Topics in Functional Data Analysis

Habilitation Thesis

David Kraus

August 2021

Masaryk University
Faculty of Science
Department of Mathematics and Statistics

Contents

1. Introduction and summary

1.1. Introduction

Functional data analysis is an active area of statistics that deals with data that can be seen as mathematical functions. These could be curves, surfaces, images etc. Due to the development of modern technology, contemporary data sets indeed often consist of data units that are complex object. A functional data set is a collection of observations of such functions (mathematically regarded as realizations of random processes, i.e., random variables in a function space), whereas more traditional data sets consist of observations of numbers or vectors. For a general background, see, e.g., Bosq (2000), Ramsay and Silverman (2005), Ferraty and Vieu (2006), Ferraty and Romain (2011), Horváth and Kokoszka (2012), Hsing and Eubank (2015) or Kokoszka and Reimherr (2017).

My research concentrates on the development of statistical methodology driven by applications. This text comprises five research articles containing my and my co-authors' contributions to the field of functional data analysis accompanied by this introductory section, which summarizes the contents of the papers. The presentation is simplified to provide only the basic ideas and results of each paper. Thus, for example, references to preceding and subsequent relevant publications are not included and results are described in a stylized way rather than as rigorous formal statements.

The papers included in the appendix are:

(A) Panaretos, V. M., Kraus, D., and Maddocks, J. H. (2010). Second-order comparison of Gaussian random functions and the geometry of DNA minicircles. *Journal of the American Statistical Association*, 105(490):670–682.

(B) Kraus, D. and Panaretos, V. M. (2012). Dispersion operators and resistant second-order functional data analysis. *Biometrika*, 99(4):813–832.

(C) Kraus, D. (2015). Components and completion of partially observed functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(4):777–801.

(D) Kraus, D. and Stefanucci, M. (2019). Classification of functional fragments by regularized linear classifiers with domain selection. *Biometrika*, 106(1):161–180.

(E) Kraus, D. (2019). Inferential procedures for partially observed functional data. *Journal of Multivariate Analysis*, 173:583–603

Four papers (A, B, C, D) have been published in the Journal of the American Statistical Association, Biometrika and Journal of the Royal Statistical Society: Series B (Statistical Methodology), which are regarded by the scientific community among the leading 5–7 journals in the field of methodological statistics. Paper E has been published in the Journal of Multivariate Analysis, which is a standard, respected journal in the field. Two papers (C, E) are single-authored, the other three are collaborative with equal contribution of each co-author. The papers have been published with peer-reviewed supplements, which are included as well.

1.2. Summary of Paper A

Paper A (Panaretos et al., 2010) studies methods of statistical inference on the covariance structure of random functions. Although its main focus is the development of statistical methodology and related theory, the motivation for this work comes from another field, namely molecular biology.

The understanding of the mechanical properties of the DNA molecule constitutes a fundamental biophysical task, as important biological processes can be affected by properties such as stiffness and shape. In addition to holding the genetic code, the DNA base-pair sequence may influence the geometric properties of the molecule. However, empirical detection of this effect on stereological data acquired through the electron microscope had previously been elusive. The data set of interest consists of closed curves (DNA minicircles obtained from short strands of DNA) in $\mathbb{R}^3$ of two types: both types have identical base pair sequences, except for a short base-pair window, where two different sequences are present (one of them, a TATA box, is of special interest). Biophysical considerations suggest this will have a significant effect on the geometry of the minicircle, and the goal is to compare these two groups to probe for such an effect.

Motivated by the need of two-sample comparison of loops, as exemplified in DNA minicircle experiments, this article considers the problem of second-order comparison of two samples of random functions, within a functional data analysis framework. In particular, given realisations of n_1 and n_2 independent copies of two continuous zero mean Gaussian processes X and Y on a compact set, we consider the problem of testing the hypothesis that their covariance operators $\mathscr{R}_X, \mathscr{R}_Y$ are equal against the alternative that they are different. Although this problem is now well studied, by the time of writing of this paper it had received relatively little attention. Our paper proposes a test based on the approximation of the Hilbert–Schmidt distance of the empirical covariance operators of the two samples of functions based on the Karhunen–Loève expansion. The asymptotic distribution of the test statistic is determined and its performance is investigated computationally. The application of our methodology to the data set of two groups of minicirles characterized by the presence or absence of a TATA box suggests the potential existence of significant differences in the two groups, which eluded previous analyses as these focused on the mean (the shape of the minicircle), whereas we detect the differences in the covariance structure (the flexibility/stiffness).

