Statistics and Computing manuscript No.
(will be inserted by the editor)

Kernel Biclustering algorithm in Hilbert Spaces

Marcos Matabuena · Juan C. Vidal

Received: date / Accepted: date

Abstract Biclustering algorithms partition data and covariates simultaneously, providing new insights in several domains, such as analyzing gene expression to discover new biological functions. This paper aims to establish a new model-free biclustering algorithm in abstract spaces using the notions of energy distance (ED) and the maximum mean discrepancy (MMD) –two distances between probability distribution in a separable Hilbert space and capable of handling complex data as curves or graphs. The proposed method can learn more general and complex cluster shapes than most existing literature approaches, usually focused on detecting mean and variance differences or other particular geometries shapes according to specific parametric distributions. Despite, the biclustering configurations of our approach are constrained to create disjoint structures at the datum and covariate levels, results are similar to state-of-the-art methods in their optimal scenarios, assuming a proper Kernel choice, outperforming them when cluster differences are concentrated in higher-order moments. Our approach has been tested in several situations that involve simulated and real-world datasets. Finally, new theoretical consistency results are established using some tools of the theory of optimal transport. 4

Keywords Biclustering · Kernel algorithms · Reproducing Kernel Hilbert Spaces · Unsupervised Learning · Funcional data analysis

1 Introduction

Cluster analysis plays a significant role in the exploratory and descriptive analysis of data in various unsupervised learning tasks and its application to multiple real-world problems (Jain et al., 1999). However, their application in many real-world applications can be limited. For instance, data distribution may be specified by latent local structures, in which the unit or individuals of each cluster are constituted for a different set of active covariates. For example, in the analysis of genetic profiles in bioinformatics, patient sub-types are characterized by specific regions of genes that can even show disjoint gene expression patterns among clusters (Madeira and Oliveira, 2004).

Biclustering algorithms (Madeira and Oliveira, 2004; Busygin et al., 2008) are a promising method to overcome the former limitation as they can create heterogeneous data groups with different covariates between clusters of individuals. In this way, they increase the interpretability and clinical meaning of the groups, making the analysis more robust to noise and mitigating the curse of dimensionality.

Biclustering methods can be classified into two categories according to the structures they are able to build (Madeira and Oliveira, 2004; Fraiman and Li, 2020). In the first category, clusters are arbitrarily positioned, and can overlap with each other (Cheng and Church, 2000; Getz et al., 2000; Bergmann et al., 2003; Tanay et al., 2004; Lazzeroni and Owen, 2002; Ben-Dor et al., 2002; Shabalin et al., 2009), while, in the second one,

Marcos Matabuena
E-mail: marcos.matabuena@usc.es

Marcos Matabuena · Juan C. Vidal
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, 15782 Santiago de Compsotela, Spain.

Juan C. Vidal
Departamento de Electrónica e Computación, Universidade de Santiago de Compostela, 15782 Santiago de Compsotela, Spain.

overlapping is not allowed and, thus, clusters are formed following a checkerboard structure in their matrix representation (Kluger et al., 2003; Lee et al., 2010; Chi et al., 2017; Flynn et al., 2020; Chen et al., 2013; Cho et al., 2004; Dhillon, 2001).

Another taxonomical classification is given by the probabilistic nature of the biclustering algorithm. On the one hand, generative hierarchical models specified through parametric distributions Flynn et al. (2020). On the other hand, we can consider non-parametric biclustering algorithms can also be defined through Anova decomposition in which clusters are constructed using different distance criteria that measure variability of the clusters (see for example Bryan et al. (2006)).

In addition, sparse biclustering algorithms have recently received more attention (Helgeson et al., 2020). These algorithms usually perform well in high-dimensional sparse structures, as they simultaneously perform the variable selection to reduce the dimensionality of the problem and the clusters construction.

However, despite the real progress in this active research area, existing methods still have limitations. For example, like in the $k-$means or probabilistic mixture algorithms, many biclustering algorithms can only detect structural changes in mean or the constructed cluster shapes following particular geometries according to parametric distributions Flynn et al. (2020). Nevertheless, variance (Chen et al., 2013) or another higher-order moment may be a critical component in real-world problems such as in establishing biological differences across patients subtypes (de Jong et al., 2019; Helgeson et al., 2020). Furthermore, it is well-known that solving a biclustering problem is generally an $NP-$hard problem (Madeira and Oliveira, 2004) and obtain efficient and reliable optimization strategies on a large scale or high dimensional settings can be challenging. Finally, the research biclustering techniques with complex statistical objects are limited to recent Euclidean functional data contributions Galvani et al. (2021). However, methods for complex statistical objects can be of great interest in contemporary applications such as personalized or precision medicine. In these context, the increasing ability to register patient health at high-order resolution calls for seeking better representations to have a more reliable characterization of clinical patient conditions Matabuena et al. (2021); Matabuena and Petersen (2021).

Kernel algorithms based on Reproducing Kernel Hilbert Spaces (RKHS) learning paradigm are a powerful data analysis modeling strategy, both in unsupervised as well as in supervised learning, to integrate and analyze complex statistical objects such as dynamic structures or functional data objects. One of the most representative examples of these techniques' modeling power is the maximum mean discrepancy (MMD) (Gretton et al., 2012), a distance between probability measures that has been successfully applied to many problems (Romano et al., 2020; Gretton et al., 2012; Wu et al., 2020). Recently in (França et al., 2020), MMD has been used to define a new non-parametric clustering algorithm in Separable Hilbert Spaces called Kernel k-groups, which improves the performance in several situations examined over traditional methods.

In this paper, we extend the Kernel k-groups approach (França et al., 2020) to biclustering, defining the first general biclustering method ($AKKB$) in separable Hilbert spaces based on the RKHS paradigm. Our method generalizes other clustering algorithms, such as the $k-$means or graph partitioning-based algorithms, and allows constructing clusters beyond the mean and variance, or other parametric shapes distributions, unlike most of the prior existing methodologies.

1.1 Outline of contributions

We summarize the main methodological contributions of this paper below:

1. To the best of our knowledge, we propose the first biclustering methodology under Reproducing Kernel Hilbert space learning paradigm that can be used to analyze complex data such as functional curves or graphs structures. At the same time, the proposed methodology allows detecting distributional differences beyond low order moments, increasing the variety of cluster shapes that algorithms can construct concerning a wide variety of existing clustering methods.
2. We propose an efficient optimization strategy that alternately applies the Kernel $k-$groups algorithm on the individual and covariate levels. For this purpose, we introduce the structural assumption that the clusters are mutually disjoint at both levels. Unlike most existing biclustering literature based on greedy optimization strategies (see for example Bryan et al. (2006))), the new optimization strategy presents theoretical guarantees.
3. We establish the consistency of our algorithm, something unusual in the biclustering literature, and discuss the mathematical connections of our proposal with other existing clustering algorithms.

2 Preliminaries

Let $(\Omega, \mathcal{F}_i, \mathbb{P}_i)$, $i \in \mathcal{I} = \{1, \ldots, p\}$, be p probability spaces with common sample space Ω. For each $i \in \mathcal{I}$, consider $\mathbb{H}_i$ a separable Hilbert space over $\mathbb{R}$ with inner product $\langle \cdot, \cdot \rangle_{\mathbb{H}_i}$, norm $\|\cdot\|_{\mathbb{H}_i}$, and σ-Field of sets $\mathcal{B}_{\mathbb{H}_i}$. Let $X^i : \Omega \to \mathbb{H}_i$ be a random element in $\mathbb{H}_i$, that is, an $(\mathcal{F}_i, \mathcal{B}_{\mathbb{H}_i})$-measurable mapping. Along this paper, we require that $\mathbb{E}\|X_i\|^2_{\mathbb{H}_i} < \infty$ and, without loss of generality, that Hilbert Space's $\mathbb{H}_i$'s are identical, which for the sake of simplicity we will denote as $\mathbb{H}$. Thus, multivariate random variables are denoted as $X = (X^1, \dots, X^p) \in \mathbb{H}^p = \mathbb{H} \times \cdots \times \mathbb{H}$ and the random sample observed as $X^{data} = \{X_i\}_{i=1}^n$, in which each X_i $(i = 1, \dots, n)$ is an independent copy of random variable X.

Let us denote by $\mathcal{J} = \{1, \dots, n\}$ the set of indices of the n observations and by $\mathcal{I} = \{1, \dots, p\}$ the set of indices of the p covariates. Since our goal is to simultaneously partition the set $\mathcal{J}$ and $\mathcal{I}$ into a fixed number of $k-$disjoints clusters, we denote by $\mathcal{A} := \{\mathcal{I}_1, \dots, \mathcal{I}_k\}$ and $\mathcal{B} := \{\mathcal{J}_1, \dots, \mathcal{J}_k\}$ the arbitrary partitions of the previous sets, i.e., $\cup_{i=1}^k \mathcal{I}_i = \mathcal{I}$, $\cup_{i=1}^k \mathcal{J}_i = \mathcal{J}$ and, $\forall i \neq j$ $\mathcal{I}_i \cap \mathcal{I}_j = \emptyset$, $\mathcal{J}_i \cap \mathcal{J}_j = \emptyset$.

For a given $x \in \mathbb{H}^p$, and the covariates selected in $\mathcal{I}_j$, let $x(\mathcal{I}_j) = (x^l)_{l \in \mathcal{I}_j} \in \mathbb{H}_{\mathcal{I}_j}$, where $\mathbb{H}_{\mathcal{I}_j}$ is a functional space in the Cartesian product of topological spaces $\times_{i \in \mathcal{I}_j} \mathbb{H}$. With this notation in hand, we define the norm of $\|x(\mathcal{I}_j)\|_{\mathcal{I}_j}$ as $\|x(\mathcal{I}_j)\|_{\mathcal{I}_j} = \sqrt{\frac{\sum_{i \in \mathcal{I}_j} \langle x^i, x^i \rangle_{\mathbb{H}}}{n_j}} = \sqrt{\frac{\sum_{i \in \mathcal{I}_j} \|x^i\|^2_{\mathbb{H}}}{n_j}}$, where the cardinally of $|\mathcal{I}_j| = n_j$. For example, suppose $\mathbb{H} = L^2([0,1]) = \{f : [0,1] \to \mathbb{R} : f$ is measurable and $\int_0^1 f^2(t)dt < \infty\}$. Then, $\|x(\mathcal{I}_j)\|_{\mathcal{I}_j} = \sqrt{\frac{\sum_{i \in \mathcal{I}_j} \langle x^i, x^i \rangle_{L^2([0,1])}}{n_j}} = \sqrt{\frac{\sum_{i \in \mathcal{I}_j} \int_0^1 (x^i(t))^2 dt}{n_j}}$.

