Distance - Based Statistics for Covariance Operators in Functional Data Analysis

Davide Pigoli

Abstract The statistical analysis of covariance operators in a functional data analysis setting is considered. Many suitable distances to compare covariance operators are presented and in particular the problem of estimating the average covariance operators among different groups is addressed. Finally, an applied problem in which this methodology has proved useful is introduced, namely, exploring phonetic relationships among Romance languages looking at covariance operators across frequencies.

Key words: Trace class Operators, Functional Data, Linguistic Data

1 Introduction

The aim of this work is to set up a framework for the comparison of covariance operators on $L^2(\Omega)$, $\Omega \subseteq \mathbb{R}$. This problem arises in Functional Data Analysis when features of curve populations lie in their covariance structure rather than in the mean function. In Section 2 some definitions and properties of operators on $L^2(\Omega)$ are recalled. Section 3 illustrates suitable distances to measure differences between covariance operators and to explore their properties. In Section 4, the application of the proposed methodology to a linguistic problem is introduced and some preliminary results are shown.

Davide Pigoli
MOX- Department of Mathematics, Politecnico di Milano , Piazza Leonardo da Vinci, 32, 20133 Milano, Italy. e-mail: davide.pigoli@mail.polimi.it

2 Some remarks on compact operators on $L^2(\Omega)$

In this section we review some properties and definitions that will be of use when describing our proposed methodology. More details and proofs can be found, e.g., in Zhu (2007).

Definition 1. Let B_1 be the closed ball in $L^2(\Omega)$, i.e. it consists in all $f \in L^2(\Omega)$ so that $||f||_{L^2(\Omega)} \leq 1$. A bounded linear operator $T : L^2(\Omega) \to L^2(\Omega)$ is compact if $T(B_1)$ is compact in the norm of $L^2(\Omega)$. A bounded linear operator T is self-adjoint if $T = T^*$.

An important property of compact operators on $L^2(\Omega)$ is the existence of a canonical decomposition. This means that two orthonormal bases $\{u_k\}_k, \{v_k\}_k$ exist so that

$$Tf = \sum_{k=1}^{+\infty} \sigma_k \langle f, v_k \rangle u_k,$$

or, equivalently,

$$Tv_k = \sigma_k u_k,$$

where $\langle .,. \rangle$ indicates the scalar product in $L^2(\Omega)$. $\{\sigma_k\}$ is called the sequence of singular value for T. If the operator is self-adjoint, a basis $\{v_k\}_k$ exists such that

$$Tf = \sum_{k=1}^{+\infty} \lambda_k \langle f, v_k \rangle v_k,$$

or, equivalently,

$$Tv_k = \lambda_k v_k$$

and $\{\lambda_k\}$ is called the sequence of eigenvalues for T.

A compact operator T is said to be *trace class* if

$$\text{trace}(T) := \sum_{k=1}^{+\infty} \langle Te_k, e_k \rangle < +\infty$$

for an orthonormal basis $\{e_k\}$. It has been proved that the definition is independent from the choice of the basis and

$$\text{trace}(T) = \sum_{k=1}^{+\infty} \sigma_k$$

where $\{\sigma_k\}_k$ are singular values for T. We indicate with $S(L^2(\Omega))$ the space of the trace class operator on $L^2(\Omega)$.

A compact operator T is said to be Hilbert-Schmidt if its Hilbert-Schmidt norm is bounded, i.e.

$$||T||_{HS}^2 = \text{trace}(T^*T) < +\infty.$$

This is a generalization of the Frobenius norm for finite-dimensional matrices.

Definition 2. A bounded linear operator R on $L^2(\Omega)$ is said to be unitary if

$$||Rf||_{L^2(\Omega)} = ||f||_{L^2(\Omega)} \quad \forall f \in L^2(\Omega)$$

We indicate with $SO(L^2(\Omega))$ the space of unitary operators on $L^2(\Omega)$.

Let now $\mathbf{f}$ be a random variable which takes values in $L^2(\Omega)$, $\Omega \subseteq \mathbb{R}$, such that $\mathbb{E}[||\mathbf{f}||^2_{L^2(\Omega)}] < +\infty$. Then, the covariance operator $C_{\mathbf{f}}(s,t) = \mathrm{cov}(\mathbf{f}(s), \mathbf{f}(t))$ is a trace class compact operator on $L^2(\Omega)$ (see Bosq, 2000, Section 1.5).

3 Distances between covariance operators

In this section novel distances to compare trace class compact operators are proposed. These are generalizations to the functional setting of metrics that have been proved useful for the case of positive definite matrices.