Let us give a more detailed description of the contents of Paper A.

Since this work is data-driven, the paper first explains the scientific background and questions in molecular biology and the source, properties and pre-processing of available data. To perform a functional data analysis of the minicircles it is required to register the data. Each curve has thus been centered and scaled, so that the center of mass is at zero and the length of the curve is one. Since the data were obtained by electron microscopy of the minicircles imbedded in a liquid, the reconstructed curves are not aligned (they are subject to a random unobservable orthogonal transformation). We describe a procedure that rigidly aligns curves by their intrinsic characteristics: each curve was individually aligned using the coordinate system induced by its moments of inertia tensor. We thus arrive at a functional data set consisting of smooth curves indexed by the arc length taking values in $\mathbb{R}^3$ (corresponding to the coordinates on the three principal axes of inertia).

We assume that we have two independent collections $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ of iid Gaussian processes on $[0,1]$, considered as random elements of the Hilbert space $\mathcal{L}^2[0,1]$ of coordinate-wise square-integrable $\mathbb{R}^3$-valued functions with the inner product $\langle f, g\rangle = \int_0^1 f(t)^\mathsf{T} g(t) dt$ (but everything readily extends to more general cases). Assuming, without loss of generality, that the mean functions are zero, the processes are characterized by their respective covariance kernels $R_X(s,t) = \operatorname{cov}\{X_i(s), X_i(t)\}) = \mathsf{E}\{X_i(s)X_i(t)^\mathsf{T}\}$, and $R_Y(s,t)$, respectively. Associated with the covariance kernel is the covariance operator $\mathscr{R}_X\colon \mathcal{L}^2[0,1] \to \mathcal{L}^2[0,1]$ defined as $\mathscr{R}_X(f)(t) = \operatorname{cov}\{\langle X_i, f\rangle, X_i(t)\} = \int_0^1 R_X(t,s) f(s) ds$. The Karhunen–Loève theorem allows for a representation of the process by a stochastic Fourier series with respect to the orthonormal eigenfunctions $\{\varphi_X^{(j)}\}_{j=1}^\infty$ of the operator $\mathscr{R}_X$,

$$X_i(t) = \sum_{j=1}^{\infty} \sqrt{\lambda_X^{(j)}}\, \xi_{ij} \varphi_X^{(j)}(t),$$

where $\{\lambda_X^{(j)}\}_{j=1}^{\infty}$ is the nonincreasing sequence of corresponding eigenvalues and $\{\xi_{ij}\}$ is an iid array of standard Gaussian random variables. The empirical covariance kernel may be used to "optimally" reduce infinite-dimensional inferential problems to multivariate ones. Letting $\hat{R}_X$ stand for the empirical covariance kernel, we denote its eigenvalues by $\hat{\lambda}_X^{k,n_1}$ and its eigenfunctions by $\hat{\varphi}_X^{k,n_1}$. The finite-dimensional reduction is then achieved by retaining a finite number of principal components $\langle X_i - \bar{X}, \hat{\varphi}_X^{k,n_1}\rangle$, $k = 1, \ldots, K$ in lieu of each X_i, and similarly for the second sample. The dimension reduction afforded by the Karhunen–Loève expansion is the tool we employ to construct our test.

We wish to test the null hypothesis $\mathscr{R}_X = \mathscr{R}_Y$ against the alternative $\mathscr{R}_X \neq \mathscr{R}_Y$. We propose the use of a test statistic based on the norm of the difference of the two empirical covariance operators. The Hilbert–Schmidt norm of a trace-class operator $\mathscr{R}$ is defined as

$$\|\mathscr{R}\|_{\mathrm{HS}} = \sqrt{\int_0^1 \int_0^1 \mathrm{trace}\{R(s,t)^{\mathsf{T}} R(s,t)\} ds dt}.$$

A test may be based on the squared Hilbert–Schmidt distance $\|\hat{\mathscr{R}}_X - \hat{\mathscr{R}}_Y\|_{\mathrm{HS}}^2$. The sampling distribution of this quantity will depend on the unknown covariance operators even asymptotically. To be able to "normalize" the test statistic, we employ the property that for any orthonormal system $\{e_i\}$ of $\mathcal{L}^2[0,1]$, we have

$$\|\mathscr{R}\|_{\mathrm{HS}}^2 = \sum_{i=1}^{\infty} \|\mathscr{R} e_i\|^2 = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \langle \mathscr{R} e_i, e_j\rangle^2$$

