

Functional classification on convex function spaces

Classificazione di funzioni in spazi funzionali convessi

Angela De Sanctis, Francesca Fortuna

Abstract The problem of clustering functional data is considered. In particular we refer to cases in which the functional form of the observations is known in advance. In this setting, contrary to the classical functional approach, the approximation of the function underlying the data is not required for functional clustering methods. However, clustering functional data is a difficult task because the function space is, generally, of infinite dimension. Thus, the distance among functions may have infinity solutions. For this reason, we restrict the space of the functions to a closed and convex subset in an Hilbert space. In this proposed setting an L^2 metric is applied combined clustering algorithms for finite dimensional data.

Abstract Il lavoro affronta il problema della classificazione di dati funzionali. In particolare si fa riferimento ai casi in cui la forma funzionale delle osservazioni é nota a priori. Contrariamente all'approccio funzionale classico, in questi casi, non é necessario approssimare la funzione sottostante i dati per procedere alla loro classificazione. Tuttavia, classificare i dati funzionali é un compito difficile in quanto lo spazio funzionale é in genere infinito; di conseguenza, una metrica ad esso applicata potrebbe avere infinite soluzioni. Per questo motivo, restringiamo lo spazio delle funzioni ad un sottospazio chiuso e convesso di Hilbert in modo da applicare una metrica L^2 insieme ad algoritmi di classificazione tipici degli spazi finiti.

Key words: Functional classification, L^2 distance, convex function space

Angela De Sanctis
University G. D' Annunzio, Chieti-Pescara, Italy e-mail: a.desanctis@unich.it

Francesca Fortuna
University G. D' Annunzio
Chieti-Pescara, Italy e-mail: francesca.fortuna@unich.it

1 Introduction

Functional data analysis (FDA) (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006) addresses problems in which the observations are described by functions rather than finite dimensional vectors.

The curves in a functional data set often present a variety of distinctive patterns corresponding to different shapes and variation (Chiou and Li, 2007) that can be identified by clustering the functions (Tarpey, 2007; Abraham et al., 2003; James and Sugar, 2003).

However, clustering functional data is generally a difficult task because of the infinite dimensional space data belong (Jacques and Preda, 2004). For this reason, most approaches are based on dimension reduction before clustering. In particular, the infinite dimension problem has been reduced to a finite one by approximating data with elements from some finite dimensional space, such as coefficients of functional data expansion (Abraham et al., 2003) or a given number of principal component scores (Peng and Muller, 2008). Then, classical heuristic clustering algorithms can be performed. It is common to refer to this procedure with the term two-stage approach.

Alternatively, nonparametric methods can be applied for clustering functional data. They consist generally in defining specific distances or dissimilarities among curves and then apply clustering algorithms for finite dimensional data (Ferraty and Vieu, 2006; Ieva et al., 2013).

Finally, model based clustering methods can be performed assuming a probabilistic distribution on some finite dimensional coefficients describing data (Ray and Mallick, 2006; Sam et al., 2011).

The disadvantage of the above methods is that clustering results can differ depending on how the curves are fit to the data. For example, nonparametric methods can be assimilated to raw-data clustering (Bouveryon and Brunet, 2012) or to a two-stage method, depending on whether the distance is approximated using directly the discrete observations of curves or using an approximation of the curves into a finite basis. Moreover, estimating curves using different sets of basis functions corresponds to different linear combinations of the data and many clustering algorithms, such as K-means, are not invariant to linear transformations of the data (Tarpey, 2007).

2 Functional classification on convex function spaces

This paper focuses on a particular aspect of FDA, called parametric FDA (De Sanctis and Di Battista, 2012; Di Battista and Fortuna, 2013), that is, when the functional form underlie the observations is known in advance. For this reason, the approximation by means of basis functions is not required. In these cases, functions belong to a parametric family of functions, say S , which is a subset of $L^p(\mu)$ where μ is a positive measure of an arbitrary measure space X .

Starting from n parametric functional data, $f_1(x), f_2(x), \dots, f_n(x)$, we aim to identify a set of homogeneous clusters in $L^p(\mu)$ by determining a partition of the space according to minimal distance. Since the functional observations belong to an infinite dimensional space, the equivalence between norms and distances, typical of finite dimensional Euclidean spaces, fails. In this context, the choice of the preliminary norm becomes crucial (Ferraty and Vieu, 2006).