Distance between kernels in $L^2(\Omega)$

Every covariance operator S on $L^2(\Omega)$ can be associated with an integral kernel $s(x,y) \in L^2(\Omega \times \Omega)$, so that

$$Sf = \int_\Omega s(x,y) f(y) \mathrm{d}y, \quad \forall \ f \in L^2(\Omega).$$

Thus, distance between covariance operators can be naturally defined with the distance between kernels in $L^2(\Omega)$,

$$d_L(S_1, S_2) = ||s_1 - s_2||_{L^2(\Omega)} = \sqrt{\int_\Omega \int_\Omega (s_1(x,y) - s_2(x,y))^2 \mathrm{d}x \mathrm{d}y}.$$

This distance is correctly defined, since it inherits all the properties of the distance in the Hilbert space $L^2(\Omega)$. However, it does not exploit in any way the particular structure of the covariance operators and therefore it may not highlight the significant differences between covariance structures.

Spectral distance

A second possibility is to see the covariance operator as an element of $\mathfrak{L}(L^2(\Omega))$, the space of the linear bounded operators on $L^2(\Omega)$. It follows that the distance between S_1 and S_2 can be defined as the operator norm of the difference. We recall that the norm of a self-adjoint bounded linear operator on $L^2(\Omega)$ is defined as

$$||T||_{\mathfrak{L}(L^2(\Omega))} = \sup_{v \in L^2(\Omega)} \frac{|\langle Tv, v\rangle|}{||v||^2_{L^2(\Omega)}}$$

and for a covariance operator it coincides with the absolute value of the first (i.e. largest) eigenvalue. Thus,

$$d_{\mathfrak{L}}(S_1,S_1) = ||S_1 - S_2||_{\mathfrak{L}(L^2(\Omega))} = |\widetilde{\lambda}_1|$$

where $\widetilde{\lambda}_1$ is the first eigenvalue of the operator $S_1 - S_2$. $d_{\mathfrak{L}}(.,.)$ generalizes the matrix spectral norm which is often used in the finite dimensional case (see, e.g., El Karoui, 2008). This distance takes into account the spectral structure of the covariance operators, but it seems somehow restrictive to focus only on the behavior on the first mode of variation.

Procrustes size-and-shapes distance

In Dryden et al. (2009), a Procrustes size-and-shape distance is proposed to compare two positive definite matrices. Our aim is to generalize this distance to the case of covariance operators on $L^2(\Omega)$. Let S_1 and S_2 be two trace class covariance operators on L^2. We define the Procrustes distance in $S(L^2(\Omega))$ as

$$d_P(S_1,S_2)^2 = \inf_{R\in SO(L^2(\Omega))} ||L_1 - L_2R||_{HS}^2 = \inf_{R\in SO(L^2(\Omega))} \text{trace}((L_1-L_2R)^*(L_1-L_2R)),$$

where $||.||_{HS}$ indicates the Hilbert-Schmidt norm on $L^2(\Omega)$ and L_i are so that $S_i = L_i^* L_i$. The evaluation of the Procrustes distance asks for the solution of a minimization problem. However, an analytical solution is available and the distance has therefore an expression based on the canonical decomposition of the operator $L_2^* L_1$. The unitary operator $\widetilde{R}$ that minimizes $||L_1 - L_2R||_{HS}^2$ is defined by

$$\widetilde{R}v_k = u_k \quad \forall k = 1,\dots,+\infty.$$

where $\{u_k\}_k, \{v_k\}_k$ are the orthogonal bases in the canonical decomposition of $L_2^* L_1$.

Proposition 1. *The Procrustes distance in $S(L^2(\Omega))$ is*

$$d_P(S_1,S_2)^2 = ||L_1||_{HS}^2 + ||L_2||_{HS}^2 - 2\sum_{k=1}^{+\infty} \sigma_k$$

where σ_k are the singular values of the compact operator $L_2^* L_1$.

Square root operator distance

We can also generalize the square root matrix distance (see Dryden et al., 2009) to compare $S_1, S_2 \in S(L^2(\Omega))$. Since $S_i^{1/2}$ is an Hilbert-Schmidt operator,

$$d_R(S_1,S_2) = ||S_1^{1/2} - S_2^{1/2}||_{HS}$$

This is a special case of the Procrustes distance above, when no unitary transformation is allowed.