$\|\hat{\mathscr{R}}_X - \hat{\mathscr{R}}_Y\|_{\mathrm{HS}}^2$. In practice, we need to truncate the series to obtain a finite-dimensional reduction and choose the contrasts $\{e_i\}$ so that the truncation retains the bulk of the norm. For each of the two empirical operators, the optimal contrasts will coincide with their eigenfunctions, but we need to use a common basis. We thus choose the eigenfunctions $\hat{\varphi}_{XY}^{k,N}$ of the empirical covariance operator of the pooled sample as a compromise for the common coordinate system. Our proposed test statistic is a linear combination of the terms $\langle (\hat{\mathscr{R}}_X - \hat{\mathscr{R}}_Y)\hat{\varphi}_{XY}^{i,N}, \hat{\varphi}_{XY}^{j,N}\rangle^2$, $i, j = 1, \ldots, K$ with weights corresponding to their asymptotic covariance structure. Theorem 1 in the paper shows that under the null hypothesis and certain assumptions, this test statistic is asymptotically chi-squared distributed with $K(K+1)/2$ degrees of freedom, which is the basis of a hypothesis test.

The paper then introduces a modified test statistic that can be useful when one a priori knows that the eigenfunctions of both covariance operators are are equal. Then one can focus only on the diagonal terms (those with $i = j$), which leads to a test statistic with asymptotic chi-squared distribution with K degrees of freedom. Furthermore, we consider variance-stabilized variants of these statistics, where we apply a log transformation to the diagonal terms and Fisher's z-transformation to the off-diagonal terms. We then discuss methods to choose the truncation level.

To assess the behaviour of the proposed tests under the null hypothesis and under various alternatives we carry out a number of simulations. We consider one situation with equal covariance functions and several alternative configurations. The general and

diagonal test statistics are considered under various fixed choices of K and with automatically selected K. The study provides a useful insight into the performance and capabilities of the tests depending on the type of deviation from the null hypothesis. Next, we present an analysis of the data set of DNA minicircles. First, we show both graphically and numerically that there is no important difference between the means of the two types of curves. Then we focus on the comparison of their second order properties. The analysis shows a significant difference on the third (most important) principal axis of inertia and also jointly in the plane given by the third and second axis.

A proof of Theorem 1 is provided in the Appendix. Additional plots and tables are available in a supplementary file. In addition, the supplementary file contains a more detailed study of the problem of comparing the complete spectrum.

1.3. Summary of Paper B

Paper B (Kraus and Panaretos, 2012) focuses on the second-order structure of a random function, which is key to understanding the nature of the functional observations that it induces, as it is inextricably linked with the smoothness properties of the stochastic fluctuations of the function. These second-order properties are encapsulated in the covariance operator. The link with the smoothness properties of the random function is then given by the Karhunen–Loève expansion, which provides an optimal Fourier representation of the random function, using a basis comprised by the eigenfunctions of this operator. A natural inference problem is that of comparing the covariance structures of two samples of functional data, in order to decide whether they share the same fluctuation properties. We focus on situations where the data are not Gaussian, and indeed may be characterized by the presence of influential observations. The infinite-dimensional nature of the data means that an observation can be atypical in many ways, the deviation from the mean being only one; observations close to the mean may contain unusual frequency components. Detection of such observations via exploratory techniques may be non-trivial. Such influential observations might significantly influence the estimation of the covariance, and, even more profoundly, the quality of the estimators of its spectrum. The sensitivity of the empirical covariance operator and its spectrum to the presence of influential observations can have an impact on testing procedures for the covariance operator.

To cope with these issues, this paper introduces a class of operators that we term dispersion operators that are implicitly defined through a variational problem, motivated by M-estimators of location for the tensor product of the centred functional observations. It is then proposed that these operators be used as proxies for the covariance operator, when inferences on the second-order structure are to be drawn for non-Gaussian and potentially contaminated functional samples. The implicit definition of a dispersion operator gives rise to a score equation, as the dispersion operator is a zero of the Fréchet derivative of the variational problem with respect to the operator argument. This functional score equation is then used as a basis to construct a test for the second-order comparison of two functional samples. The test is based on the distance of the functional score equation under the null hypothesis from zero, measured by an appropriately

renormalized Hilbert–Schmidt distance. This work is motivated by and illustrated on a data set of DNA strands, which indeed is contaminated by atypical curves.

