Clustering measure-valued data with Wasserstein barycenters

Research Article

G. N. Domazakis\textsuperscript{a,d}, D. Drivaliaris\textsuperscript{e}, S. Koukoulas\textsuperscript{f}, G. I. Papayiannis\textsuperscript{a,c}, A. E. Tsekrekos\textsuperscript{b}, and A. N. Yannacopoulos\textsuperscript{a}

\textsuperscript{a}Athens University of Economics & Business, Department of Statistics, Stochastic Modeling and Applications Laboratory, Athens, GR
\textsuperscript{b}Athens University of Economics & Business, Department of Accounting and Finance, Athens, GR
\textsuperscript{c}Hellenic Naval Academy, Section of Mathematics, Piraeus, GR
\textsuperscript{d}University of Sussex, Department of Mathematics, Brighton, UK
\textsuperscript{e}University of the Aegean, Department of Financial and Management Engineering, Chios, GR
\textsuperscript{f}University of the Aegean, Department of Geography, Mytilene, GR

Abstract

In this work, clustering schemes for measure-valued data are proposed, i.e. data that their structure can be more efficiently represented as probability measures instead of points on \(\mathbb{R}^d\), employing the concept of probability barycenters as defined with respect to the Wasserstein metric. Such type of clustering approaches are highly appreciated in many fields where the observational/experimental error is significant (e.g. astronomy, biology, remote sensing, etc.) or the data nature is more complex and the traditional learning algorithms are not applicable or effective to treat them (e.g. network data, interval data, high frequency records, matrix data, etc.). Under this perspective, each observation is identified by an appropriate probability measure and the proposed clustering schemes rely on discrimination criteria that utilize the geometric structure of the space of probability measures through core techniques from the optimal transport theory. The advantages and capabilities of the proposed approach are illustrated through an application in finance where eurozone countries are separated into groups according to their observed government bond yield curves with very interesting findings.

Keywords: clustering; geodesics; measure-valued data; Wasserstein barycenter; yield curves;

1 Introduction

The majority of statistical learning methods refer to data which are considered as points on some high-dimensional Euclidean space (e.g. \(\mathbb{R}^d\) for \(d > 1\)), and treat the problem of clustering them into like classes depending on their distance or similarity occurring from their location on the underlying space. Such methods have led to the development of powerful algorithms, as for example the K-means algorithm, which are widely used in statistical learning\textsuperscript{19} \textsuperscript{34}. On the other hand, the implementation of these algorithms has highlighted the importance of geometrical features such as dimension or the notion of distance employed in the underlying space on the results of the clustering process \textsuperscript{1} \textsuperscript{2} \textsuperscript{3} \textsuperscript{4}.

However, in the new era of data science, the nature of data structures has significantly been enriched and evolved so that in many cases data can no longer be sufficiently represented as elements of some high dimensional Euclidean space like \(\mathbb{R}^d\). As a result, traditional methods in

\*Corresponding author: Patission Str. 76, Athens, 10434, Greece. E-mail: gpapagiannis@aueb.gr
statistical learning may not be appropriate for analysing such types of complex data. Examples of such cases can be functional data \cite{12, 22, 31}, data sampled at high frequency \cite{17, 25}, interval data \cite{15, 16, 21}, spatial data \cite{18}, uncertain data \cite{1}, matrix-valued data \cite{33}, etc.

In this paper, we propose a framework under which complex data, which may not sufficiently be represented as vectors in $\mathbb{R}^d$, can be treated as elements of some space of probability measures, which when endowed with an appropriate metric structure, the Wasserstein metric, can lead to a better representation. After recognizing the appropriate underlying space for the data, we develop a clustering scheme in Wasserstein space, based on its geometry and concept of the barycenter in order to characterize the cluster centers as well as the concept of geodesic to quantify similarity of observations within a particular cluster. Moreover, indices that (a) quantify the homogeneity of the induced clusters and (b) reveal the appropriate selection for number of clusters are developed relying on geometric arguments on the space of probability measures. The proposed method is applied to an illustrative example in quantitative finance and in particular, the clustering of sovereign bonds to groups and then the comparison with the credit classes that have been defined by the financial industry experts.

2 Clustering in the Wasserstein space

In this section, we investigate clustering techniques that can be used for learning purposes in the space of probability measures (Wasserstein space) with the selected type of learning procedure (supervised or unsupervised) to be determined by the application. In particular, an extension of the well known K-means algorithm within the framework of probability measures is considered. We start by presenting some motivation from certain classes of complex data, which are more suitably treated as measure valued data, and introduce the problem of clustering such observations in the space of probability measures. Then, the extension of the original K-means algorithm to the probabilistic setting is discussed, i.e. the case where the observations to be clustered are probability distributions adopting the metric framework of Wasserstein spaces, in order to replace the usual Euclidean framework commonly employed in clustering studies. Employing the geometry of Wasserstein space, and in particular its geodesic structure, we propose a clustering criterion which quantifies the compactness of clusters and assists in the choice of the number of clusters.

2.1 Motivation and general framework

Consider a collection of data points $x_1, x_2, \ldots, x_n$ on $\mathbb{R}^d$, corresponding to some type of measurements. The standard clustering procedure would involve the application of a clustering algorithm, e.g. the K-means algorithm, considering these observations as points on the Euclidean space $\mathbb{R}^d$, endowed with the standard Euclidean metric, and the observations will be clustered according to their distance from fictitious centers located at certain positions in $\mathbb{R}^d$, in terms of the Euclidean distance. Such clustering procedures have been proved very successful in understanding the structure of data in statistical learning procedures. Even in the case where the data can be efficiently represented as elements of $\mathbb{R}^d$, the choice of metric with which this space is endowed may affect the clustering result (see e.g. \cite{20}).

However, there are many instances where the data under consideration cannot naturally be represented on a linear space, either because of special constraints or because of their collection type and sampling method. In such cases the K-means algorithm has to be modified in order to take into account the non-linearity of the underlying space.

A large class of data types such as the ones described above, can be understood as measures and will henceforth be referred to as measure-valued data. In order to motivate this statement, we present below four particular examples.
Example 2.1 (Data that can be understood as probability measures). In many cases, we can assume that what we actually need from our observations, is not the value of the observation as such, but the collective behaviour of relevant data. For example, one may wish to consider the distribution of a physical quantity (e.g. temperature, wind speed) for a certain time period and wish to understand the differences between the distributions of the same quantity on different time periods or locations. Under this perspective, the individual data related to the particular time periods or locations will be used to obtain (empirical) estimations of the probability measures describing the quantity. Wishing to study the aggregate characteristics, which are described by the aforementioned probability measure, one may consider as observed data the induced probability measures, therefore leading to the framework of measure-valued data. Issues such as clustering or classification of these probability measures arise naturally (e.g. in remote sensing).

Example 2.2 (Data point clouds). Often, spatial data can be understood as data point clouds that carry some heterogeneity. A particular example can be the onset and evolution of an epidemic, in which the spatial location and severity of several incidents is of interest. Certain spatial locations may be hit with different severity and to classify such regions into clusters and assess the spatio-temporal allocation of risk. Data point clouds can be identified as empirical probability measures using convex combination of Dirac measures of the incidents taking into account location, frequency and severity or smoothed representations using appropriate kernel-based estimators.

Example 2.3 (Matrix-valued data). Other types of data can be perceived as measure-valued data, hence making feasible their treatment as elements of a Wasserstein space. An example of particular interest are matrix valued data and in particular correlation or covariance matrix data. As well known Gaussian measures on $\mathbb{R}^d$ are uniquely determined by their mean and their covariance matrix hence there is an one-to-one correspondence between the space of centered Gaussian measures and the manifold of positive-definite matrices (the natural space for covariance matrices) (see, e.g. [26, 37]). Hence, one can metrize the later using the Wasserstein metric in the space of Gaussian measures. This metric is usually called the Bures-Wasserstein metric ([6]) and can be calculated explicitly in terms of formula (1) with zero locations. As a result, the proposed algorithm in Section 2.2 can be directly implemented without any modifications for clustering matrix-valued data. In fact, some of the applications presented in paper can be understood as clustering matrix valued data, as for example the application on bond returns in Section 3.1.