The biclustering algorithm introduced here assumes that partitions at individual and covariate levels are mutually disjoint. Then, in order to express in a compact way the underlying optimization problem, we define the set $\mathcal{C}^{\mathcal{A},\mathcal{B}} := \{\mathcal{C}_1, \dots, \mathcal{C}_k\}$, where each $\mathcal{C}_l = \{X_i(\mathcal{I}_l) : i \in \mathcal{J}_l\}$ depends on the selected partitions $\mathcal{A}$ and $\mathcal{B}$. Our biclustering algorithm measures the similarity between individuals within a subset of covariates using a symmetric positive definite kernel, $\mathcal{K}_{\mathcal{I}_j} : \mathbb{H}_{\mathcal{I}_j} \times \mathbb{H}_{\mathcal{I}_j} \to \mathbb{R}^+$. We assume that $\mathcal{K}_{\mathcal{I}_j}$'s can be expressed as $\mathcal{K}_{\mathcal{I}_j}(x(\mathcal{I}_j), y(\mathcal{I}_j)) = f(\|x(\mathcal{I}_j) - y(\mathcal{I}_j)\|_{\mathcal{I}_j})$, where $f(\cdot)$ is a enough regular function. Along this paper, we will use the Gaussian Kernel, $\mathcal{K}_{\mathcal{I}_j}(x(\mathcal{I}_j), y(\mathcal{I}_j)) = e^{-\frac{\left\|x(\mathcal{I}_j) - y(\mathcal{I}_j)\right\|_{\mathcal{I}_j}}{\sigma^2}}$, where $\sigma > 0$ is the bandwidth parameter of the kernel.

With this kernel construction, the RKHS space associated with the local kernel, $\mathbb{H}_{\mathcal{K}_{\mathcal{I}_j}}$, is universal, that is, $\mathbb{H}_{\mathcal{K}_{\mathcal{I}_j}}$ is dense on $\mathcal{C}(\mathbb{H}_{\mathcal{I}_j}) = \{f : \mathbb{H}_{\mathcal{I}_j} \to \mathbb{R} :$ f is continuos$\}$ and equiped with the usual supremum norm $\|\cdot\|_\infty$, which guarantees to approximate any continuous function from this functional space. Also $\mathbb{H}_{\mathcal{K}_{\mathcal{I}_j}}$ is characteristics, for any distribution F defined on $\mathcal{H}_{\mathcal{I}_j}$, the map $\phi^{\mathcal{K}_{\mathcal{I}_j}} : x \to \int \mathcal{K}_{\mathcal{I}_j}(x, \cdot) F(dx)$ is injective. If the kernel is characteristics, we can characterize the equality in distribution or statistical independence in several samples (Sriperumbudur et al., 2011). Finally, we denote the norm of RKHS generated by the Kernel $\mathcal{K}_{\mathcal{I}_j}$, $\mathbb{H}_{\mathcal{K}_{\mathcal{I}_j}}$, as $\|\cdot\|_{\mathcal{K}_{\mathcal{I}_j}}$.

3 Mathematical models

3.1 Preliminary models

3.1.1 Maximum mean discrepancy and energy distance

Maximum mean discrepancy (MMD) and energy distance (ED) are two families of statistical distances between arbitrary random elements that take values in separable Hilbert Spaces. With their increase in popularity at the beginning of this century, data analysis methods derived from these distances are nowadays an essential tool in a vast amount of applications and statistical modeling tasks, e.g., hypothesis testing.The equivalence between these two families of distance, at the population and finite sample levels, was established in a series of recent papers using the connections between metrics of negative-type and define positive symmetry kernels.

Let be $X \sim F, Y \sim G$ be two $\mathbb{H}^p$-random variables that hold $\mathbb{E}(\|X\|^2) < \infty$ and $\mathbb{E}(\|Y\|^2) < \infty$. The ED of order $\alpha \in (0,2]$ can be defined as:

$$\begin{aligned} \epsilon_\alpha(X,Y) = 2\mathbb{E}(\|X - Y\|^\alpha) - \mathbb{E}(\left\|X - X'\right\|^\alpha) - \\ \mathbb{E}(\left\|Y - Y'\right\|^\alpha), \end{aligned} \tag{1}$$

where X' and Y' are i.i.d random variables copies of X, Y.

ED can generalize to obtain a more general family of statistical distances in separable Hilbert Spaces. For this purpose, it is enough to consider an arbitrary semimetric of negative-type $\rho(\cdot,\cdot)$ (see Berg et al. (1984) for a formal definition). In such case, $\epsilon_\rho(X,Y)$ is defined by analogy as:

$$\begin{aligned} \epsilon_\rho(X,Y) = 2\mathbb{E}(\rho(X,Y)) - \mathbb{E}(\rho(X,X')) \\ - \mathbb{E}(\rho(Y,Y')), \end{aligned}$$

which is equivalent to Eq. (1) when $\rho(x,y) = \|x - y\|^\alpha$.

Consider the kernel $\mathcal{K}:\mathbb{H}^p\times\mathbb{H}^p\to\mathbb{R}^+$ symmetric and define positive. We define:

$$\begin{aligned}
&MMD(X,Y)^2_{\mathcal{K}}\\
&=\left\|\phi^{\mathcal{K}}_X-\phi^{\mathcal{K}}_Y\right\|^2_{\mathbb{H}^p_{\mathcal{K}}}\\
&=\left\|\int\mathcal{K}(x,)F(dx)-\int\mathcal{K}(y,)G(dy)\right\|^2_{\mathbb{H}^p_{\mathcal{K}}}\\
&=\langle\int\mathcal{K}(x,)F(dx),\int\mathcal{K}(x,)F(dx)\rangle_{\mathbb{H}^p_{\mathcal{K}}}\\
&\quad+\langle\int\mathcal{K}(y,)G(dy),\int\mathcal{K}(y,)G(dy)\rangle_{\mathbb{H}^p_{\mathcal{K}}}\\
&\quad-2\langle\int\mathcal{K}(x,)F(dx),\int\mathcal{K}(y,)G(dy)\rangle_{\mathbb{H}^p_{\mathcal{K}}}\\
&=\mathbb{E}(\mathcal{K}(X,X'))+\mathbb{E}(\mathcal{K}(Y,Y'))-2\mathbb{E}(\mathcal{K}(X,Y)),
\end{aligned}\tag{2}$$

where $\phi^{\mathcal{K}}_X\in\mathbb{H}^p_{\mathcal{K}}$ is know in the literature as the kernel mean embedding of the random variable X generated by the kernel $\mathcal{K}$. If for any distribution function F, the mapping $\phi^{\mathcal{K}}_X:x\in\mathbb{H}^p\to\int\mathcal{K}(x,)F(dx)$ is inyective, then we can construct omnibus tests that characterize the equality in distribution.

Given two samples i.i.d $\{X_i\}_{i=1}^{n_1}\sim F$ and $\{Y_i\}_{i=1}^{n_2}\sim G$, $n_1+n_2=n$, we can straightforward estimate the empirical counterpart, replacing the true distribution functions F and G by the empirical distributions F_{n_1} and G_{n_2}. More specifically, we obtain:

$$\begin{aligned}
\tilde{\epsilon}_\rho(X,Y)=&\frac{2}{n_1n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\rho(X_i,Y_j)-\frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\rho(X_i,X_j)\\
&-\frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}\rho(Y_i,Y_j),
\end{aligned}$$

and

$$\begin{aligned}
M\tilde{M}D(X,Y)_{\mathcal{K}}=&\frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\mathcal{K}(X_i,X_j)\\
&+\frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}\mathcal{K}(Y_i,Y_j)\\
&-\frac{2}{n_1n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\mathcal{K}(X_i,Y_j)\\
&.
\end{aligned}$$

The election of ρ or the $\mathcal{K}$ in the ED or MMD is critical in the finite sample performance in the different modeling tasks. However, it is not easy to establish a general criterion in a general set-up since each selected semi-metric characterizes distributional differences, giving more or less priority to a specific moment in the distance computation. Moreover, this information is not available in practice, and difficult to obtain from expert knowledge or prior studies.

Consider k ($k>2$) random variables, $X^1\sim F^1,\dots,X^k\sim F^k$ and k i.i.d random samples, $X^{1,data}=\{X_i^{1,data}\}_{i=1}^{n_1}\sim F^1$, …, $X^{k,data}=\{X_i^{k,data}\}_{i=1}^{n_k}\sim F^k$, $n_1+\dots+n_k=n$. Our goal is to define an ED statistics to compare distributional differences between the k-random samples. For this purpose, we can use the standard ED statistics between each pair of random samples as follows:

$$\begin{aligned}
&\tilde{\epsilon}_\rho(X^1,X^2,\dots,X^k)\\
&=\sum_{m=1}^{k}\sum_{r\neq m}^{k}\frac{2n_mn_r}{n}\tilde{\epsilon}_\rho(X^{m,data},X^{r,data})\\
&=\sum_{m=1}^{k}\sum_{r\neq m}^{k}\frac{2n_mn_r}{n}\left[\frac{2}{n_mn_r}\sum_{i=1}^{n_m}\sum_{j=1}^{n_r}\rho(X_i^m,X_j^r)\right.\\
&\qquad\left.-\frac{1}{n_m^2}\sum_{i=1}^{n_m}\sum_{j=1}^{n_m}\rho(X_i^m,X_j^m)-\frac{1}{n_r^2}\sum_{i=1}^{n_r}\sum_{j=1}^{n_r}\rho(X_i^r,X_j^r)\right],
\end{aligned}\tag{3}$$

where we applied by convenience the crossed 2-empirical energy distance with specified size standardization related to group sizes of samples.

Hereinafter we will summarize the notation as follows, where $m=r=1,\dots,k$:

$$\begin{aligned}
&\rho(X^{m,data},X^{r,data})\\
&=\sum_{i=1}^{n_m}\sum_{j=1}^{n_r}\frac{1}{n_mn_r}\rho(X_i^{m,data},X_j^{r,data}).
\end{aligned}$$

So,

$$\begin{aligned}
&\tilde{\epsilon}_\rho(X^1,X^2,\dots,X^k)\\
&=\sum_{m=1}^{k}\sum_{r\neq m}^{k}\frac{2n_mn_r}{n}\Big[2\rho(X^{m,data},X^{r,data})\\
&\qquad-\rho(X^{m,data},X^{m,data})\\
&\qquad-\rho(X^{r,data},X^{r,data})\Big].
\end{aligned}\tag{4}$$

It should be remarked that we can establish an important connection between Eq. (3) and Eq. (4), and the empirical Gini mean distance. In particular, it is held that:

$$\tilde{\epsilon}_\rho(X^1,X^2,\dots,X^k)+\sum_{n\ =1}^{k}\frac{n_n}{n}\rho(X^{n,data},X^{n,data})\tag{5}$$

is constant. This connection is essential to define a new notion of biclustering algorithm in the context of RKHS, as it allows to use the Gini mean distance or the energy distance in the clustering definition.

