Biometrika (2014), **101**, 2, *pp.* 409–422
Printed in Great Britain
doi: 10.1093/biomet/asu008
Advance Access publication 17 April 2014

Distances and inference for covariance operators

By DAVIDE PIGOLI

Department of Statistics, University of Warwick, Coventry, CV4 7AL, U.K.
d.pigoli@warwick.ac.uk

JOHN A. D. ASTON

Statistical Laboratory, University of Cambridge, Wilberforce Rd, Cambridge, CB3 0WB, U.K.
j.aston@statslab.cam.ac.uk

IAN L. DRYDEN

School of Mathematical Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD, U.K.
ian.dryden@nottingham.ac.uk

and PIERCESARE SECCHI

Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milan, Italy
piercesare.secchi@polimi.it

Summary

A framework is developed for inference concerning the covariance operator of a functional random process, where the covariance operator itself is an object of interest for statistical analysis. Distances for comparing positive-definite covariance matrices are either extended or shown to be inapplicable to functional data. In particular, an infinite-dimensional analogue of the Procrustes size-and-shape distance is developed. Convergence of finite-dimensional approximations to the infinite-dimensional distance metrics is also shown. For inference, a Fréchet estimator of both the covariance operator itself and the average covariance operator is introduced. A permutation procedure to test the equality of the covariance operators between two groups is also considered. Additionally, the use of such distances for extrapolation to make predictions is explored. As an example of the proposed methodology, the use of covariance operators has been suggested in a philological study of cross-linguistic dependence as a way to incorporate quantitative phonetic information. It is shown that distances between languages derived from phonetic covariance functions can provide insight into the relationships between the Romance languages.

Some key words: Distance metric; Functional data analysis; Procrustes analysis; Shape analysis.

1. Introduction

Datasets are increasingly becoming available that are best described as functional. In recent years, many statistical techniques have been proposed to deal with functional data (see, e.g., Ramsay & Silverman, 2005; Ferraty & Vieu, 2006; Horváth & Kokoszka, 2012). However, this

body of work has mainly focused on mean functions, with little attention paid to the analysis of the covariance operator, which is either directly or indirectly of interest in its own right.

Recent work (Panaretos et al., 2010; Fremdt et al., 2013) has examined the testing of equality of covariance structures from two groups of functional curves by defining a test statistic through the Karhunen–Loève expansions of the two covariance structures. These methods are therefore based on the Hilbert–Schmidt metric, exploiting the immersion of the space of covariance operators in the Hilbert–Schmidt space. However, such an extrinsic approach ignores the geometry of the space of covariance operators.

We consider the problem of defining possible metrics for covariance operators. Making use of various distances between covariance operators, some of which are newly introduced, this paper develops a formal two-sample test for comparing covariance operators and proposes an exploratory technique based on the Fréchet mean and extrapolation.

Analysis of the covariance operator arises in many applied contexts. In the linguistic analysis of human speech, the main interest is often not in the mean of speech frequency intensity but rather in the variations that can be found within the language. In § 5, we show that different languages can be compared and even predicted by using functional distances, allowing a quantitative analysis of comparative philological relations based on speech recordings rather than discrete textual analysis.

2. DISTANCES BETWEEN COVARIANCE OPERATORS

In this section we consider the functional extension of metrics that have proved useful for positive-semidefinite matrices (Dryden et al., 2009). For a more detailed discussion of the underlying Hilbert space model for functional data, see Horváth & Kokoszka (2012, pp. 21–36).

Let f be a random function taking values in $L^2(\Omega)$, where $\Omega \subseteq \mathbb{R}$, such that $E(\|f\|^2_{L^2(\Omega)}) < +\infty$. The covariance operator C_f is defined, for $g \in L^2(\Omega)$, by $C_f g(t) = \int_\Omega c_f(t', t) g(t')\,\mathrm{d}t'$, where $c_f(t', t) = \mathrm{cov}\{f(t'), f(t)\} = E([f(t') - E\{f(t')\}][f(t) - E\{f(t)\}])$. Then C_f is a trace-class, self-adjoint, compact operator on $L^2(\Omega)$ with nonnegative eigenvalues (see, e.g., Bosq, 2000, § 1.5). Any compact operator T has a canonical decomposition that implies the existence of two orthonormal bases $\{u_k\}$ and $\{v_k\}$ for $L^2(\Omega)$ such that $Tf = \sum_k \sigma_k \langle f, v_k\rangle u_k$ or, equivalently, $Tv_k = \sigma_k u_k$, where $\langle v, v\rangle$ denotes the inner product in $L^2(\Omega)$. The sequence $\{\sigma_k\} \in \mathbb{R}$ is called the sequence of singular values for T. If the operator is self-adjoint, there exists an orthonormal basis $\{v_k\}$ such that $Tf = \sum_k \lambda_k \langle f, v_k\rangle v_k$ or, equivalently, $Tv_k = \lambda_k v_k$, and the sequence $\{\lambda_k\} \in \mathbb{R}$ is called the sequence of eigenvalues for T. A compact operator T is said to be trace class if $\mathrm{tr}(T) = \sum_k \langle Te_k, e_k\rangle < +\infty$ for every orthonormal basis $\{e_k\}$. In the case of nonnegative operators, it can be shown that this definition is independent of the choice of basis, and for nonnegative-definite self-adjoint operators, the trace is equivalent to the sum of the eigenvalues. Let $S\{L^2(\Omega)\}$ denote the space of trace-class operators on $L^2(\Omega)$. Finally, a compact operator T is said to be Hilbert–Schmidt if its Hilbert–Schmidt norm is bounded, i.e., $\|T\|^2_{\mathrm{HS}} = \mathrm{tr}(T^*T) < +\infty$, which is a generalization of the Frobenius norm for finite-dimensional matrices.

Not all matrix-based distances are extendable in the functional case. Two popular metrics for finite-dimensional covariance matrix analysis are the log-Euclidean metric and the affine invariant Riemannian metric. While both would appear to be natural candidates for generalization to covariance operators, their generalization is not straightforward due to the natural trace-class structure of the covariance operator, which implies that the eigenvalues λ_i in descending order are summable, i.e., $\sum_i \lambda_i < \infty$, and this implies that $\lambda_i \to 0$ as $i \to \infty$. The log-Euclidean distance (Arsigny et al., 2006) for two positive-definite matrices M_1 and M_2 is defined as

$d_{\log}(M_1, M_2) = \|\log(M_1) - \log(M_2)\|$, where $\|\cdot\|$ indicates the Frobenius norm and $\log(\cdot)$ the matrix logarithm, i.e., given a spectral decomposition $M = VDV^{-1}$, $\log(M) = V\log(D)V^{-1}$ with $[\log(D)]_{ii} = \log([D]_{ii})$ and $[\log(D)]_{ij} = 0$ for $i \neq j$. The matrix logarithm is not extendable to infinite-dimensional trace-class operators, as the $\log([D]_{ii})$ tend to minus infinity. The affine invariant Riemannian metric (Pennec et al., 2006) for positive-definite matrices is defined as $d_{\mathrm{Riem}}(M_1, M_2) = \|\log(M_1^{-1/2} M_2 M_1^{-1/2})\|$, which requires calculation of the inverse. Even for a positive-definite compact operator, the inverse is unbounded in general (Zhu, 2007, § 1.3).