Example 2.4 (Uncertain data). In many cases the observations of a dataset are not that certain due to the effect of experimental or observational error, noise to data, inconsistencies of the monitoring equipment, etc. The uncertainty of the measurements introduces some fuzziness as to the exact location of the points on the Euclidean space. One way to take into account this fuzziness is to substitute the data points with probability measures $\mu_1, \mu_2, \ldots, \mu_n$, each one centered at its location vector $m_i = x_i \in \mathbb{R}^d$, with a covariance matrix $S_i \in \mathbb{R}_{+}^{d \times d}$, with $\mathbb{R}_{+}^{d \times d}$ denoting the set of symmetric and positive definite matrices of dimension $d$, modeling the uncertainty as to the exact location of the point. If geometric intuition is of any use, we are substituting each data point $x_i \in \mathbb{R}^d$, with an ellipsoid in $\mathbb{R}^d$, located at $m_i = x_i$, whose principal axes are indicated by the eigenvectors of the covariance matrix $S_i$, and we assign a probability that the actual measurement is located within this ellipsoid in terms of the probability measure $\mu_i$. The different dispersion matrix $S_i, S_j$ corresponding to the different probability measures $\mu_i, \mu_j$ model the possible differential uncertainty assigned to each measurement. This is a very reasonable and plausible assumption. For example, each measurement may be obtained by a different source or a different measurement procedure, e.g. they may correspond to measurements of a quantity performed at different geographic locations, where remoteness
introduces observational errors (one can easily consider applications in observational astronomy, meteorology, medical imaging, remote sensing, etc).

Motivated by the above examples we consider a collection of measures \( \mu_i \) for \( i = 1, 2, ..., n \) that arose from an observation process or some statistical experiments. We will restrict our attentions to the case of the family of Location-Scale probability laws, and assume that each probability measure is sufficiently characterized by a location and dispersion pair, \((m_i, S_i)\). In particular, we will treat our observations as elements in the set of probability measures with finite moments of second order denoted by \( \mathcal{P}_2(\mathbb{R}^d) \), endowed with the topology induced by the quadratic Wasserstein metric (see e.g. \cite{32, 33}). This is a natural choice for metrizing the space of probability measures consistent with the weak-* topology, with the Wasserstein metric providing an appropriate notion of distance between two probability measures, taking efficiently into account the geometrical structure and non-linear nature of the underlying space.

We now proceed to the question of clustering these measure-valued observations. Since clustering requires a notion of distance which is appropriate for quantifying divergence between the objects under consideration, we propose the use of the Wasserstein metric space for this purpose. For example, this choice of metric space setting, can accommodate all the applications mentioned in Examples \ref{example1}, \ref{example2}. For instance, concerning Example \ref{example2} our proposal properly takes into account the observational uncertainty (which may be of different degrees depending on the observation) of the original measurements and when combined with the K-means clustering algorithm is expected to lead to more robust results. Note that in the limit as \( S_i \to 0 \), the probability measure \( \mu_i \) degenerates to Dirac mass \( \delta_{x_i} \) on \( \mathbb{R}^d \), and the Wasserstein distance \( W_2(\mu_i, \mu_j) \) degenerates to the standard Euclidean distance \( \|x_i - x_j\|_2 \) in \( \mathbb{R}^d \). Similar comments can be made also for the other examples.

We close the discussion by indicating why we expect that a K-means algorithm based on the Wasserstein distance leads to different results. This is best illustrated, within an example in which in terms of a limiting procedure we may pass from a description of data in Euclidean space to its description in measure space; the example of uncertain data (Example \ref{example2}) is ideally suited for this purpose. Consider that at some point in the course of the clustering procedure we have two potential cluster centers \( \bar{x}_{c1}, \bar{x}_{c2} \in \mathbb{R}^d \) and an observation \( y \in \mathbb{R}^d \) that has to be assigned to one of these two centers. Assume without loss of generality that \( \|y - \bar{x}_{c1}\|_2 < \|y - \bar{x}_{c2}\|_2 \) so that observation \( y \) is assigned to the cluster \( c_1 \) centered at \( \bar{x}_{c1} \). Consider now the case where observational uncertainty is assigned to the observations. Since the potential cluster centers are obtained dynamically as centroids of subsets of the original observations, the uncertainty concerning the observations propagates to some uncertainty as to the true location of the centers \( \bar{x}_{c1}, \bar{x}_{c2} \). In the framework described above, this uncertainty is modeled by considering both the observation to be assigned to a particular cluster, as well as the centers of the two potential clusters the observations belong to, as probability measures \( \mu, \bar{\mu}_{c1}, \bar{\mu}_{c2} \in \mathcal{P}_2(\mathbb{R}^d) \), with locations and dispersions \((m, S) = (x, S), (\bar{m}_1, \bar{S}_1) = (\bar{x}_{c1}, \bar{S}_{c1}), (\bar{m}_2, \bar{S}_2) = (\bar{x}_{c2}, \bar{S}_{c2}) \), respectively, with the covariance matrices reflecting the extent and the nature of the observational uncertainty.

Let us assume that these probability measures belong to the same Location-Scatter family, say without loss of generality the Gaussian family, which is a good choice for modeling observational uncertainty on account of the central limit theorem. Taking this uncertainty into account we must necessarily use the Wasserstein distance between the resulting probability measures as a criterion of assigning the point \( \mu \in \mathcal{P}_2(\mathbb{R}^d) \) to one of the two clusters centered at the points \( \bar{\mu}_{c1}, \bar{\mu}_{c2} \in \mathcal{P}_2(\mathbb{R}^d) \). Using this criterion we will assign the point \( \mu \) to the cluster \( c \in \{c_1, c_2\} \) by setting \( c = \arg\min_{c \in \{c_1, c_2\}} \{W_2(\mu, \bar{\mu}_{c1}), W_2(\mu, \bar{\mu}_{c2})\} \). According to \ref{example2} (please see Section \ref{example2}), we have that

\[
W_2^2(\mu, \bar{\mu}_{c_j}) = \|m - \bar{m}_{c_j}\|_2^2 + Tr \left( S + \bar{S}_{j} - 2(S^{1/2}\bar{S}_{j}S^{1/2})^{1/2} \right), \quad j = 1, 2.
\]

From this expression it becomes clear how uncertainty may alter the assignment to a particular cluster. Suppose that \( \|m - \bar{m}_{c_1}\|_2 < \|m - \bar{m}_{c_2}\|_2 \), so that the standard Euclidean K-means
assigns this observation to cluster \(c_1\). If
\[
Tr \left( S + \bar{S}_1 - 2(S^{1/2} \bar{S}_1 S^{1/2})^{1/2} \right) > Tr \left( S + \bar{S}_2 - 2(S^{1/2} \bar{S}_2 S^{1/2})^{1/2} \right),
\]
then using the Wasserstein distance the same observation will be assigned to cluster \(c_2\) instead. This difference in assignment is attributed as an effect of observational uncertainty and will lead to more robust clustering of the uncertain data. As a nice geometric intuitive approach consider ellipsoids of different orientations that have to be accommodated to different clusters. Clearly, the orientation of each ellipsoid plays a role as to which cluster it is appended to.