3.1.2 Model free clustering in RKHS

Any clustering algorithm aims to create $k-$groups such that elements in the same group are similar, while elements between different groups are the most different from each other. According to França et al. (2020), a natural clustering algorithm can be defined with ED by selecting the partition of X^{data}, the size k, and the set of clusters $\{\hat{\mathcal{C}}_1,\ldots,\hat{\mathcal{C}}_k\}$ that maximize the following optimization problem with multi-sample energy statistics:

$$\{\hat{\mathcal{C}}_1,\ldots,\hat{\mathcal{C}}_k\} = \arg \max_{\{\mathcal{C}_1,\ldots,\mathcal{C}_k\}} \tilde{\epsilon}_\rho(\mathcal{C}_1,\ldots,\mathcal{C}_k). \tag{6}$$

Using a semi-metric ρ selected beforehand, this optimization problem finds the partition of X^{data} which maximizes the group-separation criteria based on multi-sample energy distance statistics. It is know that given a semi-metric ρ of negative type, one can associate a symmetric, positive, characteristic Kernel $\mathcal{K}$, as follows, $\mathcal{K}(x,y) = \frac{1}{2}[\rho(x,x_0)+\rho(y,x_0)-\rho(x,y)]$, $\forall x,y \in \mathbb{H}^p$, where $x_0 \in \mathbb{H}^p$ is an arbitrary fixed point (Berg et al., 1984; Sejdinovic et al., 2013). Following França et al. (2020), and using the above relation, the previous optimization problem can be written as:

$$\{\hat{\mathcal{C}}_1,\ldots,\hat{\mathcal{C}}_k\} = \arg \min_{\{\mathcal{C}_1,\ldots,\mathcal{C}_k\}} \sum_{i=1}^{k} \frac{1}{n_{\mathcal{C}_i}} \underbrace{\sum_{x\in\mathcal{C}_i}\sum_{y\in\mathcal{C}_i} \mathcal{K}(x,y)}_{Q_i}, \tag{7}$$

where $|\mathcal{C}_i|= n_{\mathcal{C}_i}$; and Q_i is the so-called *within Kernel dispersion* between the elements of the cluster $\mathcal{C}_i$.

Using the optimization problem defined in Eq. (7), we can see that classical clustering algorithms such as graph partitioning or Kernel $k-$means are a special case of the optimization problem defined in Eq. (6). In this paper, we are particularly interested in its relationship with the Kernel $k-$means algorithm since it motivates our biclustering formulation and our optimization strategy. In particular, the optimization problem of Eq. (7) is equivalent to:

$$\{\hat{\mathcal{C}}_1,\ldots,\hat{\mathcal{C}}_k\} = \arg \min_{\{\mathcal{C}_1,\ldots,\mathcal{C}_k\}} \sum_{i=1}^{k} \frac{1}{n_{\mathcal{C}_i}} \sum_{x\in\mathcal{C}_i} ||\phi(x) - \hat{\mu}_{\mathcal{C}_i}||^2_{\mathcal{K}}, \tag{8}$$

where $\phi(x) = \mathcal{K}(\cdot,x) \in \mathbb{H}^p_{\mathcal{K}}$, $\hat{\mu}_{\mathcal{C}_i} = \frac{1}{n_{\mathcal{C}_i}}\sum_{x\in\mathcal{C}_i}\phi(x)$.

With this relationship in mind, we can find the optimal solution of Kernel k-groups clustering algorithm (França et al., 2020), using the classical resolution strategies of Kernel k-means algorithms such as Hartigan's or Lloyd's algorithms.

Let $\mathcal{Q}_l(x) = \sum_{y\in\mathcal{C}_l}\mathcal{K}(x,y)$ be, the cost to move the datum $x \in \mathcal{H}^p$ to cluster $\mathcal{C}_l$. We can define the cost to move the datum x from cluster j to l as:

$$\Delta\mathcal{Q}^{(j\to l)}(x) = \frac{\mathcal{Q}_l^+}{n_{\mathcal{C}_l}+1} + \frac{\mathcal{Q}_j^-}{n_{\mathcal{C}_j}-1} - \frac{\mathcal{Q}_l}{n_{\mathcal{C}_l}} - \frac{\mathcal{Q}_j}{n_{\mathcal{C}_j}}, \tag{9}$$

where $\mathcal{Q}_l^+(x)$ and $\mathcal{Q}_j^-(x)$, are the cost to add x to the cluster $\mathcal{C}_l$ and to remove x from the cluster j, respectively, where:

$$\mathcal{Q}_l^+(x) = \mathcal{Q}_l + 2\mathcal{Q}_l(x) + \mathcal{K}(x,x),$$
$$\mathcal{Q}_j^-(x) = \mathcal{Q}_j - 2\mathcal{Q}_j(x) + \mathcal{K}(x,x),$$

and:

$$\mathcal{Q}_l = \sum_{x\in\mathcal{C}_l}\sum_{y\in\mathcal{C}_l}\mathcal{K}(x,y) \;\; (l=1,\ldots,k),$$

is the so-called *within Kernel dispersion* between the elements of the cluster $\mathcal{C}_l$. A resolution strategy can therefore consist of solving recurrently the following problem:

$$j^* = \arg \max_{l=1,\ldots,k|l\neq j} \Delta\mathcal{Q}^{(j\to l)}(x), \forall x \in X^{data}. \tag{10}$$

If $\Delta\mathcal{Q}^{(j\to j^*)}(x) > 0$, that is, if we can improve the objective function moving the datum x to cluster $\mathcal{C}_{j^*}$.

3.2 *AKKB* algorithm

3.2.1 Mathemathical definition

This section introduces the new biclustering kernel algorithm in Separable Hilbert Spaces inspired by the concepts of ED and MMD. Nevertheless, our final formulation can be seen as an infinite-dimensional RKHS generalization of the biclustering algorithm proposed in Fraiman and Li (2020), that is designed to detect structural differences in mean in the context of finite-dimensional Euclidean Spaces.

First, we must note that the proposed algorithm introduces specific shapes constraints to construct k-clusters at individual and covariate levels. In particular, we assume that k-clusters formed in the previous two levels are mutually disjoint. Under this structural assumption, we provide an efficient resolution strategy using the standard Kernel k-means algorithm. Although this might look restrictive, there are many real-world examples that fit this constraint, such as in genetic studies where clusters of patients are characterized by disjoint biological structures of genes Madeira and Oliveira (2004). In addition, this approach has another potential advantage in some settings as the cluster of covariables can be more interpretable. Finally, the proposed algorithm and framework represent the formal basis for handling other more general biclustering configurations and analyzing complex statistics objects with biclustering techniques. It should be mentioned, however, that the obtention of more general biclustering configurations, e.g. using greedy optimization methods, is out of the scope of this paper.

In order to propose a biclustering RKHS procedure capable of detecting general distributional differences in general separable Hilbert Spaces, we can not use the notions of ED and MMD directly, since random elements that constitute each subset of $\mathcal{C}^{A,B} = \{\mathcal{C}_1, \dots, \mathcal{C}_k\}$, where $\mathcal{C}_l = \{X_i(\mathcal{I}_l) : i \in \mathcal{J}_l\}$, can have different dimensions. However, we can exploit the connection between the ED and the Gini mean distance introduced in Eq. (5) to obtain another characterization of the clustering problem defined in Eq. (7). More specifically, we can now define our biclustering problem as follows:

$$\hat{\mathcal{C}}^{\hat{\mathcal{A}},\hat{\mathcal{B}}} = \arg \min_{\mathcal{C}^{\mathcal{A},\mathcal{B}} = \{\mathcal{C}_1,\dots,\mathcal{C}_k\}} \sum_{i=1}^{k} \frac{1}{n_{\mathcal{C}_i}} \rho_{\mathcal{I}_i}(\mathcal{C}_i, \mathcal{C}_i), \tag{11}$$

where $\rho_{\mathcal{I}_i}$ denotes a metric of negative-type, related with the Kernel $\mathcal{K}_{\mathcal{I}_i}$ previously defined in Section 2. In particular, let $\rho_{\mathcal{I}_i}(x,y) = \mathcal{K}_{\mathcal{I}_i}(x,x_0) + \mathcal{K}_{\mathcal{I}_i}(y,x_0) - 2\mathcal{K}_{\mathcal{I}_i}(x,y)$ $\forall x,y \in \mathbb{H}_{\mathcal{I}_i}$, $x_0 \in \mathbb{H}_{\mathcal{I}_i}$ be an arbitrary fixed point with $\mathcal{K}_{\mathcal{I}_i}(x,y) = f(\|x-y\|_{\mathcal{I}_i})$ and f a regular and strict monotone function. Using similar arguments to França et al. (2020), we can establish that the problem is equivalent to:

$$\hat{\mathcal{C}}^{\hat{\mathcal{A}},\hat{\mathcal{B}}} = \arg \min_{\mathcal{C}^{A,B}} \sum_{i=1}^{k} \frac{1}{n_{\mathcal{C}_i}} \sum_{x \in Ci} \|\phi_{\mathcal{I}_i}(x) - \hat{\mu}_{\mathcal{I}_i}\|^2_{\mathcal{K}_{\mathcal{I}_i}}, \tag{12}$$

where $\phi_{I_i}(x) = \mathcal{K}_{\mathcal{I}_i}(\cdot, x) \in \mathbb{H}_{\mathcal{K}_{\mathcal{I}_i}}$ and $\hat{\mu}_{\mathcal{C}_i} = \frac{1}{n_{\mathcal{C}_i}} \sum_{x \in \mathcal{C}_i} \phi_{\mathcal{I}_i}(x)$.

