task. The SKRA is tested as a feature selection/extraction strategy in classification problems. Attained results show that our supervised representation outperforms, in most of the cases, state-of-the-art strategies in terms of data interpretability and system accuracy.

- A new generalized cross-correntropy measure named GCC is proposed by taking advantage of different RKHSs. Accordingly, relevant multidimensional dependencies are highlighted automatically by considering the input data domain-varying behavior when available. the GCC measure includes both the distribution and the structure of the

input process by representing the samples using different RKHSs. Besides, a relevant feature ranking criteria is proposed on GCC to highlight the contribution of each input feature into the studied process. The GCC incorporates the KEIPV strategy for imposing the data statistical distributions constraints into each studied RKHS. Besides, GCC approach was enhanced based on an adaptive learning scheme to deal with dynamic systems. The introduced GCC measure is an extension of the well-known cross-correntropy function from a Hilbert space embedding perspective. Our approach is tested in clustering, classification, and prediction applications. Overall, the proposed GCC allows encoding input data structures and highlights relevant feature dependencies to facilitate further learning stages in terms of system accuracy and data interpretability.

8.2. Future work

We have presented a kernel-based representation framework aiming to reveal relevant patterns in machine learning tasks. However, from the achieved theoretical and experimental results, there are still many issues that can be addressed to improve the learning performance in terms of data interpretability and system accuracy. In particular, the following remarks could be of interest for future work approaches:

- In addition, there is room for improvements of the proposed KAGP methodology, namely, some information theory measures can be incorporated to deal with the local and global consistency preservation when dealing with noisy environments. Furthermore, KAGP can be tested as a relevant data representation approach to support dimensionality reduction, classification, and regression tasks. Also, a feature ranking score could be formulated from the KAGP representation to deal with feature selection algorithms [63].

- One important subject for further research is related to the extensions of the proposed KEIVP for different kernel functions. In particular, the multivariate Gaussian kernel, the polynomial, the Laplace, etc., can be adapted according to the introduced information potential variability criterion. Moreover, other functions, besides the variability, could be studied to tune a suitable RKHS. In addition, KEIVP can be tested as a tool to lead with regression and prediction tasks, aiming to reveal the main input regularities that are correlated with the output signals. In fact, KEIVP can be useful the lead with kernel-based online learning scenarios due to the RKHS can be properly adapted within a KEIVP-based strategy [119, 92]. Another research topic of interest is related to the well-known support vector machine (SVM) classifier [132]. The KEIVP could be useful to fix the kernel function in such a machine, besides, the main relations between the kernel-function and the regularization parameter in SVM can be studied into a KEIVP-based method.

- Another interesting line of work involve the extension of the introduced KEDR by employing multi-kernel representations, to deal better with complex data structures [56]. Thus, an ITL-based formulation for Gramm matrices could be proposed from different RKHSs. In addition, an enhancement of the KEDR optimization process, e.g., by taking into account second order derivatives, must be added aiming to deal properly with the non-convex behavior of the KEDR cost functional. Furthermore, an automatic selection of the trade-off parameter in T1KEDR and T2KEDR must be developed. Besides, it would be interesting to investigate the possibility of including shift invariant similarities in KEDR.

- Regarding the SKRA methodology, it would be interesting to develop an automatic strategy to fix properly the embedding dimension. Owing to the linear transformation in SKRA encodes the relevant patterns during the centered alignment procedure, it is necessary to find an embedding space that highlights the whole relevant information of the given learning task. Furthermore, other kernel function can be studied in SKRA to test the flexibility of the introduced SKRA. Besides, other kind of learning tasks can be tested into the SKRA formulation, namely, regression and prediction could be of interest. Here, an online learning scheme could be considered to reveal the temporal evolution of the relevant features [78].

- With respect to the GCC proposal it would be interesting to test the proposed representation in different machine learning tasks, e.g., dimensionality reduction and regression. Moreover, a feature selection strategy could be developed from GCC. Finally, information-theory measures can be incorporated into GCC to find relevant relationships among embedded features and to derive new theoretical relations between cross-correntropy, GCC, Hilbert space embeddings, and ITL functionals for Gramm matrices [52]. Finally, there is plenty of room of developments regarding the extension of the GCC for several input space. In this sense, kernel tensor representations could be analyzed and incorporated into the GCC approach, as new alternative to deal with multi-modal analysis [136].

9. Academic discussion

9.1. Journal and conference papers

1. Álvarez-Meza, A.M., Garcia-Vega, S., Castellanos-Domínguez. G. Identification of brain activity patterns using kernel-based feature relevance analysis. International Journal of Neural Systems. (Submitted).

2. Álvarez-Meza, A.M., Castellanos-Domínguez. G., Príncipe, J. A generalized cross-correntropy measure form different reproducing Hilbert spaces. IEEE transactions on Pattern Analysis and Machine Learning. (Submitted).

3. Álvarez-Meza, A.M., Castro-Ospina, A.E., Castellanos-Domínguez, G. Automatic graph pruning based on kernel alignment for spectral clustering. *Pattern Recognition Letters*, Elsevier, (Accepted).

4. Álvarez-Meza, A.M., Molina-Giraldo, S., Castellanos-Domínguez, G. Background modeling using object-based selective updating and correntropy adaptation. *Image and vision computing*. (Accepted).

5. L. D. Lopez-Rios, L. X. Arias-Mora, Y. Ricardo-Cespedes, L. F. Velasquez-Martinez A. M. Alvarez-Meza, and G. Castellanos-Dominguez. Kernel-based relevant feature extraction to support motor imagery classification. In *Sympoium of Image, Signal Processing, and Artificial Vision (STSIVA)*, 2015.

6. H. D. Insuasti-Ceballos, J. S. Lopez-Villa, S. Molina-Giraldo, A. M. Alvarez-Meza, and G. Castellanos-Dominguez. Bounding box pruning using background subtraction for high quality labeling in video-based object classification. In *Sympoium of Image, Signal Processing, and Artificial Vision (STSIVA), 2015.*

7. C. E. Arroyave-Gomez, J. F. Montoya-Cardona, S. Molina-Giraldo, A. M. Alvarez-Meza, and G. Castellanos-Dominguez. People detection in video streams using background subtraction and spatial-based scene modeling. In *Sympoium of Image, Signal Processing, and Artificial Vision (STSIVA), 2015.*

8. Álvarez-Meza, A.M., Velásquez-Martínez, L.F., Castellanos-Domínguez. G. "Time-series Discrimination using Feature Relevance Analysis in Motor Imagery Classification". *Neurocomputing*, Elsevier, 151:122-129, 2015.