### 2.2 The Wasserstein K-means algorithm

Motivated by the discussion in the previous section we consider the following problem: Given a collection of measure valued data (which arise from complex data e.g. of the type mentioned in Examples 2.1-2.4) cluster the data into like groups. On account of the nonlinear nature of the space of probability measures (or probability distributions) the standard K-means algorithm in Euclidean space is no longer sufficient for this task. On the other hand, clustering is heavily based on the concept of distance between the objects which are to be separated, hence an appropriate metrization of the space of probability measures has to be adopted before we may proceed further to our goal. A very appropriate concept of distance for quantifying divergences between different probability measures is the Wasserstein distance which is compatible with the topology of the space of probability measures (see [32, 35] for justification of this metric).

The Wasserstein distance is in general expressed as the standard optimal transport problem originated by [28]. Given a quadratic cost functional, for any two probability measures \(\mu, \nu \in \mathcal{P}(\mathbb{R}^d)\) the Wasserstein distance is mathematically expressed as the minimization problem

\[
W_2(\mu, \nu) = \inf_{T \in \mathcal{T}M(\nu)} \left\{ \int_{\mathbb{R}^d} \|x - T(x)\|^2 d\mu(x) \right\}^{1/2}
\]

where \(\mathcal{T}M(\nu) := \{ T : \mathbb{R}^d \to \mathbb{R}^d : X \sim \mu, T(X) \sim \nu \}\) i.e. denotes the space of transport maps that transform the random variable \(X \in \mathbb{R}^d\) which is distributed according to the probability measure \(\mu\) to the random variable \(Y := T(X) \in \mathbb{R}^d\) distributed according to the probability measure \(\nu\). In other words, the optimal transport map \(T^*\) pushes-forward the measure \(\mu\) to the measure \(\nu\), i.e. \(\nu(A) = T^*_\# \mu(A)\) for any Borel subset \(A \subset \mathbb{R}^d\).

Assume that there is a collection of \(n\) probability measures \(\mathcal{M} := \{\mu_1, \mu_2, ..., \mu_n\}\) on \(\mathbb{R}^d\). The problem under consideration is to separately this collection of measures into \(K\) clusters with the criterion of their “affinity” in the space of probability measures on \(\mathbb{R}^d\) \((\mathcal{P}(\mathbb{R}^d))\), requiring finite second moments, \(\mathcal{P}_2(\mathbb{R}^d)\), endowed with the metric structure induced by the quadratic Wasserstein metric, \(W_2(\cdot, \cdot)\). As the initial step of the clustering procedure, \(K\) points on \(\mathcal{P}_2(\mathbb{R}^d)\) (probability measures) should be chosen as the initial cluster centers (this measures could be chosen either randomly or according to some rule, e.g. the most distant measures between them). Then, the cluster membership of each measure \(\mu_i \in \mathcal{M}\) is determined according to the minimal Wasserstein distance criterion of this measure from each one of the centers (note that the minimal Wasserstein distance is equivalent to the minimal length curve connecting each point with each cluster center). Next, the cluster centers are updated according to the mean sense derived by the Wasserstein barycenter of the measures classified to each cluster, i.e. the notion of mean element in a set of probability measures employing the metric sense of the Wasserstein distance. In particular, given that the measures \(\mu_{k,i}\) for \(i = 1, 2, ..., n_k\) are the measures that have been assigned to the \(k\)-th cluster, the new \(k\)-th cluster center is determined through the solution of the minimization problem

\[
\bar{\mu}_k = \arg \min_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \frac{1}{n_k} \sum_{i=1}^{n_k} W_2^2(\nu, \mu_{k,i}),
\]
Clustering with Wasserstein Barycenters

i.e. through the determination of the Wasserstein barycenter (or just probability barycenter) of the measures that have been assigned to this particular cluster (later we will refer to this probability barycenter also as the $k$-th local barycenter). General existence, uniqueness and consistency results about Wasserstein barycenters for measures on $\mathbb{R}^d$ have been established in [3, 23, 24]. The above steps of the clustering algorithm are repeated until no significant change on the cluster centers or their memberships occurs. The algorithm is presented briefly below (Algorithm 1).

**Algorithm 1: Wasserstein K-Means**

1. Choose $K$ measures out of the original observations as the initial cluster centers, $\bar{\mu}_k$, $k = 1, \ldots, K$.

2. At step $m$ assign each of the measures $\mu_i$, $i = 1, \ldots, n$, to one of the $K$ clusters, choosing the cluster membership $k(i)$ according to the rule

   $$k(i) = \arg\min_{k \in \{1, 2, \ldots, K\}} W^2_2(\mu_i, \bar{\mu}^{(m)}_k).$$

3. Update the cluster centers $\{\bar{\mu}^{(m)}_k\}_{k=1}^K$, through the rule

   $$\bar{\mu}^{(m)}_k := \arg\min_{\nu \in P_2(\mathbb{R}^d)} \frac{1}{n_k} \sum_{i=1}^{n_k} W^2_2(\nu, \mu^{(m)}_{k,i}),$$

   for all $k = 1, 2, \ldots, K$, where $\mu^{(m)}_{k,i}$ denotes the $i$-th measure of the $n_k$ that have been assigned to the $k$-th cluster at step $m$.

4. Repeat steps 2-3 until the cluster centers do not change significantly.

2.3 The geometric interpretation of the clustering procedure

In this subsection, we discuss the geometric interpretations that can be extracted by the clustering procedure of the Wasserstein K-Means algorithm (Algorithm 1). Following the previous discussion, each one of the $K$ cluster centers under the proposed setting coincides with the Wasserstein barycenter of all the measures that belong to that cluster. Denote by $\bar{\mu}_k$ for $k = 1, 2, \ldots, K$ the cluster centers or local barycenters and as $\bar{\mu}_0$ the barycenter of all measures or the global barycenter. Since all the measures involved belong to the same space we are able to define the shortest paths that connect these points, i.e. their geodesic curves. A very interesting characterization of these curves is provided by [27] called the McCann’s interpolant and relies on the concept of optimal transport maps.

According to McCann, the geodesic curve (or shortest path) that connects two measures $\mu, \nu$ in Wasserstein space can be formulated as the set of parameterized curves

$$\gamma(t) = \{[(1-t)I + tT]#\mu\}_t, \quad t \in [0, 1],$$

where $\gamma(0) = \mu$ and $\gamma(1) = \nu$, $T$ is the optimal transport map that pushes-forward the measure $\mu$ to the measure $\nu$ and $I$ denotes the identity operator, i.e. $I#\mu = \mu$. According to this construction, the minimum length connections between the global barycenter $\bar{\mu}_0$ (i.e. the Wasserstein barycenter of all probability measures under study) and the local barycenters $\bar{\mu}_k$ for $k = 1, \ldots, K$ are characterized by the parameterized curves

$$\gamma_k(t) = \{[(1-t)I + tT^{(k)}]#\bar{\mu}_0\}_t, \quad t \in [0, 1], \quad k = 1, \ldots, K$$

where $T^{(k)}$ denotes the optimal transport map that pushes forward the global barycenter to the $k$-th local barycenter, i.e. $T^{(k)}_#\bar{\mu}_0 = \bar{\mu}_k$ and it holds that $\gamma_k(0) = \bar{\mu}_0$ and $\gamma_k(1) = \bar{\mu}_k$. Using these
shortest path connections between the local barycenters and the global barycenter we give an attempt in characterizing the cluster membership of the probability measures involved in the clustering task which will lead to a compactness index for the shaped clusters (see Section 2.5).