3.2.2 Optimization Algorithm

The mains steps for solving the optimization problem defined in Eq. (12) are listed below:

1. Start by running the Kernel k-groups algorithm introduced in the Section 3.1.2 separately at individual and covariate levels on X^{data} to obtain the initial partitions $\mathcal{B} := \{\mathcal{J}_1, \dots, \mathcal{J}_k\}$, and $\mathcal{A} := \{\mathcal{I}_1, \dots, \mathcal{I}_k\}$, respectively. If we represent the input data as a matrix X^{input}, we can use X^{input} and its transpose $(X^{input})^t$ to obtain two independent clustering for individuals and covariates, respectively. We must note that X^{input} is not always an $n \times p$ matrix. Suppose, for instance, that each covariate is a curve that takes values in $\mathbb{H} = L^2[0,1]$ and we observe the curve values in a grid on m points of interval $[0,1]$. Then, X^{input} will be a matrix $n \times (m \times p)$, where each subset of m columns saves the information about each funcional covariate.
2. From the partition $\mathcal{A} := \{\mathcal{I}_1, \dots, \mathcal{I}_k\}$ of $k-$groups of covariates, that have an associated Kernel $\mathcal{K}_{\mathcal{I}_i}$ $(i = 1, \dots, k)$ (see Section 2), we can build a new partition on individuals $\hat{\mathcal{B}} = \{\mathcal{J}_1, \dots, \mathcal{J}_k\}$ that solves the following optimization problem:

$$\hat{\mathcal{C}}^{\mathcal{A},\hat{\mathcal{B}}} = \arg \min_{\mathcal{C}^{\mathcal{A},\mathcal{B}}} \sum_{i=1}^{k} \sum_{x \in \mathcal{C}_i} \|\phi_{\mathcal{I}_i}(x) - \hat{\mu}_{\mathcal{I}_i}\|^2_{\mathcal{K}_{\mathcal{I}_i}}. \tag{13}$$

The optimal solution can be obtained as detailed in Section 3.1.2. More specifically, the challenge is to assign $x \in X^{data}$ to a new cluster using the criteria:

$$j^* = \arg \max_{l=1,\dots,k | l \neq j} \Delta \mathcal{Q}^{(j \to l)}(x), \forall x \in X^{data}. \tag{14}$$

However, we must take into account the local structure of Kernels $\mathcal{K}_{\mathcal{I}_i}$ $(l = 1, \dots, k)$. Thus:

$$\mathcal{Q}_l^+(x) = \mathcal{Q}_l + 2\mathcal{Q}_l(x) + \mathcal{K}_{\mathcal{I}_l}(x(\mathcal{I}_l), x(\mathcal{I}_l)),$$
$$\mathcal{Q}_j^-(x) = \mathcal{Q}_j - 2\mathcal{Q}_j(x) + \mathcal{K}_{\mathcal{I}_j}(x(\mathcal{I}_j), x(\mathcal{I}_j)),$$

where $\mathcal{Q}_l(x)$ and $\mathcal{Q}_l$ are redefined as follows:

$$\mathcal{Q}_l(x) = \sum_{y \in \mathcal{C}_l} \mathcal{K}_{\mathcal{I}_l}(x(\mathcal{I}_l), y),$$
$$\mathcal{Q}_l = \sum_{x \in \mathcal{C}_l} \sum_{y \in \mathcal{C}_l} \mathcal{K}_{\mathcal{I}_l}(x,y) \quad (l = 1, \dots, k).$$

Solving this problem is computationally intensive as local Kernels must be recalculated according to the normalized norm defined in Section 2.
3. Repeat step 2 but in this case changing the role of $\mathcal{B} := \{\mathcal{J}_1, \dots, \mathcal{J}_k\}$ by $\mathcal{A} := \{\mathcal{I}_1, \dots, \mathcal{I}_k\}$.
4. Repeat alternatively steps 2 and 3, T times, to get the succession of partitions $\{\mathcal{B}^k\}_{k=1}^T$ and $\{\mathcal{A}^k\}_{k=1}^T$, where the sub-index k, denotes an iteration of the algorithm. Return the solution $\mathcal{C}^{\mathcal{A}^T,\mathcal{B}^T}$.

In real-world scenarios, we know that local resolution strategies based on Kernel $k-$means algorithms can strongly depend on the initial condition and get stuck in local minima. To avoid this, we applied R random restarts in each phase of the algorithm.

3.2.3 Theory

This Section introduces some mathematical properties of our biclustering proposal as well as consistency results. Rigorous mathematical proofs are relegated to Supplementary Material.

Before starting, we present some statistical learning concepts related to clustering analysis. For this aim, we adapt the reference to the notation introduced Biau et al. (2008). Let us consider the following empirical risk function associated with our biclustering problem:

$$\mathcal{W}_n(c_n, \mathcal{A} = \{\mathcal{I}_1, \ldots, \mathcal{I}_k\}, X^{data}, \mu_n) \\ = \min_{\mathcal{C}^{\mathcal{A},\mathcal{B}} = (\mathcal{C}_1, \ldots, \mathcal{C}_k)} \frac{1}{n} \sum_{i=1}^{k} \sum_{x \in \mathcal{C}_i} \|\phi_{\mathcal{I}_i}(x) - c_n(I_i)\|^2_{\mathcal{K}_{\mathcal{I}_i}}, \tag{15}$$

where μ_n is the empirical measure of X^{data}, $\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$, and, for any $\mathcal{I}_i \in \mathcal{P}(\{1, \ldots, p\})$,

$c_n : \mathcal{I}_i \in \mathcal{P}(\{1, \ldots, p\}) \to c_n(\mathcal{I}_i) \in \mathcal{K}_{\mathcal{I}_i}$ is a set-valued function in an appropriate $RKHS$, denoting by $\mathcal{P}(\{1, \ldots, p\})$ the power set of covariate indexes $\{1, \ldots, p\}$.

The problem defined in Eq. (15) finds the partition at the individual level that minimizes the empirical risk function, from a fixed set-valued function c_n and the cluster covariates $\mathcal{A}$.

Consider now the population counterpart:

$$\mathcal{W}(c, \mathcal{A}, \mu) = \int \min_{1 \le j \le k} \left\| \phi_{\mathcal{I}_j}(x(\mathcal{I}_j)) - c(I_j) \right\|^2_{\mathcal{K}_{\mathcal{I}_j}} d\mu(x), \tag{16}$$

where μ denotes the measure of the random variable X, $x(\mathcal{I}_i) = (x^i)_{i \in \mathcal{I}_i}$, and $c : \mathcal{I}_i \in \mathcal{P}(\{1, \ldots, p\}) \to c(\mathcal{I}_i) \in \mathcal{K}_{\mathcal{I}_i}$ is a fixed set-valued function in a RKHS. Then, the optimal clustering risk can be defined as:

$$\mathcal{W}^*(\mu) = \inf_{\mathcal{A}} \inf_{c} W(\mathcal{A}, c, \mu). \tag{17}$$

We must note that for each $\mathcal{I}_i$ fixed, the infimum of $c(\mathcal{I}_i)$ is found in the mean, due to the Euclidean geometry of the quadratic problem.

We say that the partition of the set of covariates $\mathcal{A}_n$ and the cluster centers c_n are δ_n minimizer of the empirical clustering risk if:

$$\mathcal{W}_n(c_n, \mathcal{A}_n, X^{data}, \mu_n) \le W^*(\mu_n) + \delta_n \tag{18}$$

To establish consistency results, we assume that $\delta_n \to 0$ when $n \to \infty$. In practice, this means that δ_n-minimizers of the empirical clustering risk converge to the optimal risk.

Theorem 1 *Let A_n and c_n be δ_n-minimizers of the empirical clustering risk such that when $n \to \infty$, $\delta_n \to 0$. Then:*

- $\lim_{n\to\infty} \mathcal{W}(c_n, \mathcal{A}_n, X^{data}, \mu_n) = \mathcal{W}^*(\mu)$ *with probability one.*
- $\lim_{n\to\infty} \mathbb{E}(W(A_n, c_n, \mu_n)) = \mathcal{W}^*(\mu)$

Bellow, we introduce some results that provide theoretical guarantees the optimization strategy defined prior in the Section 3.2.2. In addition, we offer the computational complexity cost of our procedure.

Theorem 2 *AKKB biclustering algorithm converges in a finite number of steps.*

Theorem 3 *The complexity of AKKB biclustering algorithm is $\mathcal{O}(k^2 n^4)$, where k is the number of clusters, n is the number of data points.*

4 Results

$AKKB$[1] performance has been compared against the following state-of-the-art algorithms, using synthetics and real datasets:

- Alternating k-means biclustering (AKM)[2] (Fraiman and Li, 2020): This algorithm find local minimum by alternating the use of an adapted version of the k-means clustering algorithm between columns and rows, and is a particular case of $AKKB$.
- Profile likelihood biclustering (PL)[3] (Flynn et al., 2020): This method is based on the maximal profile likelihood using as a reference exponential family distributions. In our experiments, the algorithm was run using the Gaussian distribution family.
- Sparse biclustering (SBC)[4] (Tan and Witten, 2014): This algorithm assumes that data entries are Gaussian with a Bicluster-specific mean and equal variance that maximize the L_1-penalized log-likelihood to obtain sparse Biclusters. An optimal $\lambda-$regularization parameter is selected according to the Bayesian information criterion (BIC).

As in Fraiman and Li (2020), we selected the former algorithms in this comparison since:

1. Each datum and covariate are grouped in a unique cluster. Thus, biclustering solutions should produce non-overlapping clusters.
2. The selected biclustering methods allow explicitly specify the number of clusters at both individual and covariate levels.

Along with this paper, and according to the definition established in Section 2, $AKKB$ was fitted using a Gaussian Kernel. In the kernel definition, we consider two different kernel bandwidths, at individual and covariate levels, that we denote by σ_{data} and $\sigma_{variables}$, respectively. In both cases, the median heuristics was used to obtain an initial estimation of this kernel parameter. Supplementary material provides additional

[1] An `R` package is available at `https://github.com/anonymousclustering/biclustering` to reproduce the results.

[2] `R` package `akmbiclust`

[3] `R` package `biclustpl`

[4] `R` package `sparseBC`

results in which we analyze the impact of the bandwidth selection at both levels.

We must note that the median heuristic is one of the most popular criteria to select the bandwidth parameter with Gaussian kernels. In Ramdas et al. (2015), the author shows that the application of heuristic means strategy can maximize MMD power in different simulation scenarios. However, we do not have any guarantee about its performance in the biclustering set-ups.

Finally, a hundred restarts are applied in all algorithms used in the comparative.

4.1 Simulated data

One thousand simulations of different random process X_{ij} $(i = 1,\dots,200; j = 1,\dots,200)$ were performed. We assume that the random process X_{ij} defines two latent clusters in a block matrix 2×2. A is defined as:

$$A = (X_{ij}) = \left(\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right)$$

where:

$$A_{11} = (X_{ij})_{i=1,\dots,100}^{j=1,\dots,100} \qquad A_{12} = (X_{ij})_{i=1,\dots,100}^{j=101,\dots,200}$$
$$A_{21} = (X_{ij})_{i=101,\dots,200}^{j=1,\dots,100} \qquad A_{22} = (X_{ij})_{i=101,\dots,200}^{i=101,\dots200}$$

The following scenarios were considered to illustrate the theoretical advantages of $AKKB$:

1. Gaussian data such that clustering differences are in variance, whereas all data have the same mean:

   $$A_{11} \sim N(0,2),$$
   $$A_{22} = A_{12} = A_{21} \sim N(0,1).$$

2. Data distributed according to a uniform distribution, in which each block of the random matrix A has the same mean, but there are differences in the rest of the moments:

   $$A_{11} \sim Unif(0.3,0.7),$$
   $$A_{22} = A_{12} = A_{21} \sim Unif(0,1).$$

3. A random matrix A distributed according to uniform, standard Gaussian, and truncated Gaussian distributions, with the same first three moments. In particular:

   $$A_{11} \sim Unif(-\sqrt{3},\sqrt{3}), A_{22} \sim N(0,1),$$
   $$A_{12} = A_{21} \sim N_{truncated}(0,1,-1.8,1.8) + Unif(-0.5,0.5)$$

Table 1 Average accuracy in the three synthetic scenarios with 1000 trials.