9. Bron, E., Smits, M., Cárdenas-Peña, D., Álvarez-Meza, A.M., et. al. "Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge". *NeuroImage*, 111:562-579, 2015.

10. L. Velásquez-Martínez, A. Álvarez-Meza, G. Castellanos-Domínguez. Connectivity Analysis of Motor Imagery Paradigm using Short-Time Features and Kernel Similarities. IWINAC, *Lecture Notes in Computer Science*, 2015.

11. J. Martínez-Vargas, A. Álvarez-Meza, G. Castellanos-Domínguez. Single-channel separation between stationary and non-stationary signals using relevant information. IbPRIA, *Lecture Notes in Computer Science*, 2015.

12. D.F. Collazos-Huertas, A. Álvarez-Meza, N. Gaviria-Gómez, G. Castellanos-Domínguez. Kernel-based feature relevance analysis for ECG beat classification. IbPRIA, *Lecture Notes in Computer Science*, 2015.

13. S. García-Vega, A. Álvarez-Meza, G. Castellanos-Domínguez. Time-series prediction based on kernel adaptive filtering with cyclostationary codebooks. IbPRIA, *Lecture Notes in Computer Science*, 2015.

14. Álvarez-Meza, A.M., Molina-Giraldo, S., Castellanos-Domínguez, G. Image and video processing based on kernel representations. LAP Lambert Academic Publishing, 2015.

15. J. Hurtado-Rincon, S. Rojas-Jaramillo, Y. Ricardo-Cespedes, A. M. Alvarez-Meza, and G. Castellanos-Dominguez. Motor imagery classification using feature relevance analysis: and Emotiv-based BCI system. In *Sympoium of Image, Signal Processing, and Artificial Vision (STSIVA)-IEEE* 1-15, 2014.

16. D. F. Collazos-Huertas, A. F. Giraldo-Forero, D. Cárdenas-Peña A. M. Álvarez-Meza, and G. Castellanos-Domínguez. Functional Protein Prediction Using HMM Based Feature Representation and Relevance Analysis. CCBCOL, *Advances in Intelligent Systems and Computing*, Springer Link, 232:71-76, 2014.

17. A. Álvarez-Meza, S. Molina-Giraldo, G. Castellanos-Dominguez. Correntropy-based Adaptive Learning to Support Video Surveillance Systems. *22nd International Conference on Patttern Recognition (ICPR)*, 2590-2595, 2014.

18. D. Cárdenas-Peña, M. Orbes-Arteaga, A. Castro-Ospina, A. Álvarez-Meza, G. Castellanos-Dominguez. A Kernel-based Representation to Support 3D MRI Unsupervised Clustering. *International Conference on Pattern Recognition (ICPR)*, 3203-3208, 2014.

19. J. Martínez-Vargas, C. Castro-Hoyos, A. Álvarez-Meza, C. Acosta-Medina, G. Castellanos-Dominguez, Recursive Separation of Stationary Components by Subspace Projection and Stochastic Constraints. *International Conference on Patttern Recognition (ICPR)*, 3469-3474, 2014.

20. A. Álvarez-Meza, G. Castellanos-Dominguez, J. Príncipe. Functional Relevant Multichannel Kernel Adaptive Filter for Human Activity Analysis. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

21. Álvarez-Meza, A.M., Cárdenas-Peña, D., Castro-Ospina, A., Castellanos-Domínguez. G. Tensor-Product Kernel-based Representation encoding Join MRI View Similarity. In *36th Annual International Conference of the IEEE Engineering in Medicine & Biology Society – EMBC*, 2014.

22. Álvarez-Meza, A.M., Cárdenas-Peña, D., Castellanos-Domínguez. G. MRI Discrimination by inter-slice similarities and kernel-based centered alignment. Bio-inspired intelligence - IWOBI 2014.

23. Castro-Ospina, A., Álvarez-Meza, A.M., Castellanos-Domínguez. G. Compactly supported graph building for spectral clustering. Bio-inspired intelligence - IWOBI 2014.

24. García-Vega, S., Álvarez-Meza, A.M., Castellanos-Domínguez. G. Neural Decoding using Kernel-based Functional Representation of ECoG Recordings. CIARP, *Lecture Notes in Computer Science*, Springer Link, 8827:247-254, 2014.

25. Cárdenas-Peña, D., Álvarez-Meza, A.M., Castellanos-Domínguez. G. Kernel-based Image Representation for Brain MRI Discrimination. CIARP, *Lecture Notes in Computer Science*, Springer Link, 8827:343-350, 2014.

26. Álvarez-Meza, A.M., Castro-Ospina, A., Castellanos-Domínguez. G. Spectral Clustering using Compactly Supported Graph Building. CIARP, *Lecture Notes in Computer Science*, Springer Link, 8827:327:334, 2014.

27. Álvarez-Meza, A.M., Cárdenas-Peña, D., Castellanos-Domínguez. G. Unsupervised Kernel Function Building using Maximization of Information Potential Variability. CIARP, *Lecture Notes in Computer Science*, Springer Link, 8827:335:342, 2014.

28. García-Vega, S., Álvarez-Meza, A.M., Castellanos-Domínguez. G. Estimation of Cyclostationary Codebooks for Kernel Adaptive Filtering. CIARP, *Lecture Notes in Computer Science*, Springer Link, 8827:351:358, 2014.

29. Álvarez-Meza, A. M., Valencia-Aguirre, J., Daza-Santacoloma, G., Acosta-Medina, C. D., and Castellanos-Domínguez, G. Video Analysis based on Multi-Kernel Representation with Automatic Parameter Choice. *Neurocomputing*, Elsevier, 100: 117–126, 2013.

30. A. Álvarez-Meza, L. Velasquez-Martinez, G. Castellanos-Domínguez. Feature relevance analysis supporting automatic motor imagery discrimination in EEG based BCI systems. In *35th Annual International Conference of the IEEE Engineering in Medicine & Biology Society – EMBC*, 2013.