Several proximity measures can be used. A suitable measure of distance between two functions, $f_i(x)$ and $f_j(x)$ is the L^p distance:

$$d(f_i(x), f_j(x)) = \|f_i(x) - f_j(x)\|_{L^p} = \left(\int_X |f_i(x) - f_j(x)|^p d\mu(x) \right)^{\frac{1}{p}} \quad (1)$$

for which different p can be chosen.

However, Ferraty and Vieu (2006) have highlighted that semi metric spaces are more flexible with respect to normed or metric ones. In this case, the distance between two functions can be expressed as follows:

$$d_l(f_i(x), f_j(x)) = \left(\int_X (f_i^{(l)}(x) - f_j^{(l)}(x))^2 dx \right)^{\frac{1}{2}} \quad (2)$$

where $f^{(l)}(x)$ is the l -th derivative of $f(x)$.

We recall that a semi-metric can be defined to be a metric but such that $d(f_i(x), f_j(x)) = 0 \nRightarrow f_i(x) = f_j(x)$.

Indeed equation (2) represents a semi-metric for $l > 0$; in fact, by considering $f^{(0)}(x) = f(x)$ the L^2 metric is obtained.

However, it is not always appropriate to use a semi-metric because it emphasizes features of the derivatives rather than of the curves themselves.

The main problem caused by the infinite dimensionality of the spaces, is that there may be infinitely many solutions for the distance in (1). For this reason, we restrict the space of the functions to closed and convex subsets, S , in Hilbert spaces (De Sanctis et al., 2014). We recall that $L^2(\mu)$ is the space of square integrable measurable functions on X whose norm is induced by the inner product:

$$(f_i, f_j) = \int_X f_i(x) \overline{f_j(x)} d\mu(x) \quad (3)$$

whenever $f_i(x), f_j(x) \in L^2(\mu)$. An essential property of Hilbert space is that the distance of a point to a closed set is always attained. In fact, if S is a closed convex set in a Hilbert space, H , and $h \in H$, then, there exists a unique point $s \in S$ that minimize the distance between h and a point in S (Rudin, 1986):

$$d(h, S) = \|h - s\| = \min\{\|h - s\| : s \in S\} \quad (4)$$

Whereas, in a metric space the distance from a point t to a set S is:

$$d(t, S) = \inf\{\|t - s\| : s \in S\} \quad (5)$$

In the case in which the function space S is nonconvex, we can resort to a biunivocal transformation, T , between S and a convex subset S' in $L^p(X)$ (De Sanctis et al., 2014). In particular, let $f_1(x), f_2(x), \dots, f_n(x)$ be functional observation belong to a nonconvex functional space S and $g_1(x), g_2(x), \dots, g_n(x)$ the corresponding functions in a convex subset S' , defined by T . Let $\bar{g}(x)$ a functional cluster mean defined as the mean of the corresponding function, then the L^2 distance between the functions, $g_i(x)$, and the functional cluster mean is:

$$d_2(g_i(x), \bar{g}(x)) = \left(\int_X (g_i(x) - \bar{g}(x))^2 dx \right)^{\frac{1}{2}} \quad (6)$$

The main advantage of the parametric FDA approach is that it is possible to compute the functional distance directly on the explicit form of the functions when the function space is convex in $L^2(X)$. Otherwise, we refer to the corresponding functions, defined by the transformation T and then return to the original space through the inverse transformation, T^{-1} . For this reason, clustering results not depend on how the curves are fit to the data, contrary to the classical FDA approach.