Let \( \mu_i \) be an arbitrary measure that needs to be assigned to one of the \( K \) clusters. It is clear from the K-means algorithm, that the cluster where the measure \( \mu_i \) will be assigned to, is the cluster whose center \( \bar{\mu}_k \) will have the minimal (Wasserstein) distance from \( \mu_i \), or in other words, the geodesic curve that connects the measure \( \mu_i \) with the specified local barycenter \( \bar{\mu}_k \) must have the smallest length comparing to the lengths of each one of the other geodesic curves \( \{ \gamma_k(\cdot) \} \) for all \( k = 1, 2, \ldots, K \). Since the measure \( \mu_i \) is assigned to the \( k \)-th cluster, then the \( k \)-th geodesic curve connecting the global barycenter \( \bar{\mu}_0 \) with the \( k \)-th local barycenter \( \bar{\mu}_k \) may reveal more information regarding this assignment. This extra information is revealed by the point on the geodesic \( \gamma_k(t) \) that the measure \( \mu_i \) is closest, i.e. the projection of measure \( \mu_i \) to the geodesic curve \( \gamma_k(t) \) (projection in terms of the Wasserstein distance). The closest point location alongside this particular geodesic curve is represented by \( \tau_i \in [0, 1] \), called the registration parameter, and its value is obtained through the solution of the optimization problem (registration problem)

\[
\tau_i := \arg \min_{t \in [0, 1]} W_2^2(\mu_i, \gamma_k(t)).
\]

The registration of measure \( \mu_i \) to the geodesic curve \( \gamma_k \), can be considered as some sort of interpolation scheme, which allows us to approximate this measure by a measure considered as an interpolant between the global barycenter (i.e. mean of all observations) \( \bar{\mu}_0 \) and the barycenter of the particular cluster to which the particular observation (measure \( \mu_i \)) is assigned to by the K-means algorithm. The registration parameter \( \tau_i \) can be thought as a similarity index, indicating whether the derived approximation is closer to the general global barycenter \( \bar{\mu}_0 \) or the barycenter \( \bar{\mu}_k \) of the \( k \) cluster. In particular, \( \tau_i = 1 \) may indicate coincidence of \( \mu_i \) with the cluster center while \( \tau_i = 0 \) may indicate coincidence with the global barycenter. As a result of that, the value of the parameter \( \tau_i \) can be conceived as a quantification of the membership of the projection of \( \mu_i \) to the particular cluster \( k \) it is assigned to. The membership is stronger the closer \( \tau_i \) is to 1.

Figure 1: Cluster membership characterization of a point \( \mu_i \in \mathcal{P}(\mathbb{R}^d) \) using its projections to the local geodesic curve connecting the global barycenter \( (\mu_0 \in \mathcal{P}(\mathbb{R}^d)) \) with the local barycenters that \( \bar{\mu}_k \) deviates less.

However, the considerations above are not completely valid since the measure \( \mu_i \) will not coincide with any other measure unless their distance is zero. Although the value \( \tau_i \) itself may provide a first sense of similarity of the particular measure with the cluster center, without
taking into account also the distance between them, any conclusion could be proved misleading. Therefore, both \( \tau_i \) and \( \sigma_i := W_2(\mu_i, \mu_k) \) (distance of the measure \( \mu_i \) from the \( k \)-th local barycenter) should be taken into account in order to better characterize the membership of measure \( \mu_i \). Clearly, \( \sigma_i \) is the component that determines the cluster membership of \( \mu_i \) through the rule \( k(i) := \arg \min_{k=1,2,...,K} \sigma_i \), however the parameter \( \tau_i \) provides a first sense of similarity between \( \mu_i \) and \( \mu_k \) if the later do not coincide. Another characteristic that contains important information is the distance between \( \mu_i \) and the point’s projection on the geodesic curve, i.e. \( \tilde{\sigma}_i := W_2(\mu_i, \gamma_k(\tau_i)) \) (see Figure 1). Clearly, the parameters \( (\tau_i, \sigma_i, \tilde{\sigma}_i) \) contain all the important information regarding the goodness of fit of \( \mu_i \) to the \( k \)-th cluster and its contribution to the cluster’s variability. On Section 2.5 we construct a cluster compactness index based on the information provided by these parameters.

Note that \( (\sigma_i, \tilde{\sigma}, \tau_i) \) endow also with a fuzzy character the whole clustering procedure. For example, assigning \( \mu_i \) to a specific cluster for which the registration parameter \( \tau_i \) has value close to 0, can have the interpretation that the cluster center is closer to this specific observation but the cluster membership is extremely ambiguous since the two measures are not that similar. On the other hand, an assignment to a cluster accompanied by a registration parameter value close to 1 means that the cluster center is closer to this specific observation but the information provided by these parameters.

2.4 The proposed clustering procedure for the Location-Scale family of measures

In this section we discuss the special but very common case in practice where the probability measures involved are members of the Location-Scatter family. In the proposed clustering approach, this particular parametric family is of great interest since the afore mentioned results can be expressed in explicit or semi-explicit form. First, we summarize the results of [4] regarding the Wasserstein distance between two probability measures on \( \mathbb{R}^d \) belonging to the same Location-Scatter family as well as an efficient scheme for the calculation of the Wasserstein barycenter for a collection of probability measures from distributions of this type.

In general, a Location-Scatter family is a family of probability laws on \( \mathbb{R}^d \), induced by positive definite and affine transformations of the form \( AX + m \), where \( A \in \mathcal{M}_{++}^{d \times d} \) (\( \mathcal{M}_{++}^{d \times d} \) denotes the set of positive definite and symmetric matrices with \( d \) diagonal elements), \( m \in \mathbb{R}^d \) and \( X \) is a random vector in \( \mathbb{R}^d \) following the probability law \( \mu_0 \in \mathcal{P}_{2,ac}(\mathbb{R}^d) \) (the set of probability measures on \( \mathbb{R}^d \) with finite second-order moments and absolutely continuous with respect to the Lebesgue measure). This family is denoted by \( \mathcal{LS}(\mu_0) := \{ \mathcal{L}(AX + m), A \in \mathcal{M}_{++}^{d \times d}, m \in \mathbb{R}^d, \mu_0 = \mathcal{L}(X) \} \). An alternative parametrization of a location-scatter family can also be obtained in terms of the location vector \( m_0 \) and the dispersion matrix \( S_0 \) of the defining measure \( \mu_0 \), in terms of \( \mu_0^* := \mathcal{L}(S_0^{-1/2}(X_0 - m_0)) \). Location-Scatter families constitute a large class of probability laws in \( \mathbb{R}^d \) including important cases such as the Gaussian family. The quadratic Wasserstein distance of two probability measures of the same Location-Scatter family \( \mu_1, \mu_2 \), each characterized in terms of the location vectors and covariance-matrices \( (m_1, S_1), (m_2, S_2) \), is given explicitly in terms of the expression

\[
W_2^2(\mu_1, \mu_2) = \| m_1 - m_2 \|^2 + Tr \left(S_1 + S_2 - 2(S_1^{1/2}S_2S_1^{1/2})^{1/2} \right)
\]

Importantly, this exact result is also a lower bound for the Wasserstein distance of any pair of probability measures \( \mu_1, \mu_2 \), with locations and covariances \( (m_1, S_1), (m_2, S_2) \), not necessarily belonging to the same Location-Scatter family.