31. L. Velásquez-Martínez, A. Álvarez-Meza, G. Castellanos-Domínguez. Motor Imagery Classification for BCI Using Common Spatial Patterns and Feature Relevance Analysis. IWINAC, *Lecture Notes in Computer Science* 7931:365:374, 2013.

32. S. Molina-Giraldo, J. Carvajal, A. Alvarez-Meza, G. Castellanos-Dominguez. Video Segmentation based on Multi-kernel Learning and Feature Relevance Analysis for Object Classification. ICPRAM, SciTePress, 2013.

33. A. Álvarez-Meza, C. Acosta-Medina, G. Castellanos-Dominguez. Automatic Singular Spectrum Analysis for Time-Series Decomposition. *ESANN proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2013.

34. D. Peluffo-Ordóñez, S. García-Vega, A. Álvarez-Meza, G. Castellanos-Domínguez. Kernel Spectral Clustering for Dynamic Data. CIARP, *Lecture Notes in Computer Science*, Springer Link, 8258:238-245, 2013.

35. S. García-Vega, A. Álvarez-Meza, G. Castellanos-Domínguez. MoCap Data Segmentation and Classification Using Kernel Based Multi-channel Analysis. CIARP, *Lecture Notes in Computer Science*, Springer Link, 8259:495-502, 2013.

36. A. Castro-Ospina, A. Álvarez-Meza, G. Castellanos-Domínguez. Automatic Graph Building Approach for Spectral Clustering. CIARP, *Lecture Notes in Computer Science*, Springer Link, 8258:190-197, 2013.

37. S. Molina-Giraldo, A. Álvarez-Meza, J. García-Álvarez, G. Castellanos-Domínguez. Video Segmentation Framework by Dynamic Background Modelling. ICIAP, *Lecture Notes in Computer Science* 8156:843:852, 2013.

38. Carvajal-Gonzalez, J., Álvarez-Mesa, A. M., and Castellanos-Domínguez, G. Feature Selection by Relevance Analysis for Abandoned Object Classification. In *17th Iberoamerican Conference on Pattern Recognition: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - CIARP*, Lecture Notes in Computer Science, Springer Link, 2012.

39. Valencia-Aguirre, J., Álvarez-Mesa, A. M., Daza-Santacoloma, G., Acosta-Medina, C. D., and Castellanos-Domínguez, G. Human Activity Recognition by Class Label LLE. In *17th Iberoamerican Conference on Pattern Recognition: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - CIARP*, Lecture Notes in Computer Science, Springer Link, 2012.

40. Álvarez-Mesa, A. M., Daza-Santacoloma, G., and Castellanos-Domínguez, G. Biomedical Data Analysis by Supervised Manifold Learning. In *34th Annual International Conference of the IEEE Engineering in Medicine & Biology Society – EMBC*, 2012.

41. Sepúlveda-Cano, L.M., Álvarez-Mesa, A. M., and Castellanos-Domínguez, G. Training using Short–time Features for OSA Discrimination. In *34th Annual International Conference of the IEEE Engineering in Medicine & Biology Society – EMBC*, 2012.

42. Molina, S., Álvarez, A.M., Peluffo, D., Castellanos, G. Image segmentation based on multikernel learning and feature relevance analysis. In *Advances in Artificial Intelligence – IBERAMIA 2012*, Lecture Notes in Computer Science, Springer Link, 2012.

43. Ramírez, D., Molina, S., Álvarez, A.M., Daza, G., Castellanos, G. Kernel Based Hand Gesture Recognition using Kinect Sensor. In *Symposium of Image, Signal Processing, and Artificial Vision (STSIVA), 2012*, Medellín – Colombia.

9.2. Awarded papers

2012: *Geographical World Finalist - Latin America Best Student Paper.* "Biomedical Data Analysis by Supervised Manifold Learning". 34th Annual International Conference of the IEEE Engineering in Medicine Ponencia:Biomedical - EMBC - 2012, San Diego, USA.

2014: *Best award paper.* "Kernel-based image representation for brain MRI discrimination". 19th Iberoamerican Congress on Pattern Recognition - CIARP, 2014, Puerto Vallarta, Mexico.

9.3. Software prototypes

- Gait interface. Certificado de registro de soporte logico - software, 13-38-395, Ministerio del interior (Colombia) 2013.
- EEG signal preprocessing and analysis (Análisis y limpieza de examenes EEG). Certificado de registro de soporte logico - software, 13-47-297, Ministerio del interior (Colombia) 2015.
- REMI - Relevant feature extraction for motor imagery discrimination: A BCI system. *software registration in progress.*
- Kernel-based relevant representation for video-based activity recognition. *software registration in progress.*

Bibliography

[1] P. Aljabar, R. Heckemann, A. Hammers, J. Hajnal, and D. Rueckert, "Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy." *NeuroImage*, vol. 46, no. 3, pp. 726–38, 2009.

[2] A. M. Álvarez, G. Daza, and G. Castellanos, "Biomedical data analysis by supervised manifold learning," in *34th IEEE EMBS Annual International Conference*, 2012.

[3] A. Álvarez-Meza, A. Castro-Ospina, and G. Castellanos-Dominguez, "Spectral clustering using compactly supported graph building," in *CIARP*. Springer, 2014, pp. 327–334.

[4] A. Álvarez-Meza, J. Valencia-Aguirre, G. Daza-Santacoloma, C. Acosta-Medina, and G. Castellanos-Domínguez, "Video analysis based on multi-kernel representation with automatic parameter choice," *Neurocomputing*, no. 0, pp. –, 2012.

[5] A. Alvarez-Meza, L. Velasquez-Martinez, and G. Castellanos-Dominguez, "Time-series discrimination using feature relevance analysis in motor imagery classification," *Neurocomputing*, vol. 151, pp. 122–129, 2015.

[6] A. Álvarez-Meza, J. Valencia-Aguirre, G. Daza-Santacoloma, and G. Castellanos-Domínguez, "Global and local choice of the number of nearest neighbors in locally linear embedding," *Pattern Recognition Letters*, vol. 32, pp. 2171 – 2177, 2011.

[7] C. Alzate and J. Suykens, "Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA," *IEEE Trans. on Pat. Anal. and Mach. Intelligence*, vol. 32, no. 2, pp. 335–347, 2010.