We now recapitulate the contributions of Paper B in more detail.

First, the paper introduces the notion of a dispersion operator as a substitute for the usual covariance operator that is more suitable for contaminated data while still characterizing the second-order structure of the random function. To describe the second-order properties of a random element X in a separable Hilbert space $\mathcal{H}$ (without loss of generality $L^2[0,1]$), one typically considers the covariance operator

$$\mathscr{C} = \mathsf{E}\{(X-\mu)\otimes(X-\mu)\},$$

where $\otimes$ stands for the tensor product and $\mu = \mathsf{E}(X)$ is the mean. The covariance operator can be seen as the Hilbert–Schmidt operator that solves the variational problem

$$\min_{\mathscr{R}\in\mathrm{HS}(\mathcal{H},\mathcal{H})} \mathsf{E}\{\|(X-\mu)\otimes(X-\mu)-\mathscr{R}\|^2\}$$

($\mathrm{HS}(\mathcal{H},\mathcal{H})$ are Hilbert–Schmidt operators from $\mathcal{H}$ to $\mathcal{H}$). The empirical covariance operator can be represented as the solution to the above optimization problem with expectation computed with respect to the empirical distribution of the data. This being essentially a least squares problem, both the empirical covariance operator and methods based on it will be sensitive to the presence of atypical observations in the dataset. We obtain procedures pertaining to the second-order structure of X that are more resistant to departures from normality and to the presence of influential observations by replacing the squared norm in the variational problem defining the covariance by a less sensitive loss function. This gives rise to a new class of second-order characteristics, which we call dispersion operators. Within this class, the most useful new choice of the loss function leads to what we call the spatial dispersion operator. It is defined via M-estimation of the location of $(X-\mu)\otimes(X-\mu)$ as

$$\operatorname*{arg\,min}_{\mathscr{R}\in\mathrm{HS}(\mathcal{H},\mathcal{H})} \mathsf{E}\{\|(X-\mu)\otimes(X-\mu)-\mathscr{R}\| - \|(X-\mu)\otimes(X-\mu)\|\},$$

where μ is a suitable element of $\mathcal{H}$ with the interpretation of a location parameter (the spatial median is a natural choice). The empirical spatial dispersion operator minimizes the sample version of the objective. By taking Fréchet derivative we arrive at an equivalent definition of the dispersion operator as a Z-estimator solving a score equation.

Proposition 1 in the paper establishes the existence and uniqueness of the (population) dispersion operator under non-restrictive assumptions on the data-generating distribution. In Corollary 1 we show that the sample dispersion operator exists and is unique under weak assumptions on the observed data, and that it is consistent for the true dispersion operator. We continue our theoretical analysis by showing an interesting link between the spectra of the dispersion and covariance operator. Although the operators are in general different, they both carry useful information on second-order properties. Proposition 2 shows that the dispersion operator has the same set of eigenfunctions as the covariance operator.

Having defined the notion of a dispersion operator, we then construct a two-sample second-order test based upon it. Let there be two independent random samples of functions, whose location parameters are μ_1, μ_2 and dispersion operators are $\mathscr{R}_1, \mathscr{R}_2$. The goal is to test the null hypothesis H_0: $\mathscr{R}_1 = \mathscr{R}_2$ against the general alternative H_1: $\mathscr{R}_1 \neq \mathscr{R}_2$. We propose to employ the general idea of score tests, that is, to base the test on the estimating score for the general model, without assuming H_0, evaluated at the null estimate of the parameter. As the centres μ_1, μ_2 are not restricted under the null hypothesis, they can be estimated separately. On the other hand, the common null estimator of the dispersion is estimated by $\hat{\mathscr{R}}$, which minimizes a combination of objectives for each sample under the restriction induced by the null. Equivalently, $\hat{\mathscr{R}}$ solves a score equation under H_0. After a reparametrization, we arrive at a score operator whose component corresponding to the difference between the two dispersion operators reflects the validity of H_0. When the null hypothesis holds, the score operator is expected to be close to the zero operator, otherwise it should be far from the zero operator. To perform the test, we need to measure its distance from the zero operator and assess the significance of the resulting test statistic. We especially develop one way of doing it. It is based on spectral truncation of the score operator, which is an infinite dimensional object (a Hilbert–Schmidt operator on $\mathcal{H}$). We use a projection of this operator on a finite dimensional subspace, in particular the one defined by the tensor products of the eigenfunctions of the dispersion operator. The test statistic is then obtained by combining the projection coefficients in a quadratic form. Theorem 1 establishes the weak convergence of the score operator to a mean zero Gaussian random operator under the null hypothesis and provides a consistent estimator of its covariance operator (which is an operator on operators). Then it provides the asymptotic null distribution of the score test statistic.