Scenario	$AKKB$	SBC	PL	AKM
1	1	0.52	0.535	1
2	0.92	0.54	0.534	0.78
3	0.91	0.515	0.505	0.83

Table 2 Average accuracy in three real dataset, where N denotes the sample size and p the number of covariates.

Data set (n,p)	$AKKB$	SBC	PL	AKM
West-2001 (49,1198)	0.85	0.51	0.81	0.81
Chowdary-2006 (42,182) 2	0.98	0.65	0.65	0.96
Armstrong-2002-v1 3 (72,1081)	0.75	0.90	0.61	0.76

The average accuracy reached by each algorithm in the three scenarios is listed in Table 1. We can see that $AKKB$ outperforms the other algorithms in all three scenarios analyzed. As expected, SBC and PL present worse performance when cluster's differences are due to changes in moments that go beyond the mean.

4.2 Genetic expression data-sets

$AKKB$ was also tested with real biomedical examples, specifically in some genetic expression datasets of the *Schliep lab* repository[5] on which true structures are annotated. The results are introduced in Table 2.

As we can see, $AKKB$ obtains competitive results in all datasets. However, SBC outperforms them in Armstrong-2002-v1, which can indicate that an underlying sparse structure may often be present in the biological data. This might indicate that applying variable selection or dimension reduction before AKKB could increase the performance of our model.

4.3 Functional data diabetes example

Functional data analysis (Ramsay and Silverman, 2007) is a research area that has received substantial attention in recent years from the statistical community to get new insight into the analysis of objects that vary in a continuum as a temporal curve of a physiological process.

In this paper to show the versatility of our biclustering algorithm, we will analyze a functional example in which we have different functional profiles obtained from the information available from a continuous glucose monitor.

[5] Available at `https://schlieplab.org/Static/Supplements/CompCancer/datasets.htm`

More specifically, we will use data from a case-control study (Weinstock et al., 2016) that aims to analyze risk factors associated with severe hypoglycemia in patients with type I diabetes. On each day with CGM information we observe 288 spaced observations every 5 min that we denote by the index j $(j=1,\ldots,288)$, and we define the related time $t_j = 5*j$.

For each participant i, $(i=1,\ldots,n)$, we have available n_i days of continuous glucose monitoring, $n_i \in \{10,\ldots,16\}$. Let us consider the random process $X_{id}(t_j)$, $(i=1,\ldots,n;\ d=1,\ldots,n_i;\ j=1,\ldots,144)$, which models the glucose concentration of subject i, on day d, and for time t_j. Also, for each subject i, let us consider the following random processes related with the morning and afternoon:

- $X_i^{mean,morning}(t_j) = \frac{1}{n_i}\sum_{d=1}^{n_i} X_{id}(t_j)$,
- $X_i^{sd,morning}(t_j) = \sqrt{\frac{1}{n_i}\sum_{d=1}^{n_i}(X_{id}(t_j) - X_i^{mean}(t_j))^2}$,
- $X_i^{CV,morning}(t_j) = \frac{X_i^{mean}(t_j)}{X_i^{sd}(t_j)}$,
- $X_i^{Skew,morning}(t_j) = [\frac{(X_{id}(t_j)-X_i^{mean}(t_j))}{X_i^{sd}(t_j)}]^3$,
- $X_i^{curtos,morning}(t_j) = [\frac{(X_{id}(t_j)-X_i^{mean}(t_j))}{X_i^{sd}(t_j)}]^4$,
- $X_i^{mean,afternoon}(725+t_j) = \frac{1}{n_i}\sum_{d=1}^{n_i} X_{id}(725+t_j)$,
- $X_i^{sd,afternoon}(725+t_j) = \sqrt{\frac{1}{n_i}\sum_{d=1}^{n_i}(X_{id}(725+t_j) - X_i^{mean}(725+t_j))^2}$,
- $X_i^{CV,afternoon}(725+t_j) = \frac{X_i^{mean}(725+t_j)}{X_i^{sd}(725+t_j)}$,
- $X_i^{Skew,afternoon}(725+t_j) = [\frac{(X_{id}(t_j)-X_i^{mean}(725+t_j))}{X_i^{sd}(725+t_j)}]^3$,
- $X_i^{curtos,afternoon}(725+t_j) = [\frac{(X_{id}(725+t_j)-X_i^{mean}(725+t_j))}{X_i^{sd}(725+t_j)}]^4$

where $j=1,\ldots,144$. In this example, we assume that each functional covariate analyzed takes values in the space $\mathcal{H} = L^2([0,720])$.

Figures 1 and 2 show the results of running the AKKB algorithm $(k=2)$ clusters. In the graph, we can see those patients with a more stable glycemic control (color Red) and that their average daily glycemic values characterize them in the morning and afternoon. In the other group (color Blue), we see patients with higher volatility, whose covariates are those related to variables measuring different modes of data variability. The figure also shows the values of the Glycosylated hemoglobin (A1C) (A1C) variable and Fasting Plasma Glucose (FPG) variable – the primary variables to diagnose and control Diabetes Mellitus Kottgen et al. (2007); Association (2020), while each color represents the group that belongs to each patient. These graphics confirm the finding is explaining that our functional biclustering results are related to the quality of glycemic control. In general, incorporating functional information of CGM data provides new insight into diabetes data analysis (see for example Matabuena et al. (2021); Gaynanova et al. (2020)).

5 Discussion and future work

In this work, we have proposed a new formulation of the biclustering problem in Separable Hilbert Spaces inspired by the energy distance and the MMD, together with an efficient resolution algorithm based on the alternating application of the Kernel $k-$means algorithm on the data and covariate sets. Empirical results with vector data show the advantages of the method over existing methods in some settings, illustrating its usefulness in more general contexts such as functional data analysis. For instance, in the aforementioned functional diabetes example, where very few methodological approaches have been proposed.

With the choice of the appropriate Kernel, we can treat other biclustering algorithms such as k-means as a particular case, or formulate new graph-partitioning algorithms based on the methodology discussed here only by modifying the Kernel. Our algorithm can also run specific dimensionality reduction techniques such as principle component analysis or random projections. They are recommended when the data has a sparse structure. In a comprehensive analysis, we have seen that they help improve those cases. Variable selection can be an important landmark in many high-dimensional problems in which we know that Kernels, when p is growing with $n \to \infty$, can drastically deteriorate their performance (Ramdas et al., 2015).

Biclustering algorithms have been massively used in biology problems. However, their applications are not exclusive to this area, and they are also commonly used in other domains such as consumer analysis, sports, or text mining (Kasim et al., 2016).

In order to solve a biclustering algorithm in RKHS efficiently, we have introduced the constraint that the clusters formed at the subject and covariate level are mutually exclusive and that all covariates take values in a common domain. At first glance it may appear to be a hard constraint. However, the problem would be computationally prohibitive otherwise, and one would have to resort to heuristic algorithms due to the need for kernel-local computations that evaluate by brute force an objective function. Additionally, the results suggest our model's good performance in analyzing several gene expression data. This could be explained by the fact that the hypothesis is not restrictive in these domains, where genetic regions determine specific biological functions. With our clustering, we can achieve higher expressivity by identifying the covariates' sets most associated with those subtypes of patients. Naturally, some variables can

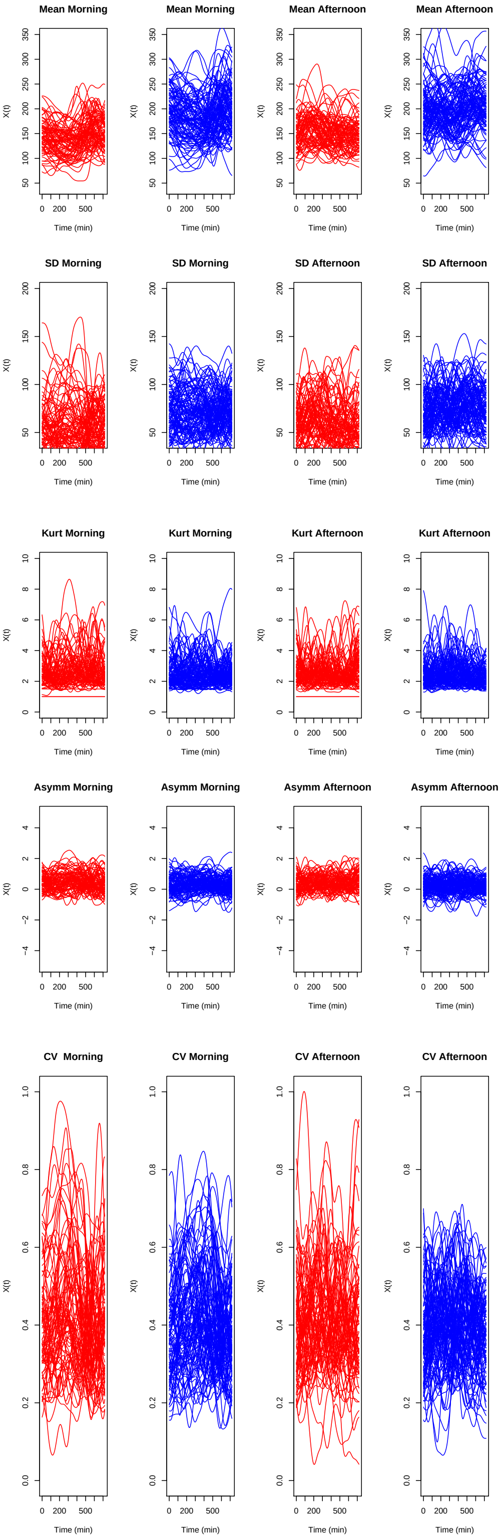

Fig. 1 Morning and afternoon functional profiles on the covariates. X^{mean}, X^{sd}, X^{CV}, X^{Skew}, X^{curtos}. Red (patients with reasonable glycaemic control). Blue (patients with worse glycaemic control).