2.1 Functional k -means

In this proposed setting an L^2 metric in function space is applied combined with a k -means algorithm for finite dimensional data (MacQueen, 1967). The latter is an iterative procedure that is initialized by fixing the number k of clusters and by selecting a set of k arbitrary and distinct initial centroids in S , $\{\phi_1^{(0)}(x), \dots, \phi_k^{(0)}(x)\}$. At the m -th algorithm iteration, $m > 0$, each function is assigned to the cluster whose centroid, at the $(m-1)$ -th iteration, is the nearest according to the chosen distance:

$$C_i^{(m)} = \arg \min_{q=1,2,\dots,k} \left(\int_X |f_i(x) - \phi_q^{m-1}(x)|^2 dx \right)^{\frac{1}{2}} \quad (7)$$

where $C_i^{(m)}$ is the m -th cluster assignment of the i -th function, $i = 1, 2, \dots, n$. Once all of the functions have been assigned to a cluster, the cluster means are updated as:

$$\phi_q^{m+1}(x) = \sum_{f_i \in c_q} \frac{f_i(x)}{n_q} \quad (8)$$

where n_q is the number of functions in the q -th cluster, c_q . This procedure continues until no function changes cluster. We remark that the algorithm is possible because S is a closed and convex subset in a Hilbert space then it contains a unique element

of smallest norm (Rudin, 1986). If the functional space is nonconvex, the original observations $f_i(x)$ are replaced with the corresponding functions $g_i(x) = T\left(f_i(x)\right)$. Next, the procedure can be transported on S by the transformation T^{-1} .

3 Application

The framework described previously, has been applied to a real data set concerning biodiversity of epiphytic lichens in Liguria region, in northwestern Italy. Lichen biodiversity provide useful information about the global conditions affecting the environment over a given area (Giordani et al., 2002). These organisms, in fact, are particularly sensitive to environmental stresses, especially with regard to pollution, eutrophication and climate change (ANPA, 2001).

Data on lichen abundance have been collected following the standards suggested by Asta et al. (2002). The survey lasted from 2002 to 2003 and involves a total of 196 epiphytic lichen species and 153 plots (excluding the plots in which the lichen frequency is equal to zero).

Figure (1) shows the location of the sample units in the region.

For every i -th environmental site ($i = 1, \dots, 153$) and for each j -th species ($j =$

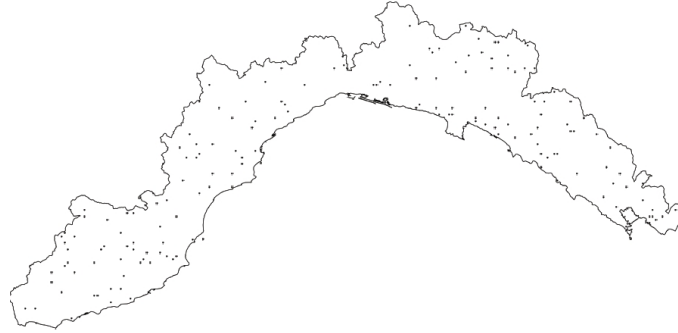


Fig. 1 Locations of the sample units in Liguria region.

$1, \dots, 196$), we consider an abundance vector, $\mathbf{N}_i = (N_{i1}, \dots, N_{is})'$, calculated as the sum of the lichen frequencies found on every t -th tree, $t = 1, 2, \dots, T$, belonging to the plot and a relative abundance vector, $\mathbf{p}_i = (p_{i1}, \dots, p_{is})'$, with $p_j = N_j / \sum_{j=1}^s N_j$, such that $0 \leq p_j \leq 1$ and $\sum_{j=1}^s p_j = 1$.

In order to evaluate lichen biodiversity we refer to a parametric families of diversity indices (Hill, 1973; Patil and Taillie, 1982), which are usually referred to as diversity profiles. They consist of a sequence of measurements allowing different aspects of community structure to be encompassed in a single diversity spectrum. In particular,

the Intrinsic diversity profile (Patil and Taillie, 1979) has been applied. It is defined as the plotting of the (l, T_l) pairs, where:

$$T_l = \sum_{j=l+1}^s p(j)^{\#} \quad l = 0, 1, \dots, s \quad (9)$$

is the right-tail sum diversity index, $p(j)^{\#}$ is the relative abundance vector ranked in descending order and l is the species abundance rank. Figure (2) displays the intrinsic profiles for the overall area with species abundances rank, l , on the abscissa. Since there are many rare species, figure (3) shows the intrinsic profiles focusing on the the first forty ordered species.

It is no possible to distinguish a site with greater diversity because the profiles cross

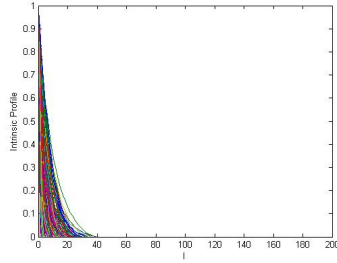


Fig. 2 Intrinsic profiles for Liguria region.