Even though in applications only finite-dimensional representations are available, these are usually not of full rank, i.e., they have zero eigenvalues. This means that the above metrics must be computed on subspaces which should be carefully chosen to avoid instability in the computation of the distance coming from small eigenvalues, while taking into account all the significant information. Indeed, the way the distances change as the dimension of the approximation increases is also an issue (Fremdt et al., 2013). We therefore use some alternative distances which are well-defined for self-adjoint trace-class operators with nonnegative eigenvalues.

A distance between covariance operators can be defined naturally by using the distance between their kernels in $L^2(\Omega \times \Omega)$. Let S_1 and S_2 be two covariance operators, and let $S_i f(t) = \int_\Omega s_i(t', t) f(t')\, \mathrm{d}t'$ for $f \in L^2(\Omega)$. Then, we can define the kernel distance as

$$d_L(S_1, S_2)^2 = \|s_1 - s_2\|^2_{L^2(\Omega\times\Omega)} = \int_\Omega \int_\Omega \{s_1(t', t) - s_2(t', t)\}^2 \, \mathrm{d}t' \, \mathrm{d}t. \tag{1}$$

This distance is well-defined, since it inherits all the properties of the distance in the Hilbert space $L^2(\Omega \times \Omega)$. Indeed, (1) is the distance induced by the Hilbert–Schmidt norm, since for Hilbert–Schmidt kernel operators one has $\|S_1 - S_2\|_{\mathrm{HS}} = \|s_1 - s_2\|_{L^2(\Omega\times\Omega)}$. Thus, the kernel distance exploits the immersion of the space of covariance operators in the Hilbert–Schmidt space, being an extrinsic metric that ignores the geometry of the space of interest. In addition, as will be seen later, the kernel distance is not constrained always to provide estimates within the space of covariance operators when used for extrapolation or prediction.

A second possibility is to regard the covariance operator as an element of $\mathfrak{L}\{L^2(\Omega)\}$, the space of bounded linear operators on $L^2(\Omega)$, in which case the distance between S_1 and S_2 can be defined as the operator norm of their difference. We recall that the norm of a self-adjoint bounded linear operator on $L^2(\Omega)$ is defined as $\|T\|_{\mathfrak{L}\{L^2(\Omega)\}} = \sup_{v\in L^2(\Omega)} |\langle Tv, v\rangle| / \|v\|^2_{L^2(\Omega)}$ and that for a covariance operator this coincides with the absolute value of the largest eigenvalue. Thus

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

where $\tilde{\lambda}_1$ is the eigenvalue of the operator $S_1 - S_2$ with the largest absolute value. The distance $d_{\mathfrak{L}}(\cdot\,, \cdot)$ generalizes the matrix spectral norm 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 is restrictive in the sense that it focuses only on the behaviour of the first mode of variation. This can describe effectively the distance between the operators only if $\tilde{\lambda}_1$ explains the majority of the variation, which is often not the case in practical applications.

Since covariance operators are trace-class operators, we can also generalize the square root matrix distance (Dryden et al., 2009) to them. Indeed, S being a self-adjoint trace-class operator, there exists a Hilbert–Schmidt self-adjoint operator

$$(S)^{1/2} f = \sum_k \lambda_k^{1/2} \langle f, v_k\rangle v_k,$$

where the λ_k are eigenvalues and the v_k eigenfunctions of S. We can therefore define the square root distance between two covariance operators S_1 and S_2 as

$$d_{\mathrm{R}}(S_1, S_2) = \|S_1^{1/2} - S_2^{1/2}\|_{\mathrm{HS}}. \tag{2}$$

The square root transformation has been shown to produce good results in the finite-dimensional setting (Dryden et al., 2009), but is also well-defined for trace-class operators. Any power greater than $1/2$ would be a possible candidate distance, since $\|S^\alpha\|_{\mathrm{HS}} < +\infty$ for all $\alpha \geqslant 1/2$. For general trace-class operators, the square root operator is the smallest power that can be defined while still ensuring finite distances, meaning that it is the closest available to the log-Euclidean distance, which is the limit when $\alpha \to 0$. In addition, it can be interpreted as a distance which takes into account the full eigenstructure of the covariance operator.

However, (2) is only one particular choice from a broad family of distances based on the mapping of two operators S_1 and S_2 from the space of covariance operators to the space of Hilbert–Schmidt operators. In general, we can consider a transformation from S_i to L_i such that $S_i = L_i L_i^*$, and define the distance to be the Hilbert–Schmidt norm of $L_1 - L_2$. It is easy to see that any such transformation is defined only up to a unitary operator R, since $(L_i R)(L_i R)^* = L_i R R^* L_i^* = L_i L_i^* = S_i$. One way to make the approach well-defined is to choose the particular unitary operator R which minimizes the distance between the two operators L_1 and L_2. In Dryden et al. (2009), a Procrustes distance is proposed to compare two positive-definite matrices, which we now generalize. Let S_1 and S_2 be two covariance operators on $L^2(\Omega)$. We define the square of the Procrustes reflection size-and-shape distance between S_1 and S_2 as

$$d_{\mathrm{P}}(S_1, S_2)^2 = \inf_{R \in \mathrm{O}\{L^2(\Omega)\}} \|L_1 - L_2 R\|_{\mathrm{HS}}^2 = \inf_{R \in \mathrm{O}\{L^2(\Omega)\}} \mathrm{tr}\{(L_1 - L_2 R)^*(L_1 - L_2 R)\},$$

where the L_i $(i = 1, 2)$ are such that $S_i = L_i L_i^*$ and $\mathrm{O}\{L^2(\Omega)\}$ is the space of unitary operators on L^2.

PROPOSITION 1. *Let σ_k be the singular values of the compact operator $L_2^* L_1$. Then*

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

The proof can be found in the Supplementary Material. The map from the space of covariance operators to the space of Hilbert–Schmidt operators associates S_i with the equivalence class of operators L_i such that $S_i = L_i L_i^*$. The Procrustes distance looks for the operator $\tilde{L}_i$ in this class that minimizes the distance between the two covariance operators, and $\tilde{L}_i$ can be non-self-adjoint.