[8] E. Amigo, J. Gonzalo, J. Artiles, and F. Verdejo, "A comparison of extrinsic clustering evaluation metrics based on formal constraints," *Information retrieval*, vol. 12, no. 4, pp. 461–486, 2009.

[9] R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. E. Elger, "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state," *Phys. Rev. E*, vol. 64, p. 061907, Nov 2001. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevE.64.061907

[10] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American mathematical society*, pp. 337–404, 1950.

[11] J. Bae, P. Chhatbar, J. Francis, J. Sanchez, and P. Jose, "Reinforcement learning via kernel temporal difference," in *33th IEEE Annual International Conference of the EMBS*, 2011.

[12] M. Beauchemin, "A density-based similarity matrix construction for spectral clustering," *Neurocomputing*, vol. 151, pp. 835–844, 2015.

[13] L. Belanche, "Developments in kernel design," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. ESANN, 2013, pp. 369–378.

[14] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[15] S. Bhattacharyya, A. Sengupta, T. Chakraborti, A. Konar, and D. Tibarewala, "Automatic feature selection of motor imagery eeg signals using differential evolution and learning automata," *Medical & biological engineering & computing*, vol. 52, no. 2, pp. 131–139, 2014.

[16] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[17] B. Blankertz, G. Dornhege, M. Krauledat, K.-R. MÃ$\frac{1}{4}$ller, and G. Curio, "The non-invasive berlin brain computer interface: Fast acquisition of effective performance in untrained subjects," *NeuroImage*, vol. 37, no. 2, pp. 539 – 550, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811907000535

[18] I. Borg and P. Groenen, *Modern Multidimensional Scaling: Theory and Applications*. Springer, 2005.

[19] A. Brockmeier, J. Choi, E. Kriminger, J. Francis, and J. Principe, "Neural decoding with kernel-based metric learning," *Neural Computation*, vol. 26, pp. –, 2014.

[20] A. J. Brockmeier *et al.*, "Information-theoretic metric learning: 2-d linear projections of neural data for visualization," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*. IEEE, 2013, pp. 5586–5589.

[21] K. Bunte, S. Haase, M. Biehl, and T. Villmann, "Stochastic neighbor embedding (sne) for dimension reduction and visualization using arbitrary divergences," *Neurocomputing*, vol. 90, pp. 23–45, 2012.

[22] M. A. Carreira-Perpiñán, "The elastic embedding algorithm for dimensionality reduction." in *ICML*. Omnipress, 2010, pp. 167–174.

[23] O. Carrera-Leon, J. M. Ramirez, V. Alarcon-Aquino, M. Baker, D. D'Croz-Baron, and P. Gomez-Gil, "A motor imagery bci experiment using wavelet analysis and spatial patterns feature extraction," in *Engineering Applications (WEA), 2012 Workshop on*. IEEE, 2012, pp. 1–6.

[24] J. Carvajal, A. Álvarez, and G. Castellanos, "Feature selection by relevance analysis for abandoned object classification," in *17th Iberoamerican Congress on Pattern Recognition, image analysis, computer vision and applications – CIARP*, 2012.

[25] Z. C. Chao., Y. Nagasaka., and N. Fujii., "Long-term asynchronous decoding of arm motion using electrocorticographic signals in monkeys," *Frontiers in neuroengineering*, vol. 3, p. 3, 2010.

[26] B. Chen, Y. Zhu, J. Hu, and J. C. Principe, "A variable step-size sig algorithm for realizing the optimal adaptive fir filter," *International Journal of Control, Automation, and Systems*, vol. 9, pp. 1049–1055, 2011.

[27] G. Chen, S. A. Jaradat, N. Banerjee, T. S. Tanaka, M. S. Ko, and M. Q. Zhang, "Evaluation and comparison of clustering algorithms in analyzing es cell gene expression data," *Statistica Sinica*, vol. 12, no. 1, pp. 241–262, 2002.

[28] W. Chen and G. Feng, "Spectral clustering with discriminant cuts," *Knowledge-Based Systems*, vol. 28, pp. 27–37, 2012.

[29] C. Christoforou, R. Haralick, P. Sajda, and L. C. Parra, "Second-order bilinear discriminant analysis," *The Journal of Machine Learning Research*, vol. 11, pp. 665–685, 2010.

[30] M. M. Churchland, M. Y. Byron, J. P. Cunningham, L. P. Sugrue, M. R. Cohen, G. S. Corrado, W. T. Newsome, A. M. Clark, P. Hosseini, B. B. Scott *et al.*, "Stimulus onset quenches neural variability: a widespread cortical phenomenon," *Nature neuroscience*, vol. 13, no. 3, pp. 369–378, 2010.

[31] R. Corralejo, R. Hornero, and D. Álvarez, "Feature selection using a genetic algorithm in a motor imagery based brain computer interface," in *IEEE EMBC*, 2011.

[32] C. Cortes, M. Mohri, and A. Rostamizadeh, "Algorithms for learning kernels based on centered alignment," *The Journal of Machine Learning Research*, vol. 13, pp. 795–828, 2012.

[33] B. R. Cowley, M. T. Kaufman, Z. S. Butler, M. M. Churchland, S. I. Ryu, K. V. Shenoy, and M. Y. Byron, "Datahigh: Graphical user interface for visualizing and interacting with high-dimensional neural activity," *Journal of neural engineering*, vol. 10, no. 6, p. 066012, 2013.

[34] N. Cristianini, J. Kandola, A. Elisseeff, and J. Shawe-Taylor, "On kernel target alignment," in *Innovations in Machine Learning*. Springer, 2006, pp. 205–256.

[35] G. Daza-Santacoloma, J. D. A.-L. no, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruíz, and G. Castellanos-Domínguez, "Dynamic feature extraction: An application to voice pathology detection," *Intelligent Automation and Soft Computing*, 2009.

[36] I. S. Dhillon, Y. Guan, and B. Kulis, "Kernel k-means: spectral clustering and normalized cuts," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 551–556.

[37] ——, *A unified view of kernel k-means, spectral clustering and graph cuts*. Citeseer, 2004.

[38] D. L. Donoho and C. Grimes, "Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data," 2003.

[39] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Mc Graw Hill, 2000.