Next, the paper presents empirical results. In a simulation study, we investigate the behaviour of the test based on the spatial dispersion and the non-resistant L^2 test under the null hypothesis without and with contamination and the impact of contamination on the power of these tests under various alternative and contamination scenarios. We also apply the proposed methodology to the data set of DNA minicircles studied in Paper A. The supplementary document contains proofs of theoretical results.

1.4. Summary of Paper C

It is standard in the field of functional data analysis to assume that all functions are observed on the same domain. In Paper C (Kraus, 2015), we develop methods of analysis for functional data that are observed incompletely in the sense that each function might be observed only on a subset of the domain, whereas no information about the curve is available on the complement of this subset.

Our work is motivated by an ambulatory blood pressure monitoring data set that is part of the "Swiss Kidney Project on Genes in Hypertension." The data set consists of automatically recorded temporal heart rate profiles of several hundred participants. Due to either the failure of the recording device or participant's discomfort some values have not been measured and the time points corresponding to unobserved values form

series (intervals) of non-negligible length. The resulting data set thus consists of partially observed curves (functional fragments). Since there is only a relatively small fraction of complete curves, removing incomplete curves would considerably reduce (and possibly destroy) the accuracy of the statistical analysis. Therefore, this type of functional data necessitates the development of special methodological approaches, which is the subject of this paper. Before the appearance of this paper, relatively little work had been published on missing data in the functional context.

In this paper we introduce a formal framework for analysing incompletely observed functional data and develop basic nonparametric, fully functional (infinite-dimensional) inferential procedures. We first focus on the main building blocks of the analysis of the second-order properties: estimation of the covariance operator and principal component analysis. We propose an estimator of the covariance operator and its eigenvalues and eigenfunctions for partially observed functions and derive their properties. We deal with the estimation of projections (principal scores) of individual incomplete functions which is especially challenging. We develop a procedure that enables to predict the value of a principal score of a function when only a fragment of the function is available and direct computation is thus impossible. Next, we propose a method that can recover the unobserved part of the function from the observed part, using the information about the distribution of the data that it learns from the sample. We develop automatic procedures for the selection of the tuning parameter of the method that is based on generalised cross-validation for incompletely observed functions. We quantify the uncertainty of the predictions of unobserved quantities and provide approximate prediction regions (intervals and bands) covering the unobserved random quantity with high probability. Simulations confirm the usefulness and good performance of the proposed methodology.

We now describe the main methodological, theoretical and numerical contributions of Paper C.

First, the paper formalizes the framework of partially observed functional data. Functional data $X_1, \ldots, X_n$ are seen as independent identically distributed random variables in the separable Hilbert space of square integrable functions on a bounded domain. Without loss of generality, we consider the space $L^2([0,1])$ with inner product $\langle f, g\rangle = \int_0^1 f(t)g(t)dt$, $f, g \in L^2([0,1])$ and norm $\|f\| = \langle f, f\rangle^{1/2}$. In traditional functional data analysis, it is assumed that the functions $X_1, \ldots, X_n$ are observed on the whole interval $[0,1]$. We consider situations where each curve X_i is observed only on a subset of $[0,1]$. Specifically, let the observation periods be $O_i \subset [0,1]$, $i = 1, \ldots, n$. Then the observed data for the ith curve are $X_i(t)$, $t \in O_i$. We collectively denote the observed part of the curve as X_{iO_i}, which can be seen as a random element of the space $L^2(O_i)$. The values of X_i on the complement of O_i, $M_i = [0,1] \setminus O_i$, are not observed; the missing part of the trajectory is denoted as X_{iM_i}. The observation periods O_i, $i = 1, \ldots, n$ are modelled as random subsets of $[0,1]$. We assume that the observation periods are independent of the functions $X_1, \ldots, X_n$, that is, the data are missing completely at random.