In applications, we observe only a finite-dimensional representation of the operators of interest. Therefore, ideally we would require the square root distance or the Procrustes size-and-shape distance between two finite-dimensional representations to be a good approximation of the distance between the infinite-dimensional operators. We consider the more general case of Procrustes distance; the square root distance is a special case where $L_i = (S_i)^{1/2}$ and R is constrained to be the identity operator. Let $\{e_k\}_{k=1}^{\infty}$ be a basis for $L^2(\Omega)$, let $V_p = \mathrm{span}(e_1, \ldots, e_p)$, and let S_i^p be the restriction of S_i on V_p, i.e.,

$$S_i^p g = \sum_{k=1}^{p} \langle g, e_k \rangle S_i e_k, \quad g \in V_p.$$

In practical situations, V_p will be the subspace containing the finite-dimensional representation of the functional data. Let us assume that as $p \to +\infty$, $L_i^p \to L_i$ with respect to the Hilbert–Schmidt norm, where $S_i^p = L_i^p L_i^{p*}$ and we can choose, for instance, $L_i = (S_i)^{1/2}$, although any choice for which convergence is guaranteed is suitable. Then, the distance between the two restricted operators satisfies

$$d_{\mathrm{P}}(S_1^p, S_2^p)^2 = \|L_1^p\|_{\mathrm{HS}}^2 + \|L_2^p\|_{\mathrm{HS}}^2 - 2\sum_{k=1}^{p} \langle \tilde{R}^{p*} L_2^{p*} L_1^p e_k, e_k \rangle.$$

Let $\{v_k\}_{k=1}^{+\infty}$ be the orthonormal basis obtained from the canonical decomposition of $L_2^* L_1$ (Zhu, 2007, § 1.3). Since $V_p \subset L^2(\Omega)$, we can choose a subset $v_1^p, \ldots, v_p^p$, with $v_k^p \in \{v_k\}_{k=1}^{+\infty}$, which is an orthonormal basis for V_p. However, these v_k^p need not be the first p elements of the basis coming from the canonical decomposition of $L_2^* L_1$, because the space V_p depends only on the original basis $\{e_k\}_{k=1}^p$ and does not depend on the covariance structure of the data. Since the subspaces V_p are nested, we can define a permutation $s : \mathbb{N} \to \mathbb{N}$ such that $\{v_{s(1)}, \ldots, v_{s(p)}\}$ forms a basis for V_p, for every p. Since the trace of an operator does not depend on the basis choice, we obtain

$$\begin{aligned} d_{\mathrm{P}}(S_1^p, S_2^p)^2 &= \|L_1^p\|_{\mathrm{HS}}^2 + \|L_2^p\|_{\mathrm{HS}}^2 - 2\sum_{k=1}^{p} \langle \tilde{R}^p L_2^{p*} L_1^p v_{s(k)}, v_{s(k)} \rangle \\ &= \|L_1^p\|_{\mathrm{HS}}^2 + \|L_2^p\|_{\mathrm{HS}}^2 - 2\sum_{k=1}^{p} \sigma_{s(k)}, \end{aligned}$$

where $\{\sigma_{s(k)}\}_{k=1}^p$ are the singular values for $L_2^* L_1$, because the action of the operator $L_2^{p*} L_1^p$ should be equal to the action of the operator $L_2^* L_1$ on every element belonging to the subspace V_p, and $v_{s(k)} \in V_p$ for $k = 1, \ldots, p$. Finally, as $L_2^* L_1$ is trace class, the series of its singular values is absolutely convergent and therefore also unconditionally convergent, i.e., convergent under any permutation. Thus, as the number p of basis functions increases, we have

$$\begin{aligned} \lim_{p \to +\infty} d_{\mathrm{P}}(S_1^p, S_2^p)^2 &= \|L_1\|_{\mathrm{HS}}^2 + \|L_2\|_{\mathrm{HS}}^2 - 2\sum_{k=1}^{\infty} \sigma_{s(k)} \\ &= \|L_1\|_{\mathrm{HS}}^2 + \|L_2\|_{\mathrm{HS}}^2 - 2\sum_{k=1}^{\infty} \sigma_k = d_{\mathrm{P}}(S_1, S_2)^2. \end{aligned}$$

3. A test for two-sample comparison of covariance structures

Let us consider two samples of random curves. Curves in the first sample, $f_1^1(t), \ldots, f_{n_1}^1(t) \in L^2(\Omega)$, are realizations of a random process with mean $\mu(t)$ and covariance operator Σ_1. Curves in the second sample, $f_1^2(t), \ldots, f_{n_2}^2(t) \in L^2(\Omega)$, are realizations of a random process with mean $\mu(t)$ and covariance operator Σ_2. We would like to test the null hypothesis $\mathrm{H}_0 : \Sigma_1 = \Sigma_2$ against the alternative $\mathrm{H}_1 : \Sigma_1 \neq \Sigma_2$. We reformulate the test using distances between covariance operators, i.e., we test $\mathrm{H}_0 : d(\Sigma_1, \Sigma_2) = 0$ against $\mathrm{H}_1 : d(\Sigma_1, \Sigma_2) > 0$. Let S_1 and S_2 be the sample covariance operators of the two groups. We use $d(S_1, S_2)$ as a test statistic, since large values of $d(S_1, S_2)$ provide evidence against the null hypothesis. We consider M random permutations

of the labels $\{1, 2\}$ on the sample curves and compute $d(S_1^{(m)}, S_2^{(m)})$ $(m = 1, \ldots, M)$, where $S_i^{(m)}$ is the sample covariance operator for the group indexed with label i in permutation m. The p-value of the test is the proportion of $d(S_1^{(m)}, S_2^{(m)})$ which are greater than or equal to $d(S_1, S_2)$. For this formulation of the permutation test, equality of mean functions is essential. However, if the two groups have different and unknown means, an approximate permutation test can be performed after the curves are centred using their sample means. This is a common strategy for testing scale parameters, such as variance, for univariate real random variables (e.g., Good, 2005, § 3.7.2).

We now apply the proposed permutation procedure to simulated data. Our main purpose is to explore the behaviour of the different distances with various modifications of the covariance structure. We use as benchmarks the empirical power of the testing procedure proposed by Panaretos et al. (2010), with the test statistic $T_N(K)$, and that proposed by Fremdt et al. (2013), with the test statistic $\hat{T}_1$. Further details of these procedures are given in the Supplementary Material. The parameter K in these methods is the number of eigendirections considered, and it is chosen to be the number of eigenvalues of the pooled covariance operator that explain at least 90% of the variability. The performance of these tests can be improved by making different choices of K and modifications of the test statistics to look for specific differences between covariance operators (Panaretos et al., 2010), but our results show that the proposed permutation procedure competes well with existing techniques in many situations of practical interest.