[40] L. Duque Muñoz, R. Pinzon Morales, and G. Castellanos Dominguez, "Eeg rhythm extraction based on relevance analysis and customized wavelet transform," in *Artificial Computation in Biology and Medicine*. Springer, 2015, pp. 419–428.

[41] L. Duque Munoz, J. Espinosa Oviedo *et al.*, "Identification and monitoring of brain activity based on stochastic relevance analysis of short-time eeg rhythms," *BioMedical Engineering OnLine*, vol. 13, no. 1, p. 123, 2014. [Online]. Available: http://www.biomedical-engineering-online.com/content/13/1/123

[42] A. Evans, D. Collins, S. R. Mills, E. D. Brown, R. L. Kelly, and T. Peters, "3D statistical neuroanatomical models from 305 MRI volumes," in *Nuclear Science Symposium and Medical Imaging Conference, 1993., 1993 IEEE Conference Record.*, 1993, pp. 1813–1817.

[43] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta, "A survey of kernel and spectral methods for clustering," *Pattern recognition*, vol. 41, no. 1, pp. 176–190, 2008.

[44] B. Fischl, D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, A. Montillo, N. Makris, B. Rosen, and A. M. Dale, "Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain." *Neuron*, vol. 33, no. 3, pp. 341–55, 2002.

[45] S. Friedland, "Convex spectral functions," *Linear and multilinear algebra*, vol. 9, no. 4, pp. 299–316, 1981.

[46] K. Fukumizu, F. R. Bach, and M. I. Jordan, "Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces," *The Journal of Machine Learning Research*, vol. 5, pp. 73–99, 2004.

[47] D. Gajic, Z. Djurovic, S. Di Gennaro, and F. Gustafsson, "Classification of eeg signals for detection of epileptic seizures based on wavelets and statistical pattern recognition," *Biomedical Engineering: Applications, Basis and Communications*, vol. 26, no. 02, p. 1450021, 2014.

[48] T. Gandhi *et al.*, "A comparative study of wavelet families for {EEG} signal classification," *Neurocomputing*, vol. 74, no. 17, pp. 3051 – 3057, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231211003158

[49] T. Gärtner, P. Flach, and S. Wrobel, "On graph kernels: Hardness results and efficient alternatives," in *Learning Theory and Kernel Machines*. Springer, 2003, pp. 129–143.

[50] M. G. Genton, "Classes of kernels for machine learning: a statistics perspective," *The Journal of Machine Learning Research*, vol. 2, pp. 299–312, 2002.

[51] S. Ghosh-Dastidar *et al.*, "Principal component analysis-enhanced cosine radial basis function neural network for robust epilepsy and seizure detection," *Biomedical Engineering, IEEE Transactions on*, vol. 55, no. 2, pp. 512–518, Feb 2008.

[52] L. G. S. Giraldo, M. Rao, and J. C. Principe, "Measures of entropy from data using infinitely divisible kernels," *CoRR*, pp. –1–1, 2012.

[53] T. Gneiting, "Compactly supported correlation functions," *Journal of Multivariate Analysis*, vol. 83, no. 2, pp. 493–508, 2002.

[54] M. Gönen and E. Alpaydin, "Multiple kernel machines using localized kernels," in *4th IAPR International Conference on Pattern Recognition in Bioinformatics*, 2009.

[55] M. Gonen and E. Alpaydin, "Localized multiple kernel regression," in *Proceedings of the 20th International Conference on Pattern Recognition (ICPR)*, 2010, pp. 1425 –1428.

[56] M. Gönen and E. Alpaydın, "Multiple kernel learning algorithms," *The Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011.

[57] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, "Measuring statistical dependence with hilbert-schmidt norms," in *Algorithmic learning theory*. Springer, 2005, pp. 63–77.

[58] T. Hanakawa, I. Immisch, K. Toma, M. A. Dimyan, P. Van Gelderen, and M. Hallett, "Functional properties of brain areas associated with motor execution and imagery," *Journal of Neurophysiology*, vol. 89, no. 2, pp. 989–1002, 2003.

[59] R. He, B.-G. Hu, W.-S. Zheng, and X.-W. Kong, "Robust principal component analysis based on maximum correntropy criterion," *Image Processing, IEEE Transactions on*, vol. 20, no. 6, pp. 1485–1494, 2011.

[60] R. He, W. S. Zheng, and B. G. Hu, "Maximum correntropy criterion for robust face recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 8, pp. 1561–1576, 2011.

[61] R. He, W.-S. Zheng, B.-G. Hu, and X.-W. Kong, "A regularized correntropy framework for robust pattern recognition," *Neural Computation*, vol. 23, no. 8, pp. 2074–2100, 2011.

[62] W. He *et al.*, "A novel emd-based common spatial pattern for motor imagery brain-computer interface," in *Biomedical and Health Informatics (BHI), 2012 IEEE-EMBS International Conference on*. IEEE, 2012, pp. 216–219.

[63] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Advances in neural information processing systems*, 2005, pp. 507–514.

[64] J. Herault, C. Jausions-Picaud, and A. Guerin-Dugue, "Curvilinear component analysis for high-dimensional data representation: I. theoretical aspects and practical use in the presence of noise." in *IWANN (2)*, vol. 1607, 1999, pp. 625–634.

[65] H. Higashi *et al.*, "Common spatio-time-frequency patterns for motor imagery-based brain machine interfaces," *Computational intelligence and neuroscience*, vol. 2013, p. 8, 2013.

[66] G. Hinton and S. Roweis, "Stochastic neighbor embedding," in *Advances in Neural Information Processing Systems 15*. MIT Press, pp. 833–840.

[67] G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Advances in neural information processing systems*, 2002, pp. 833–840.

[68] G. E. Hinton and R. S. Zemel, "Autoencoders, minimum description length, and helmholtz free energy," *Advances in neural information processing systems*, pp. 3–3, 1994.

[69] R. A. Horn and C. R. Johnson, *Topics in matrix analysis*. Cambridge: Cambridge University Press, 1991.

[70] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*. John Wiley & Sons, 2004, vol. 46.

[71] Z. Iscan, Z. Dokur, and T. Demiralp, "Classification of electroencephalogram signals with combined time and frequency features," *Expert Systems with Applications*, vol. 38, no. 8, pp. 10 499 – 10 505, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417411003162

[72] A. K. Jain and M. H. Law, "Data clustering: A user's dilemma," in *Pattern Recognition and Machine Intelligence*. Springer, 2005, pp. 1–10.