Next, the paper focuses on the estimation of the main characteristics of the distribution that generates the data, that is, the mean function and the covariance operator. Let the

mean function be $\mu = \mathsf{E}\, X_1$. The covariance operator $\mathscr{R} : L^2([0,1]) \to L^2([0,1])$ is defined as $\mathscr{R}f = \mathsf{E}\{\langle f, X_1 - \mu\rangle(X_1 - \mu)\} = \int_0^1 \rho(\cdot, t) f(t) dt$, where $\rho(s,t) = \operatorname{cov}\{X_1(s), X_1(t)\}$ is the covariance kernel of the stochastic process X_1. Like in the multivariate case, the mean function μ at point $t \in [0,1]$ can be estimated by the sample mean of observed values at this point. The estimator $\hat{\mathscr{R}}$ of the covariance operator $\mathscr{R}$ is defined through an estimator of its covariance kernel ρ. We estimate $\rho(s,t)$ by the sample covariance computed from all complete pairs of functional values at s and t. It is seen that $\hat{\mu}(t)$ is an unbiased estimator of $\mu(t)$. Similarly, if we subtract 1 in the denominator of $\hat{\rho}(s,t)$, the estimator becomes unbiased for $\rho(s,t)$. For the estimators $\hat{\mu}$ and $\hat{\mathscr{R}}$ to be consistent, we need to assume that the observation pattern asymptotically provides enough information. The exact formulation of such assumptions is provided in equations (2) and (3) in the paper. Under these weak assumptions, we obtain a consistency result in Proposition 1 of the paper. In particular, we show that the L^2 distance between the $\hat{\mu}$ and μ and the Hilbert–Schmidt distance between $\hat{\mathscr{R}}$ and $\hat{\mathscr{R}}$ converge to zero in quadratic mean (and hence in probability). Interestingly, the properties of the estimators are unaffected by the fact that the functions are observed only partially. The full (dense) observation regime, albeit only on subsets of the domain, preserves the convergence rates known for complete functional data.

The paper then focuses on principal component analysis, which is probably the most fundamental method for functional data since it provides insight into the complex covariance structure of functional data, can be used to identify main sources of variability and quantify their importance and to reduce the dimension of the data. The theoretical foundation of functional principal component analysis is the Karhunen–Loève theorem stating that there exist random variables β_{ij} and nonrandom functions φ_j such that the stochastic process X_i admits the decomposition

$$X_i(t) = \mu(t) + \sum_{j=1}^{\infty} \beta_{ij} \varphi_j(t), \quad t \in [0,1],$$

where the series converges in mean square, uniformly in t. Here φ_j, $j = 1, 2, \ldots$ are the orthonormal eigenfunctions of the operator $\mathscr{R}$ and β_{ij}, $j = 1, 2, \ldots$ are uncorrelated mean zero variables with variances λ_j, where $\lambda_1 \geq \lambda_2 \geq \cdots > 0$ are the eigenvalues of $\mathscr{R}$. Functional principal component analysis is the empirical version of the Karhunen–Loève expansion that aims to estimate the elements involved in the expansion. In the case of completely observed functional data, to estimate the eigenvalues λ_j and eigenfunctions φ_j, one performs eigen-decomposition of the usual sample covariance operator. When the functions are observed partially, one can proceed similarly and define the estimators $\hat{\lambda}_j$ and $\hat{\varphi}_j$ as the eigenvalues and eigenfunctions of the operator $\hat{\mathscr{R}}$ given by the kernel $\hat{\rho}$. The paper shows that the asymptotic properties of the empirical eigenvalues and eigenfunctions remain unchanged by the incompleteness of the observed functions. Proposition 2 in the paper establishes that first, the empirical eigenvalues are consistent estimators of the true eigenvalues and this consistency is uniform over all indices, and second, the empirical eigenfunctions are consistent estimators of the true eigenfunctions, up to the usual sign ambiguity. The rates of convergence are parametric due to the full

observation regime on subsets.