We consider two groups with the same mean function $\sin(x)$ and covariance operators Σ_1 and $\Sigma(\gamma) = [(\Sigma_1)^{1/2} + \gamma\{(\Sigma_2)^{1/2}\tilde{R} - (\Sigma_1)^{1/2}\}][(\Sigma_1)^{1/2} + \gamma\{(\Sigma_2)^{1/2}\tilde{R} - (\Sigma_1)^{1/2}\}]^*$, where $\tilde{R}$ is an operator minimizing the Procrustes distance between Σ_1 and Σ_2; see the proof of Proposition 1. For $\gamma = 0$, the two groups thus have the same covariance operator. As γ increases, the difference between the two operators increases. In the following simulations, Σ_1 and Σ_2 are the sample covariance operators for the males and females, respectively, in the Berkeley growth curve dataset (Ramsay & Silverman, 2005), rescaled to lie in $[0, 1]$. The integral kernels of Σ_1 and Σ_2 can be found in the Supplementary Material.

For the first simulation, all the curves are simulated on $[0, 1]$ with a Gaussian process. Observations are generated on a grid of $p = 31$ points with three different sample sizes, $N = 10, 20$ and 30. Each permutation test is performed with $M = 1000$ and is repeated for 5000 samples, so that we can evaluate the power of the test for different values of sample size and different degrees of violation of the null hypothesis.

Figure 1 shows the estimated power for different values of γ and N, where $\gamma = 0$ corresponds to the empirical size. Here the square root and Procrustes tests are the most powerful for most situations, and both tests have the correct empirical size, as do all tests with large enough sample sizes. However, some tests suffer from slightly inflated size when only small N is available. The test proposed by Panaretos et al. (2010) is slightly more powerful for large values of N.

For the second simulation, curves are generated by sampling the coefficients of a Lagrangian basis on $p = 31$ equispaced points on $[0, 1]$ from a multivariate t distribution with four degrees of freedom. We use the same values of N, M and γ as in the previous case. The empirical power curves for this second setting are also plotted in Fig. 1. In this scenario, the test based on the statistic $T_N(K)$ of Panaretos et al. (2010) has severely inflated empirical size, as would be expected from the violation of the Gaussian hypothesis. The test of Fremdt et al. (2013) based on $\hat{T}_1$ has correct empirical size, at least for large N, but it has a smaller empirical power compared with the permutation test based on Procrustes or square root distance, except for small values of γ. The results of additional simulations involving other scenarios for changes in the covariance structure are reported in the Supplementary Material.

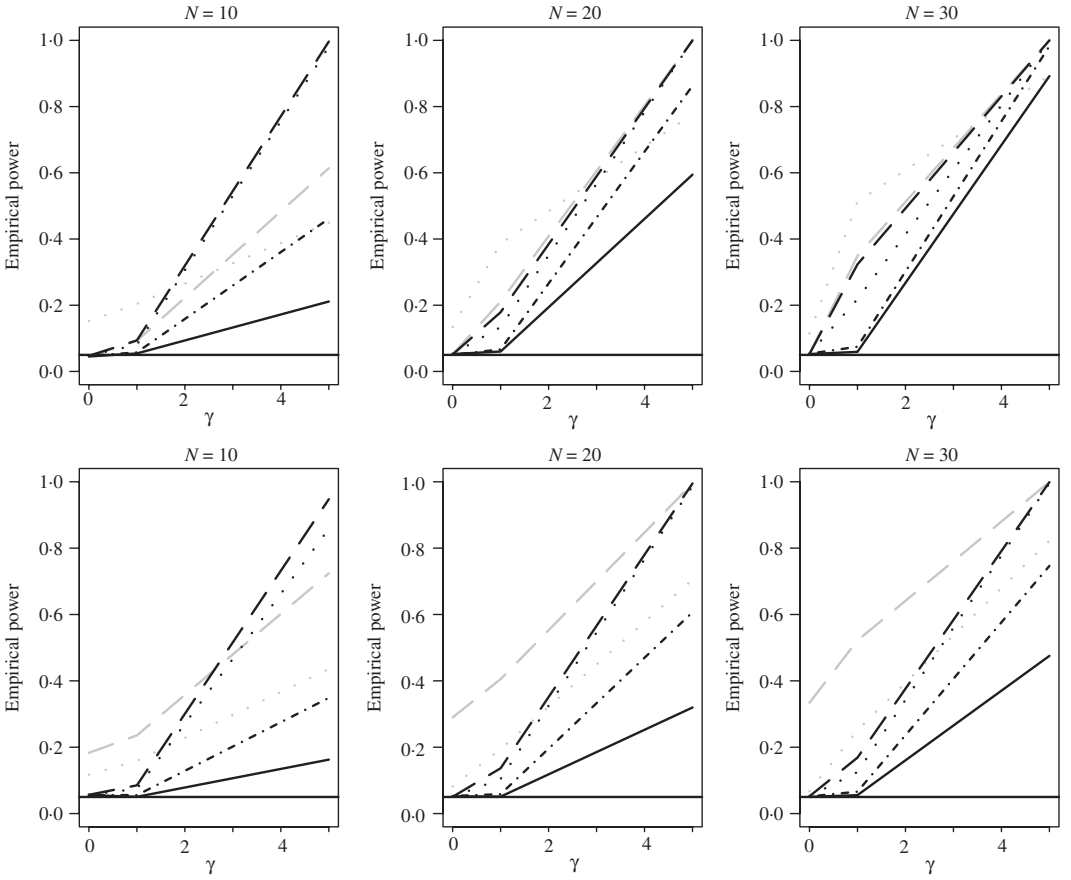

Fig. 1. Power estimated by simulation from a Gaussian process (first row) and a multivariate t distribution (second row) for different values of γ and N obtained with the parametric tests proposed by Panaretos et al. (2010) (long dashed grey line) and by Fremdt et al. (2013) (dotted grey line), along with the estimated power of the permutation test based on Procrustes size-and-shape distance (dotted black line), square root distance (long dashed black line), kernel distance (solid black line) and spectral distance (dot-dashed black line). The horizontal line shows the significance level $\alpha = 0{\cdot}05$.

4. Distance-based point estimation for covariance operators

4·1. *Fréchet averaging with square root and Procrustes distances*

We address here the problem of averaging of covariance operators. Let $S_1, \ldots, S_g$ be a sample of independent covariance operators. The most straightforward way to obtain an average operator is to calculate $\hat{\Sigma} = g^{-1}(S_1 + \cdots + S_g)$. This formula arises from the minimization of deviations measured with the Hilbert–Schmidt distance, i.e., the kernel distance. If we choose a different distance with which to compare covariance operators, it would make more sense to average covariance operators with respect to that chosen distance.