[73] F. Jamaloo *et al.*, "Discriminative csp sub-bands weighting based on dslvq method in motor imagery based bci," *Journal of Medical Signals and Sensors*, vol. 5, no. 3, pp. 26–31, 2015.

[74] R. Jenssen, J. Principe, and T. Eltoft, "Information cut and information forces for clustering," in *Neural Networks for Signal Processing, 2003. NNSP'03. 2003 IEEE 13th Workshop on*, Sept 2003, pp. 459–468.

[75] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.

[76] F. Jordan, M. Bach, and F. Bach, "Learning spectral clustering," *Advances in Neural Information Processing Systems*, vol. 16, pp. 305–312, 2004.

[77] G. Kimeldorf and G. Wahba, "Some results on tchebycheffian spline functions," *Journal of mathematical analysis and applications*, vol. 33, no. 1, pp. 82–95, 1971.

[78] J. Kivinen, A. J. Smola, and R. C. Williamson, "Online learning with kernels," *IEEE Trans. on Signal Processing*, vol. 100, no. 10, pp. 1–11, 2010.

[79] E. Kreyszig, *Introductory functional analysis with applications*. wiley New York, 1989, vol. 81.

[80] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, 1964.

[81] S. P. Kumar, N. Sriraam, P. Benakop, and B. Jinaga, "Entropies based detection of epileptic seizures with artificial neural network classifiers," *Expert Systems with Applications*, vol. 37, no. 4, pp. 3284–3291, 2010.

[82] G. R. G. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan, and W. S. Noble, "A statistical framework for genomic data fusion," *Bioinformatics*, vol. 20, no. 16, pp. 2626–2635, 2004.

[83] J. A. Lee, E. Renard, G. Bernard, P. Dupont, and M. Verleysen, "Type 1 and 2 mixtures of kullback-leibler divergences as cost functions in dimensionality reduction based on similarity preservation," *Neurocomputing*, vol. 112, no. 0, pp. 92 – 108, 2013.

[84] J. A. Lee and M. Verleysen, *Nonlinear dimensionality reduction*. Springer, 2007.

[85] ——, "Quality assessment of dimensionality reduction: Rank-based criteria," *Neurocomputing*, vol. 72, no. 7, pp. 1431–1443, 2009.

[86] ——, "On the role and impact of the metaparameters in t-distributed stochastic neighbor embedding," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 337–346.

[87] ——, "Two key properties of dimensionality reduction methods," in *Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on*. IEEE, 2014, pp. 163–170.

[88] J. A. Lee, A. Lendasse, and M. Verleysen, "Curvilinear distance analysis versus isomap." *ESANN*, vol. 2, pp. 185–192, 2002.

[89] ——, "Nonlinear projection with curvilinear distances: Isomap versus curvilinear distance analysis," *Neurocomputing*, vol. 57, pp. 49–76, 2004.

[90] S. Lemm *et al.*, "Bci competition 2003-data set iii: probabilistic modeling of sensorimotor μ rhythms for classification of imaginary hand movements," *Biomedical Engineering, IEEE Transactions on*, vol. 51, no. 6, pp. 1077–1080, 2004.

[91] L. Li, A. J. Brockmeier, J. S. Choi, J. T. Francis, J. C. Sanchez, and J. C. Príncipe, "A tensor-product-kernel framework for multiscale neural activity decoding and control," *Computational intelligence and neuroscience*, vol. 2014, p. 2, 2014.

[92] W. Liu, P. Pokharel, and J. Principe, "The kernel least-mean-square algorithm," *Signal Processing, IEEE Transactions on*, vol. 56, no. 2, pp. 543–554, 2008.

[93] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy-: Properties and applications in non-gaussian signal processing," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5286–5298, November 2007.

[94] W. Liu, J. C. Príncipe, and S. Haykin, *Kernel Adaptive Filtering: A Comprehensive Introduction*. John Wiley & Sons, Inc., 2010.

[95] Y. Lu, L. Wang, J. Lu, J. Yang, and C. Shen, "Multiple kernel clustering based on centered kernel alignment," *Pattern Recognition*, 2014.

[96] B. Mao, N. Guan, D. Tao, X. Huang, and Z. Luo, "Correntropy induced metric based graph regularized non-negative matrix factorization," in *Security, Pattern Analysis, and Cybernetics (SPAC), 2014 International Conference on*. IEEE, 2014, pp. 163–168.

[97] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *ICCV*, vol. 2, July 2001, pp. 416–423.

[98] J. D. Martínez, D. Cardenas, and G. Castellanos, "Extraction of stationary components in biosignal discrimination," in *34th IEEE EMBS Annual International Conference*, 2012.

[99] J. D. Martínez-Vargas, J. I. Godino Llorente *et al.*, "Time–frequency based feature selection for discrimination of non-stationary biosignals," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, pp. 1–18, 2012.

[100] M. Meila and J. Shi, "A random walks view of spectral segmentation," in *AI and STATISTICS*, 2001.

[101] P. Metzner, L. Putzig, and I. Horenko, "Analysis of persistent non-stationary time series and applications," *Accepted for publication in Communications in applied mathematics and computational science*, 2012.

[102] S. Molina-Giraldo, A. Álvarez-Meza, D. Peluffo-Ordóñez, and G. Castellanos-Domínguez, "Image segmentation based on multi-kernel learning and feature relevance analysis," in *IBERAMIA 2012*. Springer, 2012, pp. 501–510.

[103] R. A. Morejon and J. C. Principe, "Advanced search algorithms for information-theoretic learning with kernel-based estimators." 2004, pp. 874–884.

[104] Y. Mu and B. Zhou, "Non-uniform multiple kernel learning with cluster-based gating functions," *Neurocomputing*, vol. 74, no. 7, pp. 1095 – 1101, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231210004571

[105] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis, "Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators," in *in Advances in Neural Information Processing Systems 18*. MIT Press, 2005, pp. 955–962.