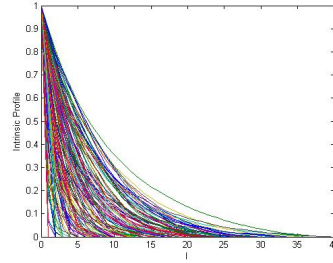


Fig. 3 Intrinsic profiles for the first 40 ordered species of Liguria.

each other. One solution could be to perform a pairwise comparison, when it is possible, but, in this way, a complete description of the study area is not provided.

Since diversity profile is a convex and decreasing curve, it can be analyzed in a functional framework (Gattone and Di Battista, 2009) and, in particular, in a parametric functional framework (Di Battista and Fortuna, 2013).

In order to identify specific common patterns among sites, a nonparametric clustering method has been applied. In particular, a functional k-means algorithm has been implemented, specifying an L^2 metric directly on the explicit known form of the functions. This is possible because the set of the sequences of intrinsic profiles in (9) is convex in the l^2 space (De Sanctis et al., 2014). The k-means procedure has been initialized by choosing three arbitrary centroids, $\phi_1^{(0)}(x), \phi_2^{(0)}(x), \phi_3^{(0)}(x)$. At each m -th iteration, the functions have been assigned to the cluster according to the minimum distance between them and the centroids:

$$d(T_{li}, \bar{T}_l) = \left(\sum_{i=1}^n |T_{li} - \bar{T}_l|^2 \right)^{\frac{1}{2}} \quad (10)$$

where the functional mean has been computed in the usual way as the average of the functions, obtaining an element of S due to the convexity property:

$$\overline{T}_l = \frac{1}{n} \sum_{i=1}^n T_{li} \quad (11)$$

The first group (dash-dot red line in figure 4) presents the grater diversity. It is composed of 61 sites which are mainly located in the province of Genoa. The second group (solid yellow line in figure 4), shows the opposite situation. It consists of 36 sites that are especially of the province of Savona. Finally, the third group (dotted blue line in figure 4), composed of 56 sites which are prevalently located in the province of Imperia, is placed between the other two, indicating a situation of medium diversity. Indeed the classification not return a clear ordered among sites.

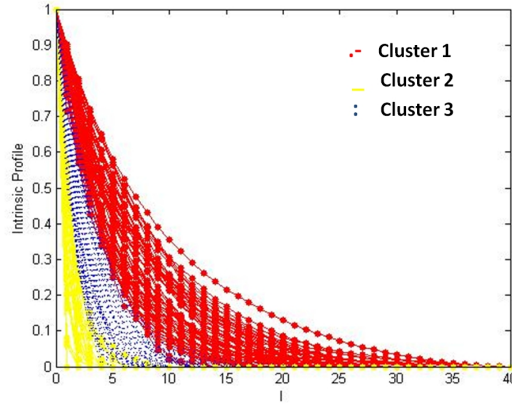


Fig. 4 Three clusters for T_{jl} for the first 40 ordered species.

In fact, some curves of the clusters continue to overlap with each other. The average intrinsic profiles of the three groups (fig. 5), however, return a clear ordering between sites without showing overlapping. The classification of the profiles, thus, returns three macro-areas in Liguria region. In particular, it is possible to distinguish an area of dominance (cluster two) with a few dominant species and, at the other extreme, an area of evenness (cluster one).

The proposed method, allows us to classify the intrinsic profiles computing the functional distance directly on the explicit form of the observations, leading to clustering results that not depend on how the curves are fit to the data. In particular, in an ecological setting, the combined used of parametric FDA and functional classification methods, allows us to identify different patterns of biodiversity when the intrinsic biodiversity profile not highlights an explicit ranking of biodiversity.

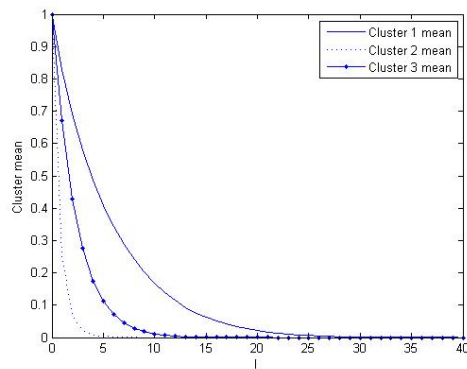


Fig. 5 Mean Intrinsic profile for the three clusters for the first 40 ordered species.