We also remind that given a collection of probability measures \( \{ \mu_i \}_{i=1,...,n} \) and a weight vector \( w = (w_1, \ldots, w_n) \in \Delta^n \) belonging to the unit simplex, the Wasserstein barycenter of the
collection for the given set of weights is defined as the minimizer of the problem:

$$\mu_B(w) := \arg \min_{\mu \in \mathcal{P}(\mathbb{R}^d)} \sum_{i=1}^n w_i W_2^2(\mu, \mu_i).$$

General conditions on the existence of the Wasserstein barycenter have been given by [3], who also studied the special case of Gaussian measures, while an alternative approach to this problem in terms of a fixed point characterization has been proposed in [4]. In both these references it has been shown independently that when the collection of measures \(\{\mu_i\}_{i=1}^n\) comes from the same Location-Scatter family, the measure \(\mu_B(w)\) (i.e. barycenter) is a measure from the same family with location and dispersion parameters determined by the pair \((m_B, S_B)\), where \(m_B\) can be explicitly expressed as

$$m_B = \sum_{i=1}^n w_i m_i,$$

and \(S_B\) is the unique solution to the matrix equation

$$S_B = \sum_{i=1}^n w_i \left( S_B^{1/2} S_i S_B^{1/2} \right)^{1/2}.$$

An iterative fixed point algorithm for the approximation of the solution of the above matrix equation has been proposed in [4] where the solution \(S_B\) is approximated through the iteration scheme

$$M_{k+1} = M_k^{-1/2} \left( \sum_{i=1}^n w_i (M_k^{1/2} S_i M_k^{1/2})^{1/2}\right)^2 M_k^{-1/2}, \quad n \geq 0,$$

for some initial condition \(M_0 \in \mathbb{R}^{d \times d}_+\).

Following the aforementioned explicit form results for the Wasserstein distance and Wasserstein barycenter in the case of Location-Scatter family, we can obtain also analytic expressions for the optimal transport maps and very convenient properties for the geodesic curves between such type of measures. Since the case of Gaussian measures is applied widely in multivariate statistics we focus on this case. Consider the Gaussian measures \(\mu_0, \mu_1 \in \mathcal{P}(\mathbb{R}^d)\) with location parameters \(m_0, m_1 \in \mathbb{R}^d\) and dispersion matrices \(S_0, S_1 \in \mathbb{R}^{d \times d}_+\) respectively. Due to the regularity of these measures, there exists a unique optimal transport map \(T\) that pushes forward measure \(\mu_0\) to \(\mu_1\), i.e. \(T_{\#} \mu_0 = \mu_1\). In the case of Gaussian measures of the same location (consider \(m_0 = m_1 = 0\)), this mapping is the symmetric and positive definite matrix \(\Lambda = S_1^{1/2} (S_1^{1/2} S_0 S_1^{1/2})^{-1/2} S_1^{1/2}\) (please see [27]). Therefore if a random variable \(X\) is distributed according to the law \(N(0, S_0)\), then its transformation \(Y := \Lambda X\) is distributed according to the law \(N(0, S_1)\). In this case the McCann’s interpolant is known explicitly and in particular, the geodesic curve connecting \(\mu_0\) and \(\mu_1\) takes the form \(\gamma(t) = \{[(1-t)I + t\Lambda]_{\#} \mu_0\}_{t \in [0,1]}\) with \(\Lambda\) as defined above. Moreover, [27] showed that each point on this geodesic curve is also a Gaussian measure and in fact, for any fixed \(t \in [0,1]\), the probability measure \(\gamma(t)\) is characterized by the location parameter \(m(t) := (1-t)m_0 + tm_1\) and the dispersion matrix \(S(t) := ((1-t)I + t\Lambda) S_0 ((1-t)I + t\Lambda)\).

Following the above explicit results, the proposed clustering procedure in the case of Gaussian measures can be done in semi-closed form. The related algorithmic procedure is described in Algorithm 2.

**Remark 2.5 (Matrix-valued data).** The above approach can be used also for the case of matrix-valued data. In particular, since any symmetric and positive definite matrix (e.g. a
Algorithm 2: Wasserstein K-Means (Location-Scale measures)

1. Choose $K$ measures out of the original observations as the initial cluster centers, $\bar{\mu}_k$, $k = 1, 2, \ldots, K$.
2. Assign each of the measures $\mu_i, i = 1, 2, \ldots, n$, to one of the $K$ clusters, selecting the cluster membership $k(i)$ according to the rule
   
   $k(i) = \arg \min_{k \in \{1, 2, \ldots, K\}} \left\{ \| m_i - \bar{m}_k \|^2 + \text{Tr} \left( S_i + \bar{S}_k - 2(S_i^{1/2}\bar{S}_k S_i^{1/2})^{1/2} \right) \right\}$.

3. Update the cluster centers by estimating the local barycenters using the new members of each cluster by applying the formulas (8)-(9).
4. Repeat steps 2-3 until the cluster centers do not change significantly.

covariance matrix) can be identified as a Gaussian measure with zero location and dispersion the matrix itself, then the Wasserstein metric can provide a useful metrization of the space of symmetric and positive-definite matrices (Bures-Wasserstein distance) [26]. Therefore, the described procedure can be used as it is (with zero locations fixed) for the separation of matrix-valued data working on the isomorphic space of Gaussian measures.

Remark 2.6 (More general probability measures). The same approach can also be used for the clustering of more general probability measures not necessarily belonging to the Location-Scatter family. However, in that case several computational issues arise. The first and most important is that there does not exist an explicit formula for the calculation of the Wasserstein distance between any two probability measures and therefore one has to solve the minimization problem stated in [2]. This problem is quite challenging computationally but can be addressed with certain regularization schemes in order to improve the convergence speed of the numerical schemes. A typical scheme of this type employs Bregman-type regularizations (see e.g. [5, 10, 13]). The second problem is the estimation of the Wasserstein barycenters where also several approaches have been considered in the literature, through discretization of the problem [11], through slicing of the cost functional [8], through modeling the measures as parametric deformations of the barycenter [7], and others. However, this remains a very active field of research and the investigation of such cases is beyond the scope of this work. However, it is a very promising direction for future work.

2.5 A similarity index and clustering criteria relying on the geodesic curves

In order to offer an alternative and more appropriate proposal to the existing clustering criteria, we construct an index which exploits the geodesic information revealed during the clustering procedure. From the registration problem (6) we are looking to employ the geometric information extracted by this task in order to provide a valid compactness (homogeneity) assessment of the estimated clusters by the Wasserstein K-Means algorithm discussed in 2.2. In particular, consider that $K$ clusters have been derived by Algorithm 1 (or 2) where each one contains $n_k$ probability measures where $n = \sum_{k=1}^{K} n_k$. For each measure which has been assigned to the $k$-th cluster (let us denote them as $\mu_{k,i}$ for $k = 1, 2, \ldots, K$ and $i = 1, 2, \ldots, n_k$), from the related registration procedure, there have been computed the pairs $(\tau_{k,i}, \bar{\sigma}_{k,i})$. In order to construct a reasonable similarity index we have first to condense the information provided by these pairs to one scalar for each measure. Below we describe the approach we follow.

Assume that $n_k$ probability measures have been assigned to the $k$-th cluster with cluster center the probability measure $\bar{\mu}_k$ and the parameter triples $\{(\tau_i, \sigma_i, \bar{\sigma}_i)\}$ have been computed for each $i = 1, 2, \ldots, n_k$. First, we need to characterize the element (measure) which is closest to the cluster’s center, i.e. we define $i_\ast := \arg \min_{i=1,2,\ldots,n_k} W_2(\mu_i, \bar{\mu}_k)$ and set $\mu_\ast = \mu_{i_\ast}$, $\sigma_\ast := \sigma_{i_\ast}$.
and $\tilde{\sigma}_k := \tilde{\sigma}_k$. In this manner, $\mu_*$ is set as the minimal element among the assigned measures in the $k$-th cluster in the sense that deviates less from the cluster’s estimated mean behaviour $\bar{\mu}_k$. Therefore, this element could be used in order to derive some comparison standards in order to better understand the characteristics of the other measures assigned to this cluster.

The approach we follow in this attempt is through the comparison and estimation of less distant points (measures) between two geodesic curves. According to that we use the projection of each measure $\mu_i$, i.e. $\gamma_k(\tau_i)$, and we calculate its projection to the geodesic curve that connects the minimal element $\mu_*$ and the original measure $\mu_i$. Through this task we cross-validate the result we obtained by the initial registration task described in 2.3.