[106] A. R. Naghsh-Nilchi and M. Aghashahi, "Epilepsy seizure detection using eigen-system spectral estimation and multiple layer perceptron neural network," *Biomedical Signal Processing and Control*, vol. 5, no. 2, pp. 147 – 157, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1746809410000054

[107] M. Nascimento and A. De Carvalho, "Spectral methods for graph clustering–a survey," *European Journal of Operational Research*, vol. 211, no. 2, pp. 221–231, 2011.

[108] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library: Coil-100," Department of Computer Science, Columbia University, New York, Tech. Rep., 1996.

[109] A. Ng, M. Jordan, Y. Weiss *et al.*, "On spectral clustering: Analysis and an algorithm," *Adv. in neural information processing systems*, vol. 2, pp. 849–856, 2002.

[110] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an algorithm," *Advances in Neural Information Processing Systems*, vol. 14, 2001.

[111] J. Orozco, S. Murillo, A. Álvarez, J. Arias, E. Trejos, J. Vargas, and G. Castellanos, "Automatic selection of acoustic and non-linear dynamic features in voice," in *INTERSPEECH*, 2011.

[112] A. R. Paiva, I. Park, and J. C. Príncipe, "A reproducing kernel hilbert space framework for spike train signal processing," *Neural Computation*, vol. 21, no. 2, pp. 424–449, 2009.

[113] J. Palanichamy and K. Ramasamy, "A novel feature selection algorithm with supervised mutual information for classification," *International Journal on Artificial Intelligence Tools*, vol. 22, no. 04, p. 1350027, 2013.

[114] I. M. Park, S. Seth, A. Paiva, L. Li, and J. Principe, "Kernel methods on spike train space for neuroscience: a tutorial," *Signal Processing Magazine, IEEE*, vol. 30, no. 4, pp. 149–160, 2013.

[115] E. Parzen, *Statistical inference on time series by Hilbert space methods.* Stanford Univ., 1959.

[116] D. H. Peluffo-Ordóñez, J. A. Lee, and M. Verleysen, "Recent methods for dimensionality reduction: A brief comparative analysis," in *22th European Symposium on Artificial Neural Networks, ESANN 2014, Bruges, Belgium, April 23-25, 2014*, 2014. [Online]. Available: http://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2014-170.pdf

[117] J. W. Pillow, Y. Ahmadian, and L. Paninski, "Model-based decoding, information estimation, and change-point detection techniques for multineuron spike trains," *Neural Computation*, vol. 23, no. 1, pp. 1–45, 2011.

[118] R. Pokharel, S. Seth, and J. C. Principe, "Mixture kernel least mean square," in *Neural Networks (IJCNN), The 2013 International Joint Conference on.* IEEE, 2013, pp. 1–7.

[119] ——, "Quantized mixture kernel least mean square," in *IJCNN*, 2014.

[120] K. Polat and S. GÃ$\frac{1}{4}$nes, "Classification of epileptiform {EEG} using a hybrid system based on decision tree classifier and fast fourier transform," *Applied Mathematics and Computation*, vol. 187, no. 2, pp. 1017 – 1026, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0096300306012380

[121] J. C. Principe, *Information theoretic learning: Rényi's entropy and kernel perspectives.* Springer, 2010.

[122] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

[123] I. Rejer, "Genetic algorithms for feature selection for brain computer interface," *International Journal of Pattern Recognition and Artificial Intelligence*, 2015.

[124] G. Rodríguez and P. J. García, "Automatic and adaptive classification of electroencephalographic signals for brain computer interfaces," *Medical systems*, vol. 36, no. 1, pp. 51–63, 2012.

[125] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[126] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on.* IEEE, 1994, pp. 138–142.

[127] J. W. Sammon, "A nonlinear mapping for data structure analysis," *IEEE Transactions on computers*, vol. 18, no. 5, pp. 401–409, 1969.

[128] L. G. Sanchez Giraldo and J. C. Principe, "Information theoretic learning with infinitely divisible kernels," *arXiv preprint arXiv:1301.3551*, 2013.

[129] ——, "Information theoretic learning with infinitely divisible kernels," *arXiv preprint arXiv:1301.3551*, 2013.

[130] I. Santamaría, P. Pokharel, and J. Principe, "Generalized correlation funtion: definition, properties, and application to blind equalization," *IEEE Trans. on Signal Processing*, vol. 54, no. 6, pp. 2187–2197, 2006.

[131] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *Computational learning theory.* Springer, 2001, pp. 416–426.

[132] B. Scholkopf and A. J. Smola, *Learning with Kernels.* Cambridge, MA, USA: The MIT Press, 2002.

[133] B. Scholkopft and K.-R. Mullert, "Fisher discriminant analysis with kernels," *Neural networks for signal processing IX*, vol. 1, p. 1, 1999.

[134] L. M. Sepúlveda, A. M. Álvarez, and G. Castellanos, "Training using short-time features for osa discrimination," in *34th IEEE EMBS Annual International Conference*, 2012.

[135] J. Shi and J. Malik, "Normalized cuts and image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 8, pp. 888–905, 2000.

[136] M. Signoretto, L. De Lathauwer, and J. A. Suykens, "A kernel-based framework to tensorial data analysis," *Neural networks*, vol. 24, no. 8, pp. 861–874, 2011.

[137] B. W. Silverman, *Density estimation for statistics and data analysis.* CRC press, 1986, vol. 26.

[138] A. Singh and J. Principe, "Kernel width adaptation in information theoretic cost functions," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, March 2010, pp. 2062–2065.

[139] L. Song, J. Bedo, K. M. Borgwardt, A. Gretton, and A. Smola, "Gene selection via the bahsic family of algorithms," *Bioinformatics*, vol. 23, no. 13, pp. i490–i498, 2007.

[140] L. Song, J. Huang, A. Smola, and K. Fukumizu, "Hilbert space embeddings of conditional distributions with applications to dynamical systems," in *Proceedings of the 26th Annual International Conference on Machine Learning.* ACM, 2009, pp. 961–968.

[141] L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt, "Feature selection via dependence maximization," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 1393–1434, 2012.

[142] V. Srinivasan, C. Eswaran, Sriraam, and N, "Artificial neural network based epileptic detection using time-domain and frequency-domain features," *Journal of Medical Systems*, vol. 29, no. 6, pp. 647–660, 2005.

[143] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, G. R. Lanckriet, and B. Schölkopf, "Kernel choice and classifiability for rkhs embeddings of probability distributions." in *NIPS*, 2009, pp. 1750–1758.