The population Fréchet mean (Fréchet, 1948) of a random element S with probability distribution μ on the space of covariance operators can be defined as $\Sigma = \arg\inf_P \int d(S, P)^2 \mu(\mathrm{d}S)$. If a sample $S_1, \ldots, S_g$ from μ is available, a least-squares estimator for Σ can be defined using the sample Fréchet mean as $\hat{\Sigma} = \arg\inf_S \sum_{i=1}^{g} d(S, S_i)^2$. Investigation of the properties of these kinds of estimators has been an active research field since the seminal work of Ziezold (1977). In the case of covariance operators, the consistency of the sample Fréchet mean is guaranteed, for the distances considered in this paper, if there exists Σ such that $E\{d(S_i, \Sigma)^2\} < +\infty$; see Huckemann (2011, Appendix A) for proofs and discussion.

In general the Fréchet mean is not unique, and therefore consistency is understood as convergence to an element of the Fréchet mean set; see Ziezold (1977), Le (1995), Le (2001) and Bhattacharya & Patrangenaru (2003) for more details.

The computation of the sample Fréchet mean $\hat{\Sigma}$ depends on the choice of the distance $d(\cdot, \cdot)$. Moreover, in practice covariance operators are often obtained from an estimation procedure applied to curve observations, and thus the sample Fréchet mean may need to be weighted to take into account the number of curves used in the estimate. Let $S_1, \ldots, S_g$ be the sample covariance operators of g different groups, with n_i curve observations each. Then the average covariance operator can be estimated as $\hat{\Sigma} = \arg\inf_S \sum_{i=1}^{g} n_i d(S, S_i)^2$. In general, this is a high-dimensional minimization problem, but some distances admit an analytical solution while for others efficient minimization algorithms are available. Note that $\hat{\Sigma}$ may not be unique for positively curved spaces, although it is unique for suitably concentrated data (Kendall, 1990; Le, 1995, 2001). It can be seen that for the square root distance d_S,

$$\hat{\Sigma} = \arg\min_S \sum_{i=1}^{g} n_i d_S(S, S_i)^2 = \left\{ G^{-1} \sum_{i=1}^{g} n_i (S_i)^{1/2} \right\}^2$$

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

For the Procrustes reflection size-and-shape distance, an analytical solution is not available. However, the Procrustes mean can be obtained by an adaptation of the algorithm proposed in Gower (1975). A description of the algorithm can be found in the Supplementary Material, as well as a simulation study comparing the behaviour of the standard average and Fréchet mean with both the square root and the Procrustes metrics. This simulation study shows that using a Fréchet mean can be beneficial in covariance estimation.

4·2. *Distance-based estimation*

Above, we chose the covariance operators as the starting point of the analysis, estimating the integral kernels of these operators with the traditional sample covariance function. However, a different approach is possible that uses the proposed distances in the estimation of the covariance operators. Specifically, an estimate of the common covariance operator for curve samples coming from g different groups could be

$$\hat{\Sigma} = \arg\min_S \sum_{i=1}^{g} \sum_{k=1}^{n_i} d\{S, (f_{ik} - \bar{f}_{i\cdot}) \otimes (f_{ik} - \bar{f}_{i\cdot})\}^2, \tag{3}$$

where $i = 1, \ldots, g$ are different groups, with $k = 1, \ldots, n_i$ curves in each group, $\otimes$ indicates the tensor product, i.e., $(f \otimes f)v = \langle f, v \rangle f$, and $\bar{f}_{i\cdot}$ is the sample mean of the ith group. If we assume that the rank-one operators $(f_{ik} - \bar{f}_{i\cdot}) \otimes (f_{ik} - \bar{f}_{i\cdot})$ are generated from the same distribution with finite second moments for all the groups, the result in Huckemann (2011) guarantees that (3) is a consistent estimator for the Fréchet mean of this distribution when $n_i \to +\infty$.

As in the case of the sample Fréchet mean, if we choose the square root distance, we get an explicit solution for problem (3) and

$$\hat{\Sigma} = \left[\left(\sum_{i=1}^{g} n_i \right)^{-1} \sum_{i=1}^{g} \sum_{k=1}^{n_i} \{(f_{ik} - \bar{f}_{i\cdot}) \otimes (f_{ik} - \bar{f}_{i\cdot})\}^{1/2} \right]^2.$$

For the Procrustes distance, problem (3) can be solved with a slight modification of the algorithm presented in the Supplementary Material. This estimator can be used, for instance, in the permutation test proposed in § 3, where the covariance operator for each group can be estimated by minimizing the same distance chosen for the test statistic.

4·3. *Interpolation and extrapolation*

Interpolation and extrapolation in nonlinear spaces has been studied in depth for the case of positive-definite matrices, for which it has been shown that simply using the Euclidean metric can be very problematic (see, e.g., Arsigny et al., 2006; Pennec et al., 2006; Dryden et al., 2009). In the infinite-dimensional case, the equivalent of a Euclidean approach would be extrapolation based on kernels. Let S_1 and S_2 be two covariance operators and $s_1(t', t)$ and $s_2(t', t)$ their integral kernels. We can obtain a path passing through these kernels as

$$s(t', t)(x) = [s_1(t', t) + x\{s_2(t', t) - s_1(t', t)\}], \quad x \in \mathbb{R}.$$

However, just as in the case of positive-definite matrices, extrapolation based on kernel distances does not always result in a valid kernel for a covariance operator; that is, the associated integral operator may not be nonnegative definite. An invalid operator could be made valid through a projection onto the space of covariance kernels, but how to choose this projection, given the possible projections available, is not immediately clear.

The square root metric and the Procrustes metric can each be associated with a geodesic which connects the two covariance operators S_1 and S_2, i.e.,

$$S_{\mathrm{R}}(x) = \left\{S_1^{1/2} + x\left(S_2^{1/2} - S_1^{1/2}\right)\right\}^* \left\{S_1^{1/2} + x\left(S_2^{1/2} - S_1^{1/2}\right)\right\}$$

and

$$S_{\mathrm{P}}(x) = \left\{S_1^{1/2} + x\left(S_2^{1/2}\tilde{R} - S_1^{1/2}\right)\right\} \left\{S_1^{1/2} + x\left(S_2^{1/2}\tilde{R} - S_1^{1/2}\right)\right\}^*, \tag{4}$$

respectively, where $x \in \mathbb{R}$ and $\tilde{R}$ is an unitary operator that minimizes $\|S_1^{1/2} - S_2^{1/2}R\|_{\mathrm{HS}}^2$. In general, analogously to the finite-dimensional case, the operator $\tilde{R}$ may not be uniquely defined if the sequence of singular values of $S_2^{1/2}S_1^{1/2}$ is degenerate (Kent & Mardia, 2001), i.e., if in our case there is more than one zero singular value. However, any choice of the operator $\tilde{R}$ provides a valid geodesic with respect to the Procrustes metric.