3.1 Averaging of covariance operators

Once appropriate distances for dealing with covariance operators have been defined, many statistical tools can be developed, conveniently generalizing traditional methods based on Euclidean distance. For the sake of brevity, here only the case of estimating the average from a sample of covariance operators is presented. Let $S_1, \dots, S_g$ be the covariance operators for g different groups. Then, a possible estimator of the common covariance operator Σ may be

$$\widehat{\Sigma} = \frac{1}{n_1 + \cdots + n_g}(n_1 S_1 + \cdots + n_g S_g).$$

However, this formula arises from the minimization of square Euclidean deviations, weighted with the number of observations. If we choose a different distance to compare covariance operators, it is more coherent to average covariance operators with respect to the chosen distance. A least square estimator for Σ can be defined for a general distance $d(.,.)$,

$$\widehat{\Sigma} = \arg\min_S \sum_{i=1}^{g} n_i d(S, S_i)^2.$$

The actual computation of the sample Frechét mean $\widehat{\Sigma}_j$ depends on the choice of the distance $d(.,.)$. In general, it asks for the solution of a high dimensional minimization problem but some distances allows for an analytic solution while for others efficient minimization algorithms are available. For those concerning Kernel distances, it is easy to see that the $L^2(\Omega)$ kernel of the Frechét average is obtained with the weighted average of the kernels $s_1(s,t), \dots, s_g(s,t)$ of the data, i.e.

$$\hat{\sigma}(x,y) = \frac{1}{n_1 + \cdots + n_g}(n_1 s_1(x,y) + \cdots + n_g s_g(x,y)).$$

For the Square root distance, the following result can be proved.

Proposition 2.

$$\widehat{\Sigma} = \arg\min_S \sum_{i=1}^{g} n_i d_S(S, S_i)^2 = \left(\frac{1}{G}\sum_{i=1}^{g} n_i S_i^{\frac{1}{2}}\right)^2. \tag{1}$$

where $G = n_1 + \cdots + n_g$.

The Procrustes mean can be obtained by an adaptation of the algorithm proposed in Gower (1975) or Ten Berge (1977). This works very well in practice if the algorithm is initialized with the estimate provided by (1).

4 Exploring phonetic relationships among Romance languages

The traditional way of exploring relationships among languages consists in looking at textual similarity. However, this often neglects any phonetic characteristics of the languages. Here a novel approach is proposed to compare languages on the basis of phonetic structure.

In particular, people speaking different languages (French, Italian, Portuguese, Iberian Spanish and American Spanish) are registered while pronouncing words corresponding with the numbers from one to ten in each language. The output of the registration for each word and for each speaker consist in the intensity of the sound over time and frequencies.

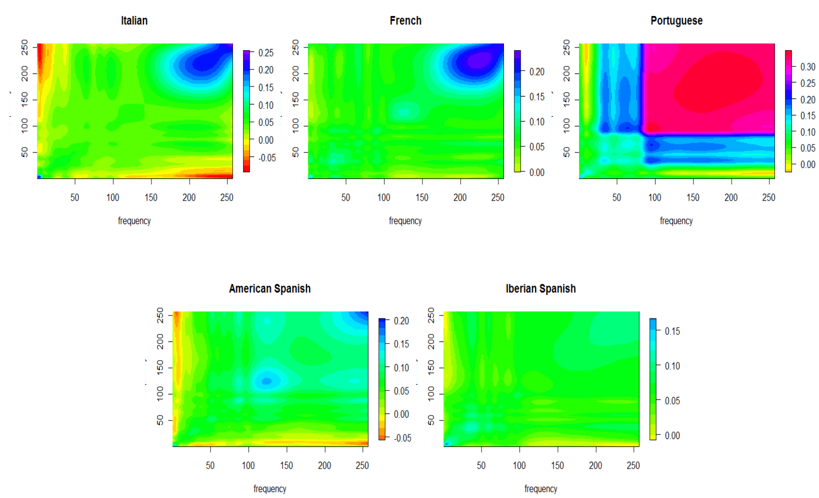

Fig. 1 Frechét average along time of covariance operators of log-spectrogram among frequencies for five Romance languages, using square root distance.

The aim is to use this data to explore linguistic hypotheses concerning the relationship among different languages. However, while many possible phonetic features may be of interest, it has been shown that covariance operators associated with frequencies can provide some phonetic insight (Hajipantelis *et al.*, 2012). Frequency covariances indeed can summarize phonetic information for the language, disregarding particular characteristics of speakers and words. For the scope of this work, we focus on the covariance operators among frequencies obtained from the log-spectrogram with estimates being obtained using the sample of all speaker of the language. We consider different time points as replicates of the same covariance operator among frequencies. It is clear that this is a major simplification of the rich structure in the data but it already leads to some interesting conclusions. Here some preliminary results are reported, focusing on the covariance operator for the word "one".