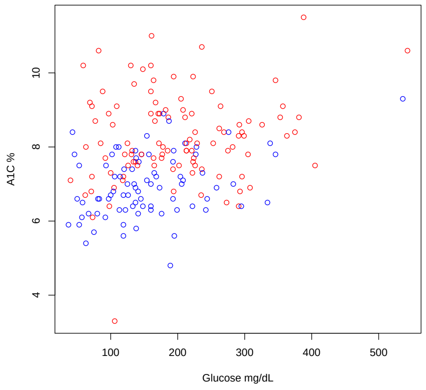

Fig. 2 Bidimensional plot: Glycosylated hemoglobin (A1C) vs. Fasting Plasma Glucose (FPG). Red (patients with reasonable glycaemic control). Blue (patients with worse glycaemic control).

be irrelevant and are noisily introduced in the estimated clusters, but this is inherent to any high-dimensional statistical procedure.

From the methodological point of view, it would be interesting to explore the possibility of solving other more general formulations of biclustering in Kernels for complex objects that can be defined in separable Hilbert spaces. From the mathematical perspective, it is not tricky to formalize these algorithms from the framework discussed here but, as we already mentioned, they are computationally prohibitive. Thus, heuristic optimization techniques are mandatory. The selection of a kernel together with its parameters is a crucial problem in the performance of the methods discussed here and remains open in a general way in kernel methods, although it is noteworthy to mention recent advances in this area last few years (Liu et al., 2020; Li et al., 2019).

Finally, a fundamental problem in clustering analysis is establishing the significance of the obtained clusters. In general, it is not straightforward, and to do it correctly, we would have to take into account the number of clusters built through the selective process (the problem of post-selection inference). In clustering, we only know the following paper (Gao et al., 2020) valid for tackling this task in hierarchical clustering algorithms to the best of our knowledge. However, perhaps using recent ideas with the MMD, incomplete U-V statistics (Lim et al., 2020) in the future, we can develop a specific procedure for the Kernel $k-$groups algorithm or, more specifically, in our $AKKB$ biclustering algorithm.

References

Association AD (2020) 2. classification and diagnosis of diabetes: Standards of medical care in diabetes—2020. Diabetes Care 43(Supplement 1):S14–S31, DOI 10.2337/dc20-S002, URL `https://care.diabetesjournals.org/content/43/Supplement_1/S14`, `https://care.diabetesjournals.org/content/43/Supplement_1/S14.full.pdf`

Ben-Dor A, Chor B, Karp R, Yakhini Z (2002) Discovering local structure in gene expression data: the order-preserving submatrix problem. In: Proceedings of the sixth annual international conference on Computational biology, pp 49–57

Berg C, Christensen JPR, Ressel P (1984) Harmonic analysis on semigroups: theory of positive definite and related functions, vol 100. Springer

Bergmann S, Ihmels J, Barkai N (2003) Iterative signature algorithm for the analysis of large-scale gene expression data. Physical review E 67(3):031902

Biau G, Devroye L, Lugosi G (2008) On the performance of clustering in hilbert spaces. IEEE Transactions on Information Theory 54(2):781–790

Bryan K, Cunningham P, Bolshakova N (2006) Application of simulated annealing to the biclustering of gene expression data. IEEE transactions on information technology in biomedicine 10(3):519–525

Busygin S, Prokopyev O, Pardalos PM (2008) Biclustering in data mining. Computers & Operations Research 35(9):2964–2987

Chen G, Sullivan PF, Kosorok MR (2013) Biclustering with heterogeneous variance. Proceedings of the national academy of sciences 110(30):12253–12258

Cheng Y, Church GM (2000) Biclustering of expression data. In: Ismb, vol 8, pp 93–103

Chi EC, Allen GI, Baraniuk RG (2017) Convex biclustering. Biometrics 73(1):10–19

Cho H, Dhillon IS, Guan Y, Sra S (2004) Minimum sum-squared residue co-clustering of gene expression data. In: Proceedings of the 2004 SIAM international conference on data mining, SIAM, pp 114–125

Dhillon IS (2001) Co-clustering documents and words using bipartite spectral graph partitioning. In: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pp 269–274

Flynn C, Perry P, et al. (2020) Profile likelihood biclustering. Electronic Journal of Statistics 14(1):731–768

Fraiman N, Li Z (2020) Biclustering with alternating k-means. arXiv preprint arXiv:200904550

França G, Vogelstein JT, Rizzo M (2020) Kernel k-groups via hartigan's method. IEEE Transactions on Pattern Analysis and Machine Intelligence

Galvani M, Torti A, Menafoglio A, Vantini S (2021) Funcc: A new bi-clustering algorithm for functional data with misalignment. Computational Statistics & Data Analysis 160:107219

Gao LL, Bien J, Witten D (2020) Selective inference for hierarchical clustering. arXiv preprint arXiv:201202936

Gaynanova I, Punjabi N, Crainiceanu C (2020) Modeling continuous glucose monitoring (cgm) data during sleep. Biostatistics

Getz G, Levine E, Domany E (2000) Coupled two-way clustering analysis of gene microarray data. Proceedings of the National Academy of Sciences 97(22):12079–12084

Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola A (2012) A kernel two-sample test. The Journal of Machine Learning Research 13(1):723–773

Helgeson ES, Liu Q, Chen G, Kosorok MR, Bair E (2020) Biclustering via sparse clustering. Biometrics 76(1):348–358

Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM computing surveys (CSUR) 31(3):264–323

de Jong TV, Moshkin YM, Guryev V (2019) Gene expression variability: the other dimension in transcriptome analysis. Physiological genomics 51(5):145–158

Kasim A, Shkedy Z, Kaiser S, Hochreiter S, Talloen W (2016) Applied biclustering methods for big and high-dimensional data using R. CRC Press

Kluger Y, Basri R, Chang JT, Gerstein M (2003) Spectral biclustering of microarray data: coclustering genes and conditions. Genome research 13(4):703–716

Kottgen A, Russell SD, Loehr LR, Crainiceanu CM, Rosamond WD, Chang PP, Chambless LE, Coresh J (2007) Reduced kidney function as a risk factor for incident heart failure: the atherosclerosis risk in communities (aric) study. Journal of the American Society of Nephrology 18(4):1307–1315

Lazzeroni L, Owen A (2002) Plaid models for gene expression data. Statistica sinica pp 61–86

Lee M, Shen H, Huang JZ, Marron J (2010) Biclustering via sparse singular value decomposition. Biometrics 66(4):1087–1095

Li CL, Chang WC, Mroueh Y, Yang Y, Póczos B (2019) Implicit kernel learning. In: The 22nd International Conference on Artificial Intelligence and Statistics, PMLR, pp 2007–2016

Lim JN, Yamada M, Jitkrittum W, Terada Y, Matsui S, Shimodaira H (2020) More powerful selective kernel tests for feature selection. In: International Conference on Artificial Intelligence and Statistics, PMLR, pp 820–830

Liu F, Huang X, Gong C, Yang J, Li L (2020) Learning data-adaptive non-parametric kernels. Journal of Machine Learning Research 21(208):1–39

Madeira SC, Oliveira AL (2004) Biclustering algorithms for biological data analysis: a survey. IEEE/ACM transactions on computational biology and bioinformatics 1(1):24–45

Matabuena M, Petersen A (2021) Distributional data analysis with accelerometer data in a nhanes database with nonparametric survey regression models. arXiv preprint arXiv:210401165

Matabuena M, Petersen A, Vidal JC, Gude F (2021) Glucodensities: A new representation of glucose profiles using distributional data analysis. Statistical methods in medical research p 0962280221998064

Ramdas A, Reddi SJ, Póczos B, Singh A, Wasserman L (2015) On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 29

Ramsay JO, Silverman BW (2007) Applied functional data analysis: methods and case studies. Springer

Romano Y, Sesia M, Candès E (2020) Deep knockoffs. Journal of the American Statistical Association 115(532):1861–1872

Sejdinovic D, Sriperumbudur B, Gretton A, Fukumizu K (2013) Equivalence of distance-based and rkhs-based statistics in hypothesis testing. The Annals of Statistics pp 2263–2291

Shabalin AA, Weigman VJ, Perou CM, Nobel AB (2009) Finding large average submatrices in high dimensional data. The Annals of Applied Statistics pp 985–1012

Sriperumbudur BK, Fukumizu K, Lanckriet GR (2011) Universality, characteristic kernels and rkhs embedding of measures. Journal of Machine Learning Research 12(7)

Tan KM, Witten DM (2014) Sparse biclustering of transposable data. Journal of Computational and Graphical Statistics 23(4):985–1008

Tanay A, Sharan R, Kupiec M, Shamir R (2004) Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. Proceedings of the National Academy of Sciences 101(9):2981–2986

Weinstock RS, DuBose SN, Bergenstal RM, Chaytor NS, Peterson C, Olson BA, Munshi MN, Perrin AJ, Miller KM, Beck RW, et al. (2016) Risk factors associated with severe hypoglycemia in older adults with type 1 diabetes. Diabetes Care 39(4):603–610

Wu K, Ding GW, Huang R, Yu Y (2020) On minimax optimality of gans for robust mean estimation. In: International Conference on Artificial Intelligence and Statistics, PMLR, pp 4541–4551

Zhang Z, Wang M, Nehorai A (2019) Optimal transport in reproducing kernel hilbert spaces: Theory and applications. IEEE transactions on pattern analysis and machine intelligence 42(7):1741–1754

A Mathematical Proofs

We start to introduce the L_2-Wasserstein distance in RKHS following the strategy established in Zhang et al. (2019). Let the input space $(\mathbb{H}^p, \mathcal{B}_{\mathbb{H}^p})$ be a measurable space with a Borel σ-algebra $\mathcal{B}_{\mathbb{H}^p}$, and $\mathbb{P}$ be the set of probability measures on $\mathbb{H}^p$ with finite-second order moments. Let $\mathcal{K}: \mathbb{H}^p \times \mathbb{H}^p \to \mathbb{R}^+$ be a positive definite kernel, and $\mathbb{H}_{\mathcal{K}}$ be the RKHS generated by $\mathcal{K}$. Let $\phi: \mathbb{H}^p \to \mathbb{H}_{\mathcal{K}}$ the corresponding embedding. For any $\mu \in \mathbb{P}$, let $\phi_\sharp \mu$ be the push-forward measure of μ[6]. Given $\mu, \eta \in \mathbb{P}$, the Wasserstein distance between push-forward measures $\phi_\sharp \mu$ and $\phi_\sharp \eta$ on $\mathbb{V}_K$ is defined as:

$$\gamma(\phi_\sharp \mu, \phi_\sharp \eta) = \left[\inf_{\substack{\pi_{\mathcal{K}} \in \\ \Pi(\phi_\sharp \mu, \phi_\sharp \eta)}} \int \|x - y\|^2_{\mathcal{K}_{I_j}} \, d\pi_{\mathcal{K}(x,y)} \right]^{1/2}, \qquad (19)$$

where $\Pi(\phi_\sharp \mu, \phi_\sharp \eta)$ is the set of joint probability measures on $\mathbb{H}^p \times \mathbb{H}^p$, with marginals $\phi_\sharp \mu$ and $\phi_\sharp \eta$.