References

- Abraham, C., Cornillon, P., Matzner-Lober, E., Molinari, N., 2003. Unsupervised curve clustering using b-splines. *Scandinavian Journal of Statistics. Theory and applications* 30(3), 581–595.
- ANPA, 2001. I.B.L. Indice di Biodiversità Lichenica. Manuali e Linee Guida. no. 2.
- Asta, J., Erhardt, W., Ferretti, M., Fornasier, F., Kirschbaum, U., Nimis, P.L., Purvis, O.W., Pirintos, S., Scheidegger, C., Haluwyn, C.V., Wirth, V., 2002. Mapping lichen diversity as an indicator of environmental quality, in: Nimis, P.L., Scheidegger, C., Wolseley, P.A. (Eds.), *Monitoring with Lichens-Monitoring Lichens*. Nato Science Program-IV, Kluwer Academic Publisher, The Netherlands, pp. 273–279.
- Bouveyron, C., Brunet, C., 2012. Model-based clustering of high-dimensional data: A review. Technical report, Laboratoire SAMM, Université Paris 1 Panthéon-Sorbonne.
- Chiou, J., Li, P., 2007. Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69, 679–699.
- De Sanctis, A., Di Battista, T., 2012. Functional analysis for parametric families of functional data. *International Journal of Bifurcation and Chaos* 22 (9), 1250226–1–1250226–6.
- De Sanctis, A., Di Battista, T., Fortuna, F., 2014. Functional statistics on function spaces. *Test*, Submitted.
- Di Battista, T., Fortuna, F., 2013. Assessing biodiversity profile through fda. *Statistica* 1, 69–85.
- Ferraty, F., Vieu, P., 2006. *Nonparametric functional data analysis*. Springer, New York.

- Gattone, S., Di Battista, T., 2009. A functional approach to diversity profiles. *Journal of the Royal Statistical Society* 58, 267–284.
- Giordani, P., Brunialti, G., Alleteo, D., 2002. Effects of atmospheric pollution on lichen biodiversity (lb) in a mediterranean region (liguria, nw italy). *Environmental Pollution* 118, 53–64.
- Hill, M., 1973. Diversity and evenness: a unifying notation and its consequences. *Ecology* 54, 427–432.
- Ieva, F., Paganoni, A., Pigoli, D., Vitelli, V., 2013. Multivariate functional clustering for the analysis of ecg curves morphology. *Journal of the Royal Statistical Society. Series C. Applied Statistics* 62(3), 401–418.
- Jacques, J., Preda, C., 2004. Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis* 71, 92–106.
- James, G., Sugar, C., 2003. Clustering for sparsely sampled functional data. *Journal of the American Statistical Association* 98, 397–408.
- MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations, in: Le Cam, L., Neyman, J. (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. CA: University of California Press, Berkeley, pp. 281–297.
- Patil, G., Taillie, C., 1979. An overview of diversity, in: Grassle, J., Patil, G., Smith, W., Taillie, C. (Eds.), *Ecological Diversity in Theory and Practice*. International Co-operative Publishing House, Fairland, MD, pp. 23–48.
- Patil, G., Taillie, C., 1982. Diversity as a concept and its measurement. *Journal of the American Statistical Association* 77, 548–567.
- Peng, J., Muller, H., 2008. Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *The Annals of Applied Statistics* 2, 1056–1077.
- Ramsay, J., Silverman, B., 2005. *Functional Data Analysis*, 2nd edn. Springer, New York.
- 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.
- Rudin, W., 1986. *Real and complex analysis*. McGraw-Hill.
- Sam, A., Chamroukhi, F., govaert, G., Aknin, P., 2011. Model-based clustering and segmentation of times series with changes in regime. *Advances in Data Analysis and Classification* 5(4), 301–322.
- Tarpey, T., 2007. Linear transformations and the k-means clustering algorithm: applications to clustering curves. *Am Stat.* 61(1), 34–40.