For every x, both the square root and the Procrustes geodesics give a valid covariance operator. However, in the case of extrapolation with the square root geodesic, this operator can be the result of the inverse square operation from the space of Hilbert–Schmidt operators to the space of covariance operators. Extrapolation in the space of Hilbert–Schmidt operators may lead to large negative eigenvalues, which result in large positive eigenvalues in the space of covariance operators. These are in general difficult to interpret, as they are an artificial effect of the choice of the square root geodesic. This effect can be avoided by using the Procrustes geodesic. We illustrate this with an artificial example involving covariance operators for boys and girls in the Berkeley growth curves study. Let S_1 and S_2 be the sample covariance operators for males and females, respectively. Figure 2 shows the minimum eigenvalue for the Hilbert–Schmidt operators $S_1^{1/2} + x(S_2^{1/2} - S_1^{1/2})$ and $S_1^{1/2} + x(S_2^{1/2}\tilde{R} - S_1^{1/2})$, for $x \in (0, 4)$. The former continuously decreases, while the latter correctly stabilizes at zero. Applying the backward map to the space of covariance operators, the square root geodesic thus yields, as an artefact, a very large positive eigenvalue, while the Procrustes geodesic does not suffer from this problem.

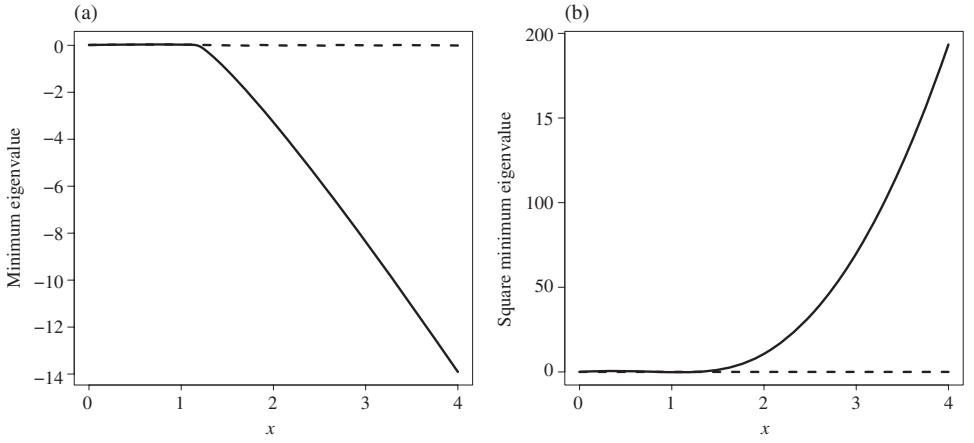

Fig. 2. (a) Minimum eigenvalue for the extrapolated Hilbert–Schmidt operator for the square root geodesic (solid line) and for the Procrustes geodesic (dashed line); (b) corresponding square eigenvalue for $S_R(x)$ (solid line) and for $S_P(x)$ (dashed line).

5. EXPLORING RELATIONSHIPS AMONG ROMANCE LANGUAGES

In this section we give an example of how the analysis of second-order structure can provide insights into a linguistic problem. The traditional way of exploring relationships across languages consists of examining textual similarity. However, this neglects phonetic characteristics of the languages. Here we propose a novel approach that involves comparing languages using their phonetic structure. This approach can add valuable linguistic information, particularly when combined with textual comparison and historical and geographical information in the linguistic analysis.

The recordings of people speaking different Romance languages are registered using the procedure of Tang & Müller (2008), with words pronounced in each language. The output of the registration for each word and for each speaker consists of the intensity of the sound over time and frequencies.

The aim is to explore phonetic relationships between languages and to compare these with existing linguistic knowledge. While the temporal aspects of each individual word are important, here we will concentrate on frequencies. Previous studies (Aston et al., 2010; Hadjipantelis et al., 2012) have indicated that covariance operators characterize languages well. The operators summarize phonetic information about the language, while disregarding characteristics of singular speakers and words. For the scope of this work, we focus on the covariance operators among frequencies in the log-spectrogram, estimated from all speakers of the language in the dataset. The spectrogram is a two-dimensional time-frequency image which gives localized time and frequency information across a word. We consider different time-points as different groups with the same covariance operator among frequencies. This is a significant simplification of the rich structure in the data, but it can generate some interesting conclusions. The main idea is that the distance between the covariance operators of frequency intensities is a good indication of the phonetic difference between languages.

Let $f_{ijk}(t) \in L^2(\Omega)$ be a realization of a random process, where $i = 1, \ldots, L$ represent different languages, $j = 1, \ldots, n$ the groups, i.e., different time-points, and $k = 1, \ldots, K_i$ the observations, i.e., individual speakers. As mentioned above, it is expected that the significant information from the different languages lies in the language-wise covariances S_i rather than in the individual observations f_{ijk}. Here results are reported for the covariance operator for the

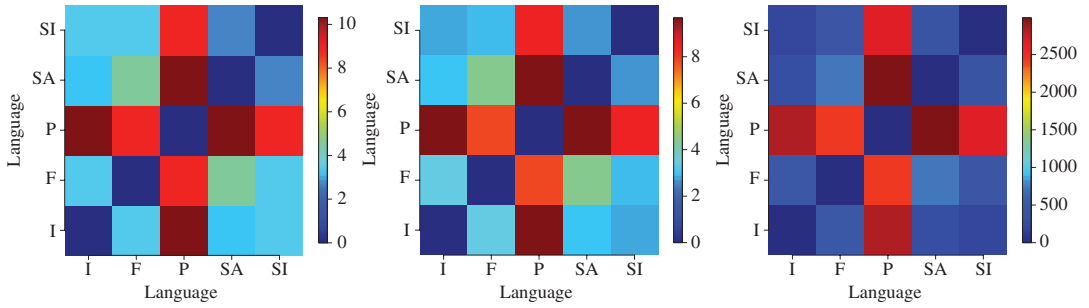

Fig. 3. Distance matrix among Fréchet estimates, obtained with the square root distance (left), Procrustes distance (centre) or kernel distance (right), where I = Italian, F = French, P = Portuguese, SA = American Spanish, and SI = Iberian Spanish.

word 'one' spoken in the different languages by a total of 23 speakers across five languages: French, Italian, Portuguese, Iberian Spanish and American Spanish. This word is similar in each language, coming from a common Latin root, but different enough to highlight changes across the languages. We estimate the covariance structure among frequencies for each language i using the estimator (3). Visual representations of the estimated covariance operator for each language are given in Fig. 4 and in the Supplementary Material. Figure 3 shows the dissimilarity matrix among estimated covariance operators, using the square root distance, Procrustes distance and kernel distance. Relationships among the covariance operators have features which are expected from linguistic hypotheses, such as strong similarity between the two varieties of Spanish, and with Italian, which are correctly found using both the square root distance and the Procrustes distance. These two distances yield essentially the same distance structure among languages, while the kernel distance matrix is slightly different and disagrees somewhat with existing linguistic knowledge. However, not all our conclusions directly support textual analysis, and thus they provide complementary information. The distance of Portuguese from both Spanish languages is greater than expected. Moreover, for historical reasons, American Spanish is expected to be closer than Iberian Spanish to Italian, but the covariance structures indicate that this is reversed.