![Figure 2: Representation of the geodesic curves involved in the double registration problem.](image)

In particular, for each $\mu_i$ assigned in cluster $k$ we define the geodesic curve $\gamma_{i,*}(s)$ connecting the measures $\mu_*$ and $\mu_i$ with $\gamma_{i,*}(0) = \mu_i$ and $\gamma_{i,*}(1) = \mu_*$. Then, we solve the reverse-registration problem

$$\min_{s \in [0,1]} W_2(\gamma_{i,*}(s), \gamma_k(\tau_i))$$

which provides the point on the geodesic curve $\gamma_{i,*}(\cdot)$ that deviates less from the point $\gamma_k(\tau_i)$ and parameterized by $s_i \in [0,1]$ which is considered as the minimizer of problem 10 for $i$. The rationale between this second projection step is that if the initial measure $\mu_i$ is quite similar with the cluster center $\bar{\mu}_k$ then at the second projection we should get a point close to $\mu_*$ i.e. $s_i \to 1$. That means that the double registration procedure will reveal if the measure $\mu_i$ really belongs to the interior of the $k$-th cluster. Observe that if $\mu_i = \mu_*$ then $\gamma_{i,*}(\cdot)$ is simplified to $\mu_*$ and the above registration problem is simplified to the reverse projection of $\gamma_k(\tau_*)$ to the initial measure $\mu_*$. For $\mu_i \neq \mu_*$ the interpretation of the minimizers is of particular interest. Let us examine the extreme cases. If $s_i = 0$ and $\tau_i = 0$ then measure $\mu_i$ can be characterized as an outlier (observation on the boundary of the cluster). The case where $s_i = 0$ and $\tau_i = 1$ cannot happen since if the initial projection is exactly on $\bar{\mu}_k$ then the solution to the reverse registration problem has only one solution on $s_i = 1$, i.e. $\mu_*$. The case $s_i = 1$ given that $\tau_i = 0$ does not belong to the feasible area of the problem unless if $\bar{\mu}_k = \bar{\mu}_0$. The case $s_i = 1$ given that $\tau_i = 1$ is probably the more “difficult” case since either the measure $\mu_i$ coincides with the minimal element or the measure $\mu_i$ belongs to the internal area of the cluster or $\mu_i$ is an outlier which characteristics deviates more from $\bar{\mu}_0$ than $\bar{\mu}_k$. It is evident that the information obtained from the minimizing pairs $(\tau_i, s_i)$ of the double registration problem (optimization problems 6 and 10) is not enough for the full characterization of the status of each point assigned in $k$-th cluster. Therefore we need to further enhance our distinction argument using the information provided by the distances $\sigma_i$ and $\tilde{\sigma}_k$ in order to provide a valid measure of homogeneity/compactness of the clusters. The above discussion justifies the definition of the following similarity index.
For each $\mu_i$ the geodesic similarity index of the $k$-th cluster center is defined as

$$\tilde{\tau}_{k,i} := \frac{s_i \tau_i \tilde{\sigma}_s}{\tau_s \tilde{\sigma}_i}$$

where $(s_i, \tau_i)$ denotes the solution to the double registration problem \cite{6,10} for $\mu_i$. This index is a rescaled version of the initial $\tau_i$, however taking into account also the effect of the true distance from the geodesic curve normalized with respect to the closest point to the cluster center. Clearly, for $\mu_i = \mu_s$ we obtain an index value equal to 1 since it is considered as the ideal case. In particular, due to its convenient scaling, each registration parameter $\tilde{\tau}_{k,i}$ can be realized as the level of certainty/ambiguity regarding the true membership of a particular observation ($\mu_{k,i}$) to the $k$-th cluster. Clearly, if $\tilde{\tau}_{k,i} \to 1$ there is a great amount of certainty that the $k$-th cluster is a very appropriate match for this observation, while if $\tilde{\tau}_{k,i} \to 0$ there is a great level of ambiguity that the $k$-th cluster is the most suitable choice for it. Note that, for the clusters that have been derived up to the time the registration parameters are computed, there is no doubt that the assigned observations to these clusters are rightly assigned there. However, small $\tilde{\tau}$ values by many observations indicate a cluster of low homogeneity (non-compact) which maybe reveals a poor initial decision for the number of clusters.

Given that $K$ different clusters have been determined and taking into account all the measures assigned to each cluster, we define the Geodesic Compactness Index (GCI) for each one of the clusters in the following manner

$$(12) \quad GCI_k(K) := \frac{1}{n_k} \sum_{i=1}^{n_k} \tilde{\tau}_{k,i} \text{ for all } k = 1, 2, \ldots, K.$$ 

Moreover, a total compactness index for the total clustering procedure is naturally defined through the weighted sum

$$(13) \quad GCI(K) := \sum_{k=1}^{K} \frac{n_k}{n} GCI_k(K) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \tilde{\tau}_{k,i}.$$ 

The interpretation of the above indices is in complete accordance with the mean tendency of the similarity indices. A value of the cluster index close to 1 is interpreted as an artifact of high homogeneity of the specific cluster while a an index value close to 0 is interpreted as very low homogeneity of the cluster. Observing low values to the majority of clusters determined indicates that the choice have been made for the number of clusters is maybe not appropriate. On the other hand, observing high values of the compactness indices for the majority of the determined clusters, indicates an appropriate choice for the number of clusters and success in the determination of compact clusters. However, note that if $K = n$ we get that $GCI_1(n) = GCI_2(n) = \ldots = GCI_n(n) = 1$ which is clearly not an indication of a good choice for the number clusters but rather a limit case.

3 Clustering EU Bond Yields

As an illustrative application we consider the task of clustering the countries of eurozone according to their credit profiles for the time period 2001-2019. The data under consideration is the weekly yield rates for each country’s zero coupon government bonds with maturity horizons from one to ten years. The data are collected, on a monthly basis, through the official DataStream database and the countries for which the clustering task is performed are those which were members of the European monetary union constantly for the time period 2001 to 2019. Note that data for Luxemburg are not available (possibly not such type of bonds have been published from its government) therefore it is excluded from the analysis. As a result,
the countries included in the analysis (followed by their abbreviations) are: Austria (AT), Belgium (BL), Finland (FI), France (FR), Germany (DE), Greece (GR), Ireland (IE), Italy (IT), Netherlands (NL), Portugal (PT) and Spain (ES).

3.1 Credit Profile Determination through Probability Measures

The available data used for the clustering task are for each particular country included in the study, the observed bond indices for each different maturity horizon as it is recorded at the end of each month. For example, for a two year period, for the country $i$ there are available 104 records for the published by the country’s government bonds with maturity horizon $m = 1, 2, ..., 10$ years. Therefore, a record at the specified time $t$ (i.e. at the end of a certain week in the time period under consideration) is the vector

$$r_t^{(i)} = \left(r_{t,1}^{(i)}, r_{t,2}^{(i)}, ..., r_{t,10}^{(i)}\right)^T \in \mathbb{R}^{10}$$

where each $r_{t,m}^{(i)}$ for $m = 1, 2, ..., 10$ denotes the recorded index for the bond with maturity horizon $m$ years. As a result, if the period under consideration consists of $n$ weeks, then for each country there is available a $(n \times 10)$ data matrix with the available records. Note that although these data could be also considered as time series data, we do not employ this approach here since it is irrelevant and outside the the scope of the present work. However, we indicatively refer to the approaches in the literature that consider the matter of clustering time series data [20, 29, 36]. In particular, in this context we realise these records as realizations of a 10-dimensional random value which can be sufficiently represented by a Location-Scatter probability measure with a certain location and dispersion matrix.