[144] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[145] Y. Tang and D. Durand, "A tunable support vector machine assembly classifier for epileptic seizure detection," *Expert systems with applications*, vol. 39, no. 4, pp. 3925–3938, 2012.

[146] M. Tangermann, K.-R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. R. Müller-Putz *et al.*, "Review of the bci competition iv," *Frontiers in neuroscience*, vol. 6, 2012.

[147] A. Teixeira, A. Matos, and L. Antunes, "Conditional renyi entropies," *Information Theory, IEEE Transactions on*, vol. 58, no. 7, pp. 4273–4277, July 2012.

[148] A. Teixeira, A. Tome, M. Bohm, C. Puntonet, and E. Lang, "How to apply nonlinear subspace techniques to univariate biomedical time series," *Instrumentation and Measurement, IEEE Transactions on*, vol. 58, no. 8, pp. 2433–2443, Aug 2009.

[149] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, p. 2319, 2000.

[150] M. Torki, A. Elgammal, and C. S. Lee, "Learning a joint manifold representation from multiple data sets," in *Proceedings of the 20th International Conference on Pattern Recognition (ICPR)*, 2010, pp. 1068 –1071.

[151] A. T. Tzallas, M. G. Tsipouras, D. Fotiadis *et al.*, "Epileptic seizure detection in eegs using time–frequency analysis," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 13, no. 5, pp. 703–710, 2009.

[152] E. D. Übeyli, "Decision support systems for time-varying biomedical signals: Eeg signals classification," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2275–2284, 2009.

[153] R. Unnikrishnan, C. Pantofaru, and M. Hebert, "A measure for objective evaluation of image segmentation algorithms," in *CVPR Workshops. IEEE Computer Society Conference on*. IEEE, 2005, pp. 34–34.

[154] J. Valencia, A. Álvarez, G. Daza, C. Acosta, and G. Castellanos, "Human activity recognition by class label lle," in *17th Iberoamerican Congress on Pattern Recognition, image analysis, computer vision and applications – CIARP*, 2012.

[155] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, p. 85, 2008.

[156] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski, "Information retrieval perspective to nonlinear dimensionality reduction for data visualization," *The Journal of Machine Learning Research*, vol. 11, pp. 451–490, 2010.

[157] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[158] X. Wang, C. Yang, and J. Zhou, "Clustering aggregation by probability accumulation," *Pattern Recognition*, vol. 42, no. 5, pp. 668–675, 2009.

[159] X. Wang and K. K. Paliwal, "Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition," *Pattern recognition*, vol. 36, no. 10, pp. 2429–2439, 2003.

[160] Y. Xu, "Kernel bayes rule," *Journal of Machine Learning Research*, vol. 14, 2013.

[161] M. Yamada, W. Jitkrittum, L. Sigal, E. P. Xing, and M. Sugiyama, "High-dimensional feature selection by feature-wise kernelized lasso," *Neural computation*, vol. 26, no. 1, pp. 185–207, 2014.

[162] P. Yang, Q. Zhu, and B. Huang, "Spectral clustering with density sensitive similarity function," *Knowledge-Based Systems*, vol. 24, no. 5, pp. 621–628, 2011.

[163] X. Yin *et al.*, "A hybrid bci based on eeg and fnirs signals improves the performance of decoding motor imagery of both force and speed of hand clenching," *Journal of neural engineering*, vol. 12, no. 3, p. 036004, 2015.

[164] A. Zajacova, S. Huzurbazar, M. Greenwood, and H. Nguyen, "Long-term bmi trajectories and health in older adults hierarchical clustering of functional curves," *Journal of aging and health*, p. 0898264315584329, 2015.

[165] A. S. Zandi *et al.*, "Automated real-time epileptic seizure detection in scalp eeg recordings using an algorithm based on wavelet packet transform," *Biomedical Engineering, IEEE Transactions on*, vol. 57, no. 7, pp. 1639–1651, 2010.

[166] L. Zelnik-Manor and P. Perona, "Self-Tuning Spectral Clustering," in *Advances in Neural Information Processing Systems 17*, vol. 2, 2004, pp. 1601—-1608.

[167] H. Zhang, M. Genton, and P. Liu, "Compactly supported radial basis function kernels," *Available at www4.stat.ncsu.edu/hzhang/research.html*, 2004.

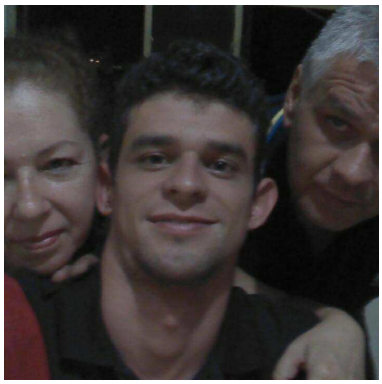
[168] H. Zhang *et al.*, "Bci competition iv–data set i: learning discriminative patterns for self-paced eeg-based motor imagery detection," *Frontiers in neuroscience*, vol. 6, 2012.

[169] X. Zhang, J. Li, and H. Yu, "Local density adaptive similarity measurement for spectral clustering," *Pattern Recognition Letters*, vol. 32, no. 2, pp. 352–358, Jan. 2011. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0167865510003181

[170] S. Zhao, B. Chen, and J. C. Principe, "An adaptive kernel width update for correntropy," in *Neural Networks (IJCNN), The 2012 International Joint Conference on*. IEEE, 2012, pp. 1–5.

[171] S. Zhao, P. Zhu, and P. Jose, "Quantized kernel least mean square algorithm," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 23, pp. 22–32, 2012.

Biographical sketch



Andres Marino Alvarez-Meza was born in 1988 in Pereira, Colombia. He received his undergraduate degree in electronic engineering (2009) and his M.Sc. degree in engineering-industrial automation (2011) from the Universidad Nacional de Colombia. Moreover, he received his Ph.D. in engineering-automatics at the same university in 2015. Between 2008 and 2015 he was appointed as research assistant and/or student in different projects related to signal processing and machine learning at the Signal Processing and Recognition Group at the Universidad Nacional de Colombia. His main research interests include machine learning and signal processing methods applied to image and video data analysis as well as bioengineering tasks.