A particularly interesting objective of the analysis is to provide insight into the change of the frequency structure along the path of language evolution. This would be inherently linked to extrapolation based on the distances we have proposed. Here, we want to compare the frequency covariance structure of a language to the structures obtained by extrapolating covariance operators of previous languages in the evolutionary path, as it is supposed from historical and geographical considerations. The Portuguese language presents a very different covariance structure from the other Romance languages, as can be seen in Fig. 4 as well as figures in the Supplementary Material. Thus, it would be of interest to compare its frequency covariance operator with the one extrapolated from the covariances of the two Spanish languages, to see if this kind of covariance was expected and whether a linear model of distance is appropriate. In the opposite direction of the evolutionary path, we also compare the Italian frequency covariance operator with that extrapolated from the two Spanish varieties.

As starting estimates, we can use those provided by any one of the three distances considered above. Here we show only results based on the kernel estimates, i.e., the classical pooled covariance functions, to highlight the effect of the different extrapolation strategies using different distances. Other choices, including ones that use estimates consistent with each extrapolation procedure, can be found in the Supplementary Material.

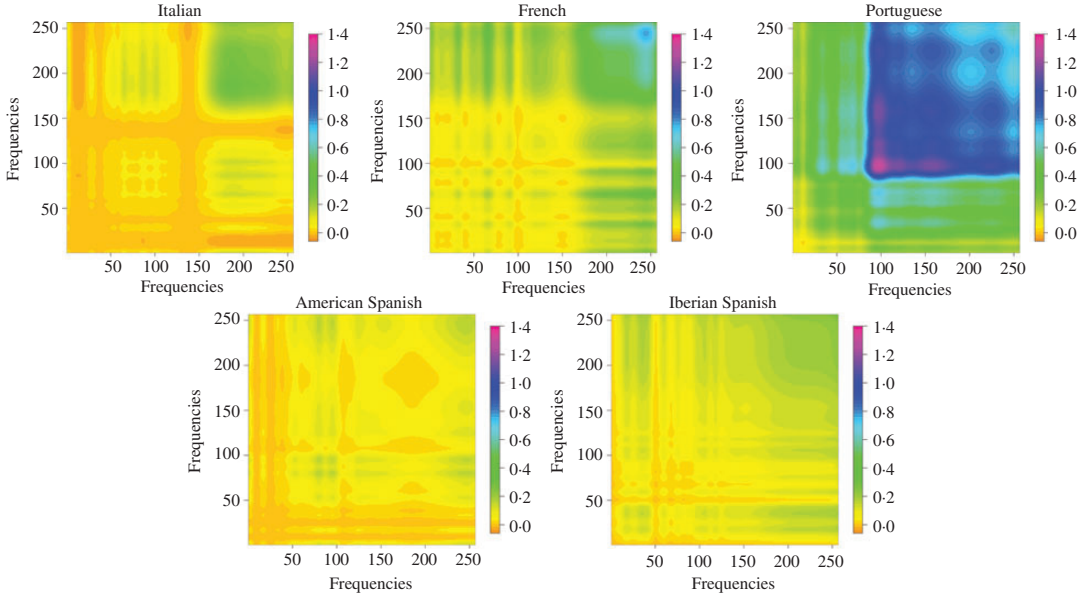

Fig. 4. Fréchet estimates of the covariance operators of the log-spectrogram among frequencies for five Romance languages, using the Procrustes distance.

The extrapolated covariance operator for Portuguese, P, is obtained with the method proposed in § 4·3, by evaluating the extrapolation line from the Iberian Spanish, SI, operator to the American Spanish, SA, operator, at $x = d(S_{SA}, S_P)/d(S_{SA}, S_{SI})$. For Italian, I, we evaluate the line from the American Spanish operator to the Iberian Spanish operator, at $x = d(S_{SI}, S_I)/d(S_{SA}, S_{SI})$.

Table 1 shows the comparison between the extrapolation starting from kernel-based estimates of the covariance operators of the two Spanish varieties and the kernel-based estimates of Portuguese and Italian. The comparison is performed with the square root distance, Procrustes distance and kernel distance. Simple extrapolation of integral kernels does not provide valid covariance operators for either Portuguese or Italian. Here the Procrustes method is shown to be better for both Portuguese and Italian, although the advantage is far greater for the former. This is to be expected, since the extrapolation for the Italian covariance operator is a short-distance extrapolation, and as such the square root mapping does not introduce large artificial effects. A truncation of negative eigenvalues to zero in the square root space, which might be thought to alleviate issues with artificial effects, performs even worse than the original square root extrapolation. Thus, extrapolation based on the Procrustes geodesic is to be preferred, as expected from the theory and seen empirically in § 4·3. Figure 5 shows results obtained with the Procrustes geodesic. The results obtained with the square root distance are given in the Supplementary Material. We can conclude that some features of the most extreme language in the family can be expected, such as higher variability in the high frequencies. Yet, unexpected features are also present, for example a higher variability in the mid-range frequencies for Portuguese; these are worthy of deeper linguistic exploration, especially using a larger corpus.

In conclusion, we have shown that the covariance structure between frequency intensities contains valuable linguistic information. Of course, this is not enough for drawing general linguistic conclusions, but we offer it as an additional tool to be used alongside existing linguistic sources of information, such as textual comparison and historical and geographical information.