The paper then moves to the most challenging contributions which are methods of inference for individual curves based on their incomplete observation. These are prediction rather than estimation problems since they aim to provide information on random targets: the principal scores β_{ij} and the missing part of the curve X_{iM_i}. In the standard situation of complete functional data, the scores are easily estimated by $\hat{\beta}_{ij} = \langle X_i - \hat{\mu}, \hat{\varphi}_j \rangle$. When the functional observations are incomplete, the direct computation of $\langle X_i - \hat{\mu}, \hat{\varphi}_j \rangle$ is impossible because the last term in the expression

$$\langle X_i - \hat{\mu}, \hat{\varphi}_j \rangle = \langle X_{iO_i} - \hat{\mu}_{O_i}, \hat{\varphi}_{jO_i} \rangle + \langle X_{iM_i} - \hat{\mu}_{M_i}, \hat{\varphi}_{jM_i} \rangle$$

is not available. In Section 3.2 of the paper we develop a procedure to estimate (or rather predict) the missing quantity $\langle X_{iM_i} - \hat{\mu}_{M_i}, \hat{\varphi}_{jM_i} \rangle$ from the observed data and establish its theoretical properties (Theorem 1, Proposition 3). We skip the description of this part in this summary and instead describe the results on prediction of the missing part of an incomplete curve.

This task of function reconstruction (completion) is studied in Section 4 of the paper. In the population version of the problem, the best prediction of X_M by a function of X_O in the sense of the mean integrated prediction squared error is the conditional expectation $\mathsf{E}(X_M|X_O)$. It is in general a nonlinear operator from $L^2(O)$ to $L^2(M)$ and similarly to the case of principal scores, we consider its best continuous linear approximation. Assuming for simplicity that the functional variable has mean zero, the minimisation problem to be solved is

$$\min_{\mathscr{A}:\|\mathscr{A}\|_\infty<\infty} \mathsf{E}\,\|X_M - \mathscr{A}X_O\|^2,$$

where the solution is looked for in the class of continuous (bounded) linear operators from $L^2(O)$ to $L^2(M)$ (by $\|\cdot\|_\infty$ we denote the operator norm). We see (by Fréchet differentiation or direct computation) that solving this minimisation is equivalent to solving the (normal) equation $\mathscr{A}\mathscr{R}_{OO} = \mathscr{R}_{MO}$. This suggests the solution $\tilde{\mathscr{A}} = \mathscr{R}_{MO}\mathscr{R}_{OO}^{-1}$ and the best linear prediction of X_M in the form $\tilde{X}_M = \tilde{\mathscr{A}}X_O$. From now on, we assume the existence of a bounded solution, that is, we assume that $\|\mathscr{R}_{MO}\mathscr{R}_{OO}^{-1}\|_\infty < \infty$. Similarly to the case of principal scores, the inverse problem $\mathscr{A}\mathscr{R}_{OO} = \mathscr{R}_{MO}$ to be solved is ill-posed, that is, small perturbations of the right-hand side $\mathscr{R}_{MO}$ can lead to large perturbations of the solution (recall that $\mathscr{R}_{OO}$ is compact, hence its inverse is unbounded); perturbations of the right-hand side indeed need to be considered since $\mathscr{R}_{MO}$ will be only estimated from the data in the sample version of the problem. Regularization (i.e., modification of an ill-posed inverse problem into a well-posed inverse problem) is necessary for a stable solution. Using ridge regularisation we obtain the solution $\tilde{\mathscr{A}^{(\alpha)}} = \mathscr{R}_{MO}(\mathscr{R}_{OO} + \alpha\mathscr{I}_O)^{-1}$ ($\alpha > 0$ is a regularization parameter, $\mathscr{I}_O$ is the identity operator of $L^2(O)$). The regularised best linear prediction equals $\tilde{X}_M^{(\alpha)} = \tilde{\mathscr{A}^{(\alpha)}}X_O$. Practically, when the sample $X_{1O_1}, \ldots, X_{nO_n}$ is observed on the subsets $O_1, \ldots, O_n$, we replace the covariance operator by its estimate and set $\hat{\mathscr{A}}_i^{(\alpha)} = \hat{\mathscr{R}}_{M_iO_i}\hat{\mathscr{R}}_{O_iO_i}^{(\alpha)-1}$. The mean function needs to be estimated as well. For the ith curve, the best linear prediction of X_{iM_i} is estimated by

$$\hat{X}_{iM_i}^{(\alpha)} = \hat{\mu}_{M_i} + \hat{\mathscr{A}}_i^{(\alpha)}(X_{iO_i} - \hat{\mu}_{O_i}).$$

Under the assumption that the optimal reconstruction operator $\mathscr{A}^{(\alpha)}$ is Hilbert–Schmidt, Theorem 1 of the paper proves the consistency of the estimated best linear reconstruction. That is, we show that, as the size of the training sample increases and the amount of regularization decreases, the L^2-distance between the theoretical best reconstruction and its regularized estimate converges to zero in quadratic mean and provide the rate of this convergence. It was later pointed out in the literature that our results are obtained under unnecessarily strong assumptions. Therefore, in a follow-up paper (Kraus and Stefanucci, 2020, not included here), we generalize the consistency result by relaxing the assumption that the true optimal linear reconstruction operator is Hilbert–Schmidt. It turns out that it is not even necessary to assume that the optimal reconstruction operator is bounded, and the ridge regularization method (which is Hilbert–Schmidt) still performs optimally in the limit. The follow-up paper explains this in the context of the Reproducing Kernel Hilbert Space theory.