Bellow, we enunciate an equivalent and computable formulation, which is fully determined by the kernel function.

Theorem 4 *Let $(\mathbb{H}^p, \mathcal{B}_{\mathbb{H}^p})$ a Borel space, and let the reproducing kernel $\mathcal{K}$ be measurable. Given $\mu, \eta \in \mathbb{P}$, we write:*

$$\gamma^{\mathcal{K}}(\mu, \eta) = \left[\inf_{\pi \in \Pi(\mu,\eta)} \int \|\phi(x) - \phi(y)\|^2 \, d\pi(x,y) \right]^{1/2}. \qquad (20)$$

where $\|\phi(x) - \phi(y)\|^2 = \mathcal{K}(x,y) + \mathcal{K}(y,y) - 2\mathcal{K}(x,y)$. Then:

- $\gamma^{\mathcal{K}}(\mu, \eta) = \gamma(\phi_\sharp \mu, \phi_\sharp \eta)$.
- If π^* is the minimized of Eq. (20), then $(\phi, \phi)_\sharp \pi*$ is a minimizer of Eq. (19), where $(\phi, \phi): \mathbb{H}^p \times \mathbb{H}^p \to \mathbb{H}_{\mathcal{K}} \times \mathbb{H}_{\mathcal{K}}$, is defined as $(\phi, \phi)(x,y) = (\phi(x), \phi(y))$.

Theorem 5 *Let A_n and c_n be δ_n-minimizers of the empirical clustering risk such that when $n \to \infty$, $\delta_n \to 0$. Then:*

- $\lim_{n\to\infty} \mathcal{W}(c_n, \mathcal{A}_n, X^{data}, \mu_n) = \mathcal{W}^*(\mu)$ *with probability one.*
- $\lim_{n\to\infty} \mathbb{E}(W(A_n, c_n, \mu_n)) = \mathcal{W}^*(\mu)$.

Now, we are in conditions to introduce the Lemmas that compose the main proof of this paper.

Lemma 1 *For any column partition $\mathcal{A} = \{\mathcal{I}_1, \ldots, \mathcal{I}_k\}$, and set-value function c, we have:*

$$|\sqrt{\mathcal{W}(\mathcal{A}, c, \mu))} - \sqrt{\mathcal{W}(\mathcal{A}, c, \eta))}| \leq \gamma^{\mathcal{K}}(\mu, \eta) \qquad (21)$$

[6] Given a probability measure μ on the input space, mapping the data through the implicit map ϕ, we are interested in the data distribution in RKHS. Such distribution is called the push-forward measure denoted as $\phi_\sharp \mu$, satisfying that for any subset A in RKHS, $\phi_\sharp \mu(A) = \mu(\phi^{-1}(A))$.

Proof Let $X \sim \mu$ and $Y \sim \eta$, achieve the infimum defining $\gamma^{\mathcal{K}}(\mu,\eta)$. Then:

$$\begin{aligned}
\sqrt{\mathcal{W}(\mathcal{A},c,\mu)} &= \left(\int \min_{1\leq j\leq k} \|\phi(x(\mathcal{I}_j)) - c(\mathcal{I}_j)\|^2_{\mathcal{K}_{\mathcal{I}_j}}\, d\mu(x)\right)^{\frac{1}{2}} \\
&= \left(\mathbb{E}(\min_{1\leq j\leq k} \|\phi(X(\mathcal{I}_j)) - c(\mathcal{I}_j)\|^2_{\mathcal{K}_{\mathcal{I}_j}})\right)^{\frac{1}{2}} \\
&\leq \mathbb{E}(\min_{1\leq j\leq k} \Bigg[\left\|\phi(X(\mathcal{I}_j)) - \phi(Y(\mathcal{I}_j))_{\mathcal{K}_{\mathcal{I}_j}}\right\|^2 \\
&\qquad\qquad + \|\phi(Y(\mathcal{I}_j)) - c(I_j)\|^2_{\mathcal{K}_{\mathcal{I}_j}}\Bigg])^{\frac{1}{2}} \\
&\leq \left(\mathbb{E}(\|\phi(X)) - \phi(Y))\|_{\mathcal{K}}\right) \\
&\quad + \left(\mathbb{E}(\min_{1\leq j\leq k} \|\phi(Y(\mathcal{I}_j)) - c(\mathcal{I}_j)\|^2_{\mathcal{K}_{\mathcal{I}_j}})\right)^{\frac{1}{2}} \\
&\leq \gamma^{\mathcal{K}}(\mu,\eta) + \sqrt{W(\mathcal{A},c,\eta)},
\end{aligned} \tag{22}$$

which implies one direction of the inequality enunciate. The other direction can be proved analogously.

Lemma 2 *Let $\mathcal{A}_n$ and c_n be a δ_n minimizers of the empirical clustering risk. Then:*

$$\sqrt{\mathcal{W}(\mathcal{A},c_n,\mu)} - \sqrt{\inf_{\mathcal{A}}\inf_{c}\mathcal{W}(\mathcal{A},c,\mu)} \leq 2\gamma^{\mathcal{K}}(\mu,\mu_n) + \sqrt{\delta_n}$$

Proof Let $\epsilon > 0$. In virtue of supremum axiom of real numbers, let $\mathcal{A}^*$ and c^* be the elements satisfying:

$$\begin{aligned}
\inf_{I}\inf_{c}\mathcal{W}(\mathcal{A},c,\mu) &\leq \mathcal{W}(\mathcal{A}^*,c^*,\mu) \\
&\leq \inf_{\mathcal{A}}\inf_{c}\mathcal{W}(\mathcal{A},c,\mu) + \epsilon.
\end{aligned} \tag{23}$$

For any $x \in \mathbb{R}$, we define $(x)_+ = \max(x,0)$. We have:

$$\begin{aligned}
&\sqrt{\mathcal{W}(\mathcal{A}_n,c_n,\mu)} - \sqrt{\inf_{\mathcal{A}}\inf_{c}\mathcal{W}(\mathcal{A},c,\mu)} \\
&\leq \sqrt{\mathcal{W}(\mathcal{A}_n,c_n,\mu)} - \sqrt{[\mathcal{W}(\mathcal{A}^*,c^*,\mu) - \epsilon)]_+} \\
&\leq \sqrt{\mathcal{W}(\mathcal{A}_n,c_n,\mu)} - \sqrt{W(\mathcal{A}^*,c^*,\mu)} + \sqrt{\epsilon} \\
&= \sqrt{\mathcal{W}(\mathcal{A}_n,c_n,\mu)} - \sqrt{\mathcal{W}(\mathcal{A}_n,c_n,\mu_n)} \\
&\quad + \sqrt{\mathcal{W}(\mathcal{A}_n,c_n,\mu_n)} - \sqrt{\mathcal{W}(\mathcal{A}*,c^*,\mu)} + \sqrt{\epsilon} \\
&\leq \sqrt{\mathcal{W}(\mathcal{A}_n,c_n,\mu)} - \sqrt{\mathcal{W}(\mathcal{A}_n,c_n,\mu_n)} \\
&\quad + \sqrt{\mathcal{W}(\mathcal{A}^*,c^*,\mu_n)} - \sqrt{\mathcal{W}(\mathcal{A}^*,c^*,\mu)} + \sqrt{\epsilon} + \sqrt{\delta_n} \\
&\leq 2\gamma^{\mathcal{K}}(\mu,\mu_n) + \sqrt{\epsilon} + \sqrt{\delta_n},
\end{aligned}$$

where we obtain the last inequality of the previosly lemma.

Lemma 3 *Under the previously conditions, we can establish the following converge results with L_2-Wasserstein distance:*

- $\lim_{n\to\infty}\gamma^{\mathcal{K}}(\mu,\mu_n) = 0$ *with probability one and*
- $\lim_{n\to\infty}\mathbb{E}(\gamma^{K}(\mu,\mu_n)) = 0$.

Proof A Weak Law of Large Numbers for Empirical Measures **?** guarantee the convergence $\mu_n \to \mu$. In addition, we know by Skorokhod's representation theorem that exist $Y_n \sim \mu_n$ and $Y \sim \mu$ of such way that $Y_n \to Y$ with probability equal one. By virtue of continuos mapping theorem mapping is verified also that $\phi(Y_n) \to \phi(Y)$. Then, with the application of triangle inequality, we can establish that:

$$\begin{aligned}
&2\|\phi(Y_n)\|^2_{\mathcal{K}} + 2\|\phi(Y)\|^2_{\mathcal{K}} - \|\phi(Y_n) - \phi(Y)\|^2_{\mathcal{K}} \\
&\quad \geq \|\phi(Y_n)\|^2_{\mathcal{K}} + \|\phi(Y)\|^2_{\mathcal{K}} - \|\phi(Y_n)\|_{\mathcal{K}}\|\phi(Y)\|_{\mathcal{K}} \\
&\quad \geq 0
\end{aligned}$$

Then, Fatou's lemma implies that:

$$\begin{aligned}
&\lim_{n\to\infty}\inf\mathbb{E}(2\|\phi(Y_n)\|^2_{\mathcal{K}} + 2\|\phi(Y)\|^2_{\mathcal{K}} - \|\phi(Y) - \phi(Y_n)\|^2_{\mathcal{K}}) \\
&\quad \geq 4\mathbb{E}(\|\phi(Y)\|^2_{\mathcal{K}}).
\end{aligned}$$

Since $\mathbb{E}(\phi(Y_n)^2) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\phi(X_i)^2 = \mathbb{E}(\phi(X))^2 = \mathbb{E}(\phi(Y))^2$, we deduce that $\lim_{n\to\infty}\mathbb{E}(\|\phi(Y_n) - \phi(Y)\|^2_{\mathbb{K}}) = 0$, which implies that $\lim_{n\to\infty}\gamma^{\mathcal{K}}(\mu,\mu_n) = 0$ with probability equal one.
For the second part of the proof, let $\Pi(\mu,\mu_n)$ be the set of joint probability measures on $\mathbb{H}^p \times \mathbb{H}^p$ with marginal probabilities μ and μ_n respectively. By definition, the L_2-Wasserman distance can be written as:

$$\gamma^{K}(\mu,\mu_n) = \inf_{\pi\in\Pi(\mu,\mu_n)}\int\left[\|\phi(x) - \phi(y)\|^2_{\mathcal{K}}\, d\pi(x,y)\right]^{1/2}. \tag{24}$$