Given that the economies are in a steady state, no major variations in these records should be observed in each month. For this reason, the entire time period 2001-19, that data are available, is split into six smaller periods: 2001-04, 2005-08, 2009-11, 2012-14, 2015-17 and 2018-19 in order to not allow very lengthy time periods where important changes may occur and significantly affect each country’s economy. These periods represent a growth period (2001-04), followed by two financial crisis intervals: 2005-08, which is the period up to the Lehman brothers collapse, and 2009-11 that is the immediate aftermath, including the ensuing debt crisis in Greece, Portugal and Ireland. These are followed by a period when certain eurozone countries were under Economic Adjustment programmes (bailout programmes, 2012-14), a period following these programmes (2015-17) and the last two recovery years (2018-19). It is evident that one may try different time intervals (6 months, 1 year periods, etc.) or use more frequent data (e.g. daily, hourly, etc.) however since this is not the main purpose of this paper we attempt to keep such details to the most convenient for the analysis and the reader framework.

Clearly, the provided data in the current form is quite hard to analyze. Therefore, we attempt for each one of the defined time periods, to condense the available data for each country in order to sufficiently and more conveniently represent the provided information. That task is performed through the identification of each data matrix by a probability measure of the Location-Scatter family of distributions, for convenience we restrict ourselves to the family of Gaussian measures. Under this approach, for a certain data matrix (i.e. provided data for a specific country $i$) the location vector $m_i \in \mathbb{R}^{10}$ represents the mean tendency of the bond indices for the particular period of time therefore it is reasonable to set

$$m_i = \bar{r}_i = (\bar{r}_{i,1}, \bar{r}_{i,2}, ..., \bar{r}_{i,10})^T \in \mathbb{R}^{10}$$

where $\bar{r}_{i,m} = \mathbb{E}[r_{t,m}^{(i)}] \in \mathbb{R}$ for $m = 1, 2, ..., 10$ denoting the mean tendency of the bond index of maturity horizon $m$ for the country $i$. Another important characteristic of the dataset is the correlation structure between the bonds of different maturity and of course their dispersion. Under the assumption that the nature of the data does not change significantly during our
sample period, then the dispersion characteristics can be sufficiently estimated by the covariance matrix

\[ S_i = \mathbb{E}[(r_t^{(i)} - \bar{r}_i)(r_t^{(i)} - \bar{r}_i)'] \in \mathbb{R}_+^{10 \times 10}. \]

Under this setting, the country’s \( i \) profile can be identified with the Gaussian probability measure \( \mu_i = N(m_i, S_i) \) by condensing the information provided by the data set to location and dispersion-correlation characteristics. This is a very natural approach in most applications considering multivariate data however, one may consider also some family of probability distributions capturing asymmetries in data (e.g. through a Copula, a meta-distribution, a vine-Copula construction, etc.). However, since in this work we investigate the case of the Location-Scatter family of distributions we feel that such a consideration, although extremely interesting, it is beyond the scope of this paper. Therefore, under the above considerations, the credit profile of a country induced by the observed bond yield data can be represented by a Gaussian measure where the location and dispersion parameters efficiently condense the available information provided for the time period under consideration and the nature of the Gaussian law reflects the assumption regarding consistency, i.e. the time period length is chosen so that no dramatic changes happen in the economies of the countries under study for its duration.

![Figure 3: Geodesic Compactness Index values per time period](image)

### 3.2 Clustering bond yield data with Wasserstein barycenters

Following the aforementioned framework and the learning methodology described in Section 2.2, we perform the clustering analysis of the credit profiles of the eurozone countries. The clustering task is performed repeatedly for the defined six time periods 2001-04, 2004-08, 2009-11, 2012-14, 2015-17 and 2018-19. First, we employ the geodesic criterion presented in Section 2.5 in order to determine a reasonable number of clusters for each time period. The results of the criterion output are graphically represented in Figure 3. We consider up to six different clusters for each time period, since more clusters may not have any interpretative value. The rationale beyond the selection of an appropriate number of clusters in each case is to choose the number of groups where a high value (comparing to the values for other group numbers) in the compactness is succeeded and adding more groups this number is not significantly increased. Having this in mind, the optimal cluster numbers per time period are: 4 for 2001-04, 3 for 2005-08, 4 for 2009-11, 4 for 2012-14, 4 for 2015-17 and 3 for 2018-19.

It seems that for the period 2001-2008 for most choices of clusters’ number, the compactness index is not higher than 0.65 (higher index value are obtained for the period 2001-04 for \( K > 4 \)
Table 1: Optimal clusters according to GCI interpretation

| Time Period | 2001-04 | 2005-08 | 2009-11 | 2012-14 | 2015-17 | 2018-19 |
|-------------|---------|---------|---------|---------|---------|---------|
| Group 1     | DE      | AT BL DE FI FR NL | AT BL DE FI FR NL | AT BL DE FI FR NL | AT BL DE FI FR IE NL | AT BL DE FI FR IE NL |
| Group 2     | AT      | GR IT   | ES IT   | ES IT   | ES IT   | ES IT   |
| Group 3     | BL ES FI FR NL | IE | IE PT   | PT      | PT      | GR IT   |
| Group 4     | GR IE IT PT | GR | GR      | GR      | GR      | GR      |

Table 2: GCI values per time period and cluster

| Year | AT | BL | DE | ES | FI | FR | GR | IE | IT | NL | PT |
|------|----|----|----|----|----|----|----|----|----|----|----|
| 2001 | AAA | AA+ | AAA | AA+ | AA+ | AAA | A | AAA | AA | AAA | AA |
| 2002 | AAA | AA+ | AAA | AA+ | AAA | AAA | A | AAA | AA | AAA | AA |
| 2003 | AAA | AA+ | AAA | AA+ | AAA | AAA | A+ | AAA | AA | AAA | AA |
| 2004 | AAA | AA+ | AAA | AAA | AAA | A | AAA | AA- | AA | AAA | AA |
| 2005 | AAA | AA+ | AAA | AAA | AAA | A | AAA | AA- | AA | AAA | AA |
| 2006 | AAA | AA+ | AAA | AAA | AAA | A | AAA | A+ | AA | AAA | AA- |
| 2007 | AAA | AA+ | AAA | AAA | AAA | A | AAA | A+ | AA | AAA | AA |
| 2008 | AAA | AA+ | AAA | AAA | AAA | A | AAA | A+ | AA | AAA | AA- |
| 2009 | AAA | AA+ | AAA | AAA | AAA | BBB+ | AA | A+ | AA | AAA | A+ |
| 2010 | AAA | AA+ | AAA | AAA | BBB+ | A | A+ | AAA | A- | AA |
| 2011 | AAA | AA+ | AAA | AAA | BBB- | AAA | A | AAA | AA | AAA | BBB |
| 2012 | AA+ | AA | AAA | BBB- | AAA | A+ | B | BBB | BBB+ | AAA | BB |
| 2013 | AA+ | AA | AAA | BBB- | AAA | B | BBB | BBB | AAA | BB |
| 2014 | AA+ | AA | AAA | BBB+ | AA+ | A | B | AA | BBB- | AAA | BB |
| 2015 | AA+ | AA | AAA | BBB+ | AA+ | AA | CCC+ | A+ | BBB- | AAA | BB |
| 2016 | AA+ | AA | AAA | BBB+ | AA+ | A | BB | A | BBB | AAA | BB |
| 2017 | AA+ | AA | AAA | BBB+ | AA+ | AA | B | A | BBB | AAA | BBB |
| 2018 | AA+ | AA | AAA | A+ | AA+ | AA | B+ | A | BBB | AAA | BBB |
| 2019 | AA+ | AA | AAA | A | AA+ | AA | BB- | A+ | BBB | AAA | BBB |

Table 3: S&P yearly credit ratings for the EU countries for the time period 2001-2019

From early on (2001-04), our proposed framework and methodology appears able to identify a cluster of the so-called PIIG countries (Portugal, Ireland, Italy and Greece), whose bond yields indicated back then that a debt crisis might ensue. During the eurozone debt crisis (2009-11) and the resulting bailout programmes (2012-14) Greece, the country with the most acute debt problem, is singled out as a different cluster, while Ireland and Portugal (countries under bailout programs) are grouped into a separate cluster. After the end of the bailout
programs in all countries (2018-19), Ireland has “moved” to the core eurozone Group 1, Greece and Italy remain the countries with the highest debt–to–GDP in the eurozone (Group 3), while the Iberia countries (Spain and Portugal) constitute a separate cluster in–between. In order to assess the validity of the obtained clustering outputs, in Figure 3 are illustrated the credit rankings of each eurozone country for the period 2001-19 on a yearly base as determined by Standard and Poors (S&P). It is evident, that the clustering approach we followed succeeded in retrieving the true situation in most of the cases if we consider the rating mechanism of S&P as an expert opinion.