Table 1. *Comparison between the kernel-based estimates of the Portuguese and Italian covariance operators and the geodesic extrapolation from kernel-based estimates of the two Spanish varieties*

	Square root geodesic extrapolation	Procrustes geodesic extrapolation	Kernel extrapolation
		Portuguese	
Square root distance	14·55	13·43	NaN
Procrustes distance	13·84	12·78	NaN
Kernel distance	3107·70	2524·99	5372·54
		Italian	
Square root distance	6·28	6·23	NaN
Procrustes distance	5·76	5·70	NaN
Kernel distance	329·88	258·45	232·99

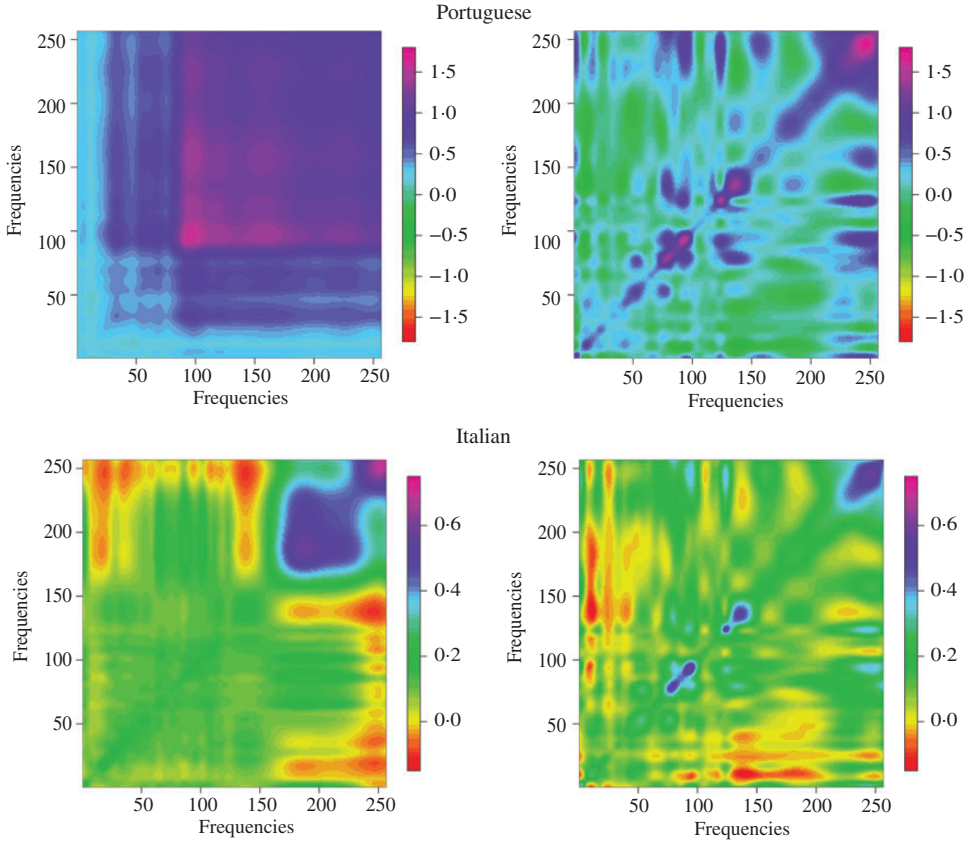

Fig. 5. Upper panels show the kernel estimate for Portuguese (left) and the corresponding Procrustes extrapolation from the two Spanish varieties using equation (4) (right). Lower panels show the kernel estimate for Italian (left) and the corresponding Procrustes extrapolation from the two Spanish varieties using equation (4) (right).

Acknowledgement

We are very grateful to the associate editor and three referees whose comments have helped to considerably strengthen the paper. We are also very much indebted to Prof. John Coleman for providing the Romance language dataset and to Pantelis Hadjipantelis for pre-processing it. This

research was partially supported by the Engineering and Physical Sciences Research Council and a Royal Society Wolfson Research Merit Award.

REFERENCES

ARSIGNY, V., FILLARD, P., PENNEC, X. & AYACHE, N. (2006). Log-Euclidean metrics for fast and simple calculus on diffusion tensors. *Magn. Resonance Med.* **56**, 411–21.

ASTON, J. A. D., CHIOU, J.-M. & EVANS, J. P. (2010). Linguistic pitch analysis using functional principal component mixed effect models. *Appl. Statist.* **59**, 297–317.

BHATTACHARYA, R. & PATRANGENARU, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds. I. *Ann. Statist.* **31**, 1–29.

BOSQ, D. (2000). *Linear Processes in Function Spaces*. New York: Springer.

DRYDEN, I. L., KOLOYDENKO, A. & ZHOU, D. (2009). Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *Ann. Appl. Statist.* **3**, 1102–23.

EL KAROUI, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36**, 2717–56.

FERRATY, F. & VIEU, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Berlin: Springer.

FRÉCHET, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. Henri Poincaré* **10**, 215–310.

FREMDT, S., STEINEBACH, J. G., HORVÁTH, L. & KOKOSZKA, P. (2013). Testing the equality of covariance operators in functional samples. *Scand. J. Statist.* **40**, 138–52.

GOOD, P. I. (2005). *Permutation, Parametric and Bootstrap Tests of Hypotheses*, 3rd edition. New York: Springer.

GOWER, J. C. (1975). Generalized Procrustes analysis. *Psychometrika* **40**, 33–50.

HADJIPANTELIS, P. Z., ASTON, J. A. D. & EVANS, J. P. (2012). Characterizing fundamental frequency in Mandarin: A functional principal component approach utilizing mixed effect models. *J. Acoust. Soc. Am.* **131**, 4651–64.

HORVÁTH, L. & KOKOSZKA, P. (2012). *Inference for Functional Data with Applications*. New York: Springer.

HUCKEMANN, S. F. (2011). Intrinsic inference on the mean geodesic of planar shapes and tree discrimination by leaf growth. *Ann. Statist.* **39**, 1098–124.

KENDALL, W. S. (1990). Probability, convexity, and harmonic maps with small image I: Uniqueness and fine existence. *Proc. Lond. Math. Soc.* **61**, 371–406.

KENT, J. T. & MARDIA, K. V. (2001) Shape, Procrustes tangent projections and bilateral symmetry. *Biometrika* **88**, 469–85.

LE, H. (1995). Mean size-and-shapes and mean shapes: A geometric point of view. *Adv. Appl. Prob.* **27**, 44–55.

LE, H. (2001). Locating Fréchet means with application to shape spaces. *Adv. Appl. Prob.* **33**, 324–38.

PANARETOS, V. M., KRAUS, D. & MADDOCKS, J. H. (2010). Second-order comparison of Gaussian random functions and the geometry of DNA minicircles. *J. Am. Statist. Assoc.* **105**, 670–82.

PENNEC, X., FILLARD, P. & AYACHE, N. (2006). A Riemannian framework for tensor computing. *Int. J. Comp. Vis.* **6**, 41–66.

RAMSAY, J. O. & SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd edition. New York: Springer.

TANG, R. & MÜLLER, H. G. (2008). Pairwise curve synchronization for high-dimensional data. *Biometrika*, **95**, 875–89.

ZHU, K. (2007). *Operator Theory in Function Spaces*, 2nd edition. Providence, Rhode Island: American Mathematical Society.

ZIEZOLD, H. (1977). On expected figures and a strong law of large numbers for random elements in quasi-metric spaces. In *Trans. 7th Prague Conf. Info. Theory, Statist. Decis. Functions, Random Proces. 8th Eur. Meeting of Statisticians*, vol. A, pp. 591–602. Dordrecht: Reidel.

[*Received July* 2012. *Revised December* 2013]