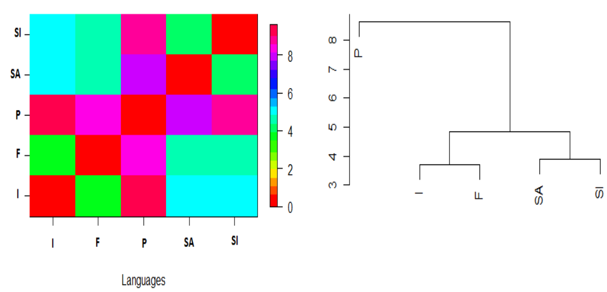

Fig. 2 Left: Distance matrix among Frechét average of Fig. 1, obtained with Square root distance. Right: Dendrogram obtained from distance matrix using an average linking, where I=Italian, F=French, P=Portuguese, SA=American Spanish, SI=Iberian Spanish.

Fig. 1 shows the covariance operator estimated for each language via Frechét averaging along time, using square root distance, for the word "one". Fig. 2 shows dissimilarity matrix among average covariance operators for each language and the correspondent dendrogram, while Fig. 3 compares a two-dimensional projection of the data obtained with a classical (metric) multidimensional scaling with the map coming from linguistic experts, containing information about historical and geographical relationship among languages. Indeed, it seems that focusing on the covariance operator captures some important information about languages. There is an overall similarity between the map predicted by experts and relationships among covariance structures. However, some unexpected features may suggest new research lines. For example, it is worth to notice that Portuguese covariance structure is a considerable distance from all the others, thus highlighting particular linguistic influences on the language.

5 Conclusions

In this work the problem of dealing with the covariance operator has been addressed. The choice of the appropriate metric is crucial in the analysis of covariance operators. Here some suitable metrics have been proposed and their properties have been highlighted. On the basis of appropriate metric, statistical methods can be developed to deal with covariance operators in the functional data analysis framework. The notable case of estimating the average from a sample of covariance operators is

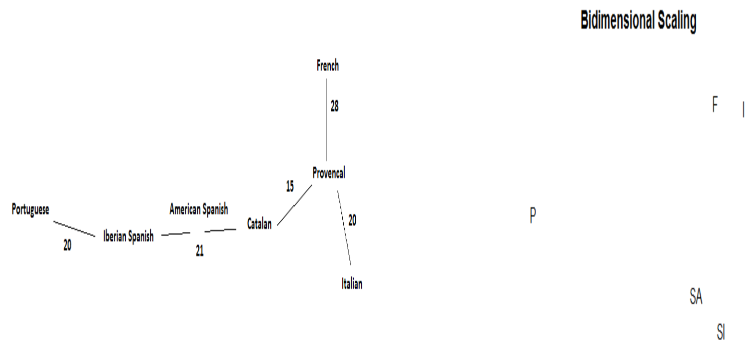

Fig. 3 Left:Map of languages built by linguistic experts using historical and geographical information. Some languages are shown for whom phonetic data are not available. Right: Bidimensional metric multidimensional scaling. The extreme behavior of the Portuguese language lead to a slightly difference configuration. Label correspond to languages: I=Italian, F=French, P=Portuguese, SA=American Spanish, SI=Iberian Spanish.

illustrated. Moreover, in many applications, the covariance operator itself is the object of interest, as illustrated by the linguistic data of Section 4. Using the square root distance between covariance operator among frequencies, some significant phonetic features of Romance languages have been found.

References

1. Bosq, D. : Linear processes in function spaces. Springer, New York (2000)
2. Dryden, I.L., Koloydenko, A., Zhou, D. : Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. Ann. Appl. Stat. **3**, 1102-1123 (2009)
3. El Karoui, N. : Operator norm consistent estimation of large-dimensional sparse covariance matrices. Ann. Stat. **36**, 2717–2756 (2008)
4. Gower, J. C.: Generalized Procrustes analysis. Psychometrika **40**,33–50 (1975)
5. Hadjipantelis, P.Z., Aston, J.A.D., Evans, J.P. : Characterizing fundamental frequency in Mandarin: A functional principal component approach utilizing mixed effect models. J. Acoust. Soc. Am. In press (2012)
6. Ten Berge, J. M. F. : Orthogonal Procrustes rotation for two or more matrices. Psychometrika **42**, 267–276 (1977)
7. Zhu, K. : Operator theory in function spaces (2nd ed.). American Mathematical Society (1977)