Let $C > 0$ be an arbitrary non-negative constant, and let $\mathcal{D}$ be the subset of $\mathbb{H}^p \times \mathbb{H}^p$ defined by:

$$\mathcal{D} = \{(x,y) \in \mathbb{H}^p \times \mathbb{H}^p : \max(\|\phi(x)\|_{\mathcal{K}}, \|\phi(y)\|_{\mathcal{K}}) \leq C\} \tag{25}$$

For any $\pi \in \Pi(\mu,\mu_n)$, we have:

$$\begin{aligned}
&\int \|\phi(x) - \phi(y)\|^2_{\mathcal{K}}\, d\pi(x,y) \\
&= \int_{\mathcal{D}} \|\phi(x) - \phi(y)\|^2_{\mathcal{K}}\, d\pi(x,y) + \int_{\mathcal{D}^C} \|\phi(x) - \phi(y)\|^2_{\mathcal{K}}\, d\pi(x,y) \\
&\leq \int_{\mathcal{D}} \|\phi(x) - \phi(y)\|^2_{\mathcal{K}}\, d\pi(x,y) \\
&\quad + 2\int_{\mathcal{D}^c} \|\phi(x)\|^2_{\mathcal{K}}\, 1\{\|\phi(x)\|_{\mathcal{K}} \\
&> C\}d\mu(x) + 2\int_{\mathcal{D}^c} \|\phi(y)\|^2_{\mathcal{K}}\, 1\{\|\phi(y)\|_{\mathcal{K}} \\
&> C\}d\mu_n(y) + 2\int_{\mathcal{D}^c} \|\phi(x)\|^2_{\mathcal{K}}\, 1\{\|\phi(x)\|_{\mathcal{K}} \\
&\leq C, \|\phi(y)\|_{\mathcal{K}} \\
&> C\}d\pi(x,y) + 2\int_{\mathcal{D}^c} \|\phi(y)\|^2_{\mathcal{K}}\, 1\{\|\phi(y)\|_{\mathcal{K}} \\
&\leq C, \|\phi(x)\|_{\mathcal{K}} \\
&> C\}d\pi(x,y).
\end{aligned} \tag{26}$$

With the application of Markov's inequality and taking the infimun on $\Pi(\mu,\mu_n)$ two sides of the previously inequality, we can see that:

$$\begin{aligned}
\mathbb{E}(\gamma^{K}(\mu,\mu_n)) \leq \mathbb{E}\Bigg[&\inf_{\pi\in\Pi(\mu,\mu_n)}\int_{\mathcal{D}} \|\phi(x) - \phi(y)\|^2_{\mathcal{K}}\, d\pi(x,y)]^{1/2} \\
&+ 8\int_{\mathcal{D}^c} \|\phi(x)\|^2_{\mathcal{K}}\, 1\{\|\phi(x)\|_{\mathcal{K}} \\
&> C\}\Bigg]\, d\mu(x).
\end{aligned} \tag{27}$$

For a fixed $C \geq 0$, the first term of right-hand side goes to 0 as $n \to \infty$ according to the first part of this lemma and the Lebesgue dominated theorem. Since in our initial assumptions, $E(\|X\|^2) \leq \infty$, the second term go to 0 so $C \to \infty$. Now, in virtue of the bound establish in Lemma 2 and the last results we can deduce as consequence this theorem.

Theorem 6 *AKKB biclustering algorithm converges in a finite number of steps.*

Proof The key point of all optimization strategy is consider alternatively the following optimization problem so at individual and covariate level:

$$j^* = \arg \max_{l=1,\dots,k|l\neq j} \Delta \mathcal{Q}^{(j\to l)}(x), \forall x \in X^{data}. \tag{28}$$

By construction, since the resolution of Eq. (28) sure that the cost function $\mathcal{Q}$ is monotonically increasing at each iteration, and there are a finite number of distinct cluster assignments, the algorithm converges in a finite number of steps.

Theorem 7 *The complexity of AKKB biclustering algorithm is $\mathcal{O}(k^2n^4)$, where k is the number of clusters, n is the number of data points.*

Proof According França et al. (2020), the complexity of kernel k-group algorithm is the order $\mathcal{O}(kn^2)$. As our algorithm consisted on resolved independently and alternatively kernel k-group algorithm at covariate and individual level, the overall complexity is $\mathcal{O}(k^2n^4) = \mathcal{O}(kn^2) * \mathcal{O}(kn^2)$.

B Analysis sensibility of kernel bandwidth parameters

To analyze the impact of kernel bandwidth parameters in our *AKKB* algorithm, we use a grid of potential candidates of σ_{data}, $\sigma_{variables}$ using as a reference the median heuristic. Then, we evaluate the model performance. For this purpose, we use some genetic expression datasets of the *Schliep lab* repository[7] on which true structures are annotated. In the analysis, we consider the following datasets: Alizadeh2, Bibtner2, Chen2, Golub2, Gordon2, Shipp2, Singh2, West2.

Bellow, we show different graphics of the results in the Figures 3-10. The axis-z of graphic, represent the precision, axis-x, σ_{data}, axis-y $\sigma_{variable}$. We can see that the selected parameter of the kernel are critical in the accuracy of our AKKB Algorithm at the data and covariate level and can modify the precision results in more than 40 % percent in some cases respect suboptimal parameter configuration.

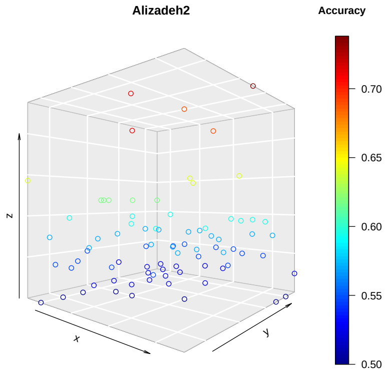

Fig. 3 Results of applying our Biclustering algorithm in the dataset Alizadeh2 varying σ_{data} and $\sigma_{variables}$ in a grid of potential values according to median heuristic criteria.

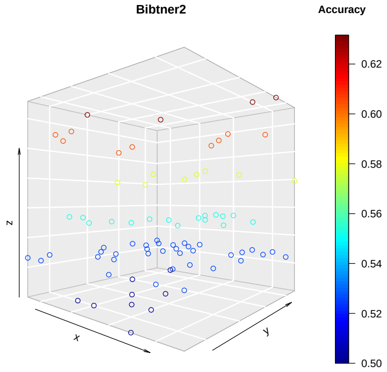

Fig. 4 Results of applying our Biclustering algorithm in the dataset Bibtner2 varying σ_{data} and $\sigma_{variables}$ in a grid of potential values according to median heuristic criteria.

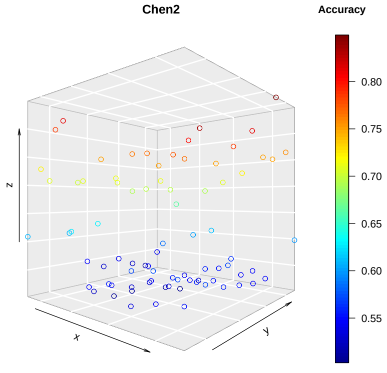

Fig. 5 Results of applying our Biclustering algorithm in the dataset Chen2 varying σ_{data} and $\sigma_{variables}$ in a grid of potential values according to median heuristic criteria.

[7] Available at `https://schlieplab.org/Static/Supplements/CompCancer/datasets.htm`

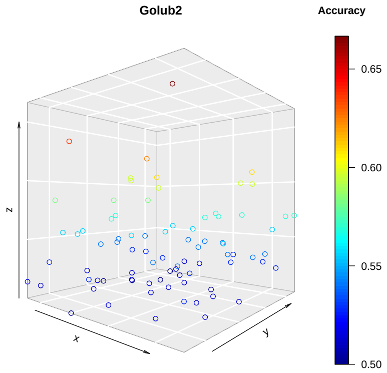

Fig. 6 Results of applying our Biclustering algorithm in the dataset Golub2 varying σ_{data} and $\sigma_{variables}$ in a grid of potential values according to median heuristic criteria.

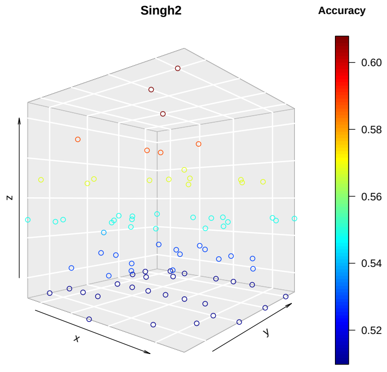

Fig. 9 Results of applying our Biclustering algorithm in the dataset Singh2 varying σ_{data} and $\sigma_{variables}$ in a grid of potential values according to median heuristic criteria.

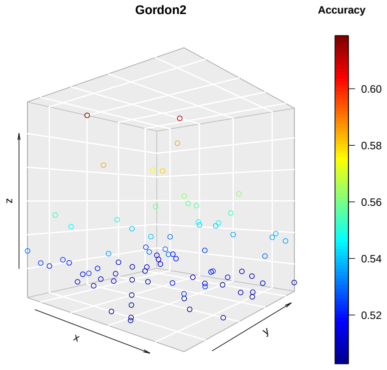

Fig. 7 Results of applying our Biclustering algorithm in the dataset Gordon2 varying σ_{data} and $\sigma_{variables}$ in a grid of potential values according to median heuristic criteria.

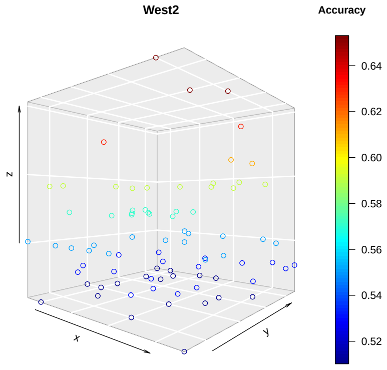

Fig. 10 Results of applying our Biclustering algorithm in the dataset West2 varying σ_{data} and $\sigma_{variables}$ in a grid of potential values according to median heuristic criteria.

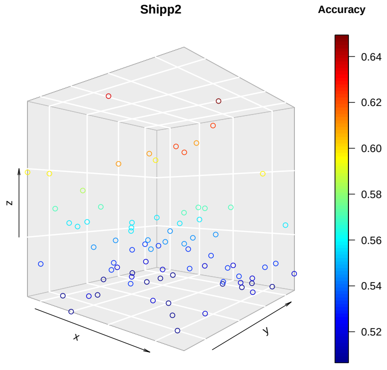

Fig. 8 Results of applying our Biclustering algorithm in the dataset Shipp2 varying σ_{data} and $\sigma_{variables}$ in a grid of potential values according to median heuristic criteria.