4 Conclusions

In this work we presented a clustering approach for measure-valued data relying on the concept of Wasserstein barycenter. In particular, we focus on the case of the Location-Scatter family of probability measures where a clustering procedure based on the K-means algorithm is proposed, accompanied by a criterion for the optimal selection of the number of clusters relying on geometric arguments of the underlying space of probability measures that we work. The proposed approach can be used as an unsupervised learning scheme with substantial success, and this fact is illustrated on the application in clustering credit profiles of EU countries considered in this paper. The proposed clustering approach seems reasonable when data under consideration evolve in time like bond data considered in the application in the paper where the proposed clustering approach seems that successfully reveals the true situation.

Acknowledgements

Georgios N. Domazakis kindly thanks Erwin Schrödinger International Institute for Mathematics and Physics for financial support and hospitality in Vienna, Austria, during the thematic programme on “Optimal Transport” in May 2019.

References

Aggarwal, C. C. (2010). Managing and mining uncertain data (Vol. 35). Springer Science & Business Media.

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In International conference on database theory (pp. 420–434).

Agueh, M., & Carlier, G. (2011). Barycenters in the Wasserstein space. SIAM Journal on Mathematical Analysis, 43(2), 904–924.

Alvarez-Esteban, P. C., del Barrio, E., Cuesta-Albertos, J., & Matrán, C. (2016). A fixed-point approach to barycenters in Wasserstein space. Journal of Mathematical Analysis and Applications, 441(2), 744–762.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., & Peyré, G. (2015). Iterative bregman projections for regularized transportation problems. SIAM Journal on Scientific Computing, 37(2), A1111–A1138.

Bhatia, R., Jain, T., & Lim, Y. (2019). On the Bures–Wasserstein distance between positive definite matrices. Expositiones Mathematicae, 37(2), 165–191.

Boissard, E., Le Gouic, T., Loubes, J.-M., et al. (2015). Distributions template estimate with wasserstein metrics. Bernoulli, 21(2), 740–759.
Bonneel, N., Rabin, J., Peyré, G., & Pfister, H. (2015). Sliced and radon wasserstein barycenters of measures. Journal of Mathematical Imaging and Vision, 51(1), 22–45.

Bouveyron, C., Girard, S., & Schmid, C. (2007). High-dimensional data clustering. Computational Statistics & Data Analysis, 52(1), 502–519.

Carlier, G., Duval, V., Peyré, G., & Schnitzer, B. (2017). Convergence of entropic schemes for optimal transport and gradient flows. SIAM Journal on Mathematical Analysis, 49(2), 1385–1418.

Carlier, G., Oberman, A., & Oudet, E. (2015). Numerical methods for matching for teams and wasserstein barycenters. ESAIM: Mathematical Modelling and Numerical Analysis, 49(6), 1621–1642.

Chiou, J.-M., & Li, P.-L. (2007). Functional clustering and identifying substructures of longitudinal data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(4), 679–699.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In Advances in neural information processing systems (pp. 2292–2300).

De Amorim, R. C., & Mirkin, B. (2012). Minkowski metric, feature weighting and anomalous cluster initializing in K-means clustering. Pattern Recognition, 45(3), 1061–1075.

de Carvalho, F. d. A., Brito, P., & Bock, H.-H. (2006). Dynamic clustering for interval data based on $l_2$ distance. Computational Statistics, 21(2), 231–250.

De Carvalho, F. d. A., & Lechevallier, Y. (2009). Partitional clustering algorithms for symbolic interval data based on single adaptive distances. Pattern Recognition, 42(7), 1223–1236.

Goutte, C., Toft, P., Rostrup, E., Nielsen, F. Å., & Hansen, L. K. (1999). On clustering fMRI time series. NeuroImage, 9(3), 298–310.

Haggarty, R., Miller, C., & Scott, E. (2015). Spatially weighted functional clustering of river network data. Journal of the Royal Statistical Society: Series C (Applied Statistics), 64(3), 491–506.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Huang, X., Ye, Y., Xiong, L., Lau, R. Y., Jiang, N., & Wang, S. (2016). Time series k-means: A new k-means type smooth subspace clustering for time series data. Information Sciences, 367, 1–13.

Irpino, A., & Verde, R. (2008). Dynamic clustering of interval data using a Wasserstein-based distance. Pattern Recognition Letters, 29(11), 1648–1658.

Jacques, J., & Preda, C. (2014). Model-based clustering for multivariate functional data. Computational Statistics & Data Analysis, 71, 92–106.

Kim, Y.-H., & Pass, B. (2017). Wasserstein barycenters over Riemannian manifolds. Advances in Mathematics, 307, 640–683.

Le Gouic, T., & Loubes, J.-M. (2017). Existence and consistency of Wasserstein barycenters. Probability Theory and Related Fields, 168(3-4), 901–917.

Liao, T. W. (2005). Clustering of time series data - A survey. Pattern recognition, 38(11), 1857–1874.
Masarotto, V., Panaretos, V. M., & Zemel, Y. (2019). Procrustes metrics on covariance operators and optimal transportation of Gaussian processes. *Sankhya A, 81*(1), 172–213.

McCann, R. J. (1997). A convexity principle for interacting gases. *Advances in mathematics, 128*(1), 153–179.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l’Académie Royale des Sciences de Paris*.

Niennattrakul, V., & Ratanamahatana, C. A. (2007). On clustering multimedia time series data using k-means and dynamic time warping. In *2007 international conference on multimedia and ubiquitous engineering (mue’07)* (pp. 733–738).

Papayiannis, G. I., & Yannacopoulos, A. N. (2018). A learning algorithm for source aggregation. *Mathematical Methods in the Applied Sciences, 41*(3), 1033–1039.

Ray, S., & Mallick, B. (2006). Functional clustering by Bayesian wavelet methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68*(2), 305–332.

Santambrogio, F. (2015). Optimal transport for applied mathematicians. *Birkhäuser, NY, 55*, 58–63.

Tuzel, O., Porikli, F. M., & Meer, P. (2007). Human detection via classification on Riemannian manifolds. *2007 IEEE Conference on Computer Vision and Pattern Recognition, 1-8*.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Villani, C. (2003). *Topics in optimal transportation* (No. 58). American Mathematical Soc.

Vlachos, M., Lin, J., Keogh, E., & Gunopulos, D. (2003). A wavelet-based anytime algorithm for k-means clustering of time series. In *In proc. workshop on clustering high dimensionality data and its applications*.

Zemel, Y., Panaretos, V. M., et al. (2019). Fréchet means and Procrustes analysis in Wasserstein space. *Bernoulli, 25*(2), 932–976.