The paper provides an estimator of the asymptotic covariance operator of the predictive distribution (error between the prediction and the target random process) and proves its consistency (Proposition 5). This enables the construction of prediction intervals. To address the problem of selection of the regularization parameter α, the paper develops a generalized cross-validation procedure for partially observed data. A simulation study is carried out to address the following goals: to investigate the performance of generalized cross-validation as a selector of the regularization parameter, to verify the validity and accuracy of the prediction intervals and bands and to explore the effect of the observation pattern. Finally, the performance of the proposed methodology is illustrated on the motivating data set of incomplete heart rate temporal profiles. Proofs of all formal statements are provided in the appendix and in a supplement.

1.5. Summary of Paper D

In Paper D (Kraus and Stefanucci, 2019), we consider classification of a functional observation into one of two groups. We formulate the theoretical (population) problem of determining the best classifier as a quadratic optimization problem on a function space, or, equivalently, as a linear inverse problem. These problems are ill-posed but, unlike in most inverse problems, this is not a complication but rather an advantage in the sense that the more ill-posed the problem is, the better optimal misclassification probability. We use regularization techniques, such as the method of conjugate gradients with early stopping and ridge regularization, to solve the optimization problem, yielding a class of regularized linear classifiers. The optimal misclassification rate is the limit along the regularization path of solutions which themselves may not converge.

We study the empirical (sample) version of the problem, where the objective function in the constrained minimization must be estimated from finite training data. We show that it is possible to construct an empirical regularization path towards the possibly non-existent unconstrained solution so that the classification error converges to its best value, possibly zero. We do this for conjugate gradient, principal component and ridge classification, in a truly infinite-dimensional manner, in the sense that the convergence takes place along a path with decreasing regularization and holds without restrictions

on the mean difference between classes. All our methodology and theory is developed in the setting of partially observed functional data, where trajectories are observed only on subsets of the domain. The principal difficulty for inference with fragments is that temporal averaging is precluded by the incompleteness of the observed functions. Our formulation as an optimization problem enables us to overcome this issue under certain assumptions because only averaging across individuals in the training data is needed, and not individual curves. We propose a domain selection strategy that looks for the best classifier with domain ranging from a minimum common domain of the training sample to the entire domain of the function to be classified. Our simulation study confirms that domain selection can considerably reduce the misclassification rate. Further simulations compare the performance of the three types of regularization. Among other findings, this study shows that the principal component and conjugate gradient classifiers often achieve comparable error rates but the latter usually needs a lower dimension of the regularization subspace, in agreement with a theoretical result we provide. Application to a data set on the geometric features of the internal carotid artery in patients with and without aneurysm demonstrates the utility of the proposed methodology.

A more detailed overview of the results of Paper D follows.

We consider classification of a Gaussian random function, X, into one of two groups of Gaussian random functions. Group 0 has mean μ_0, group 1 has mean μ_1. Both groups have covariance operator $\mathscr{R}$. We first assume that μ_0, μ_1 and $\mathscr{R}$ are known, which corresponds to the asymptotic situation with an infinite training sample. We consider the class of centroid classifiers that are based on one-dimensional projections of the form $\langle X, \psi\rangle$, where ψ is a function in $L^2(\mathcal{I})$. Given ψ, the optimal classifier based on $\langle X, \psi\rangle$ assigns X to the class $C_\psi(X)$ given by

$$C_\psi(X) = 1_{\{T_\psi(X)>0\}},$$

where $T_\psi(X) = \langle X - \bar{\mu}, \psi\rangle\langle \mu, \psi\rangle$ with $\bar{\mu} = (\mu_0 + \mu_1)/2$ and $\mu = \mu_1 - \mu_0$. The misclassification probability of this classifier is

$$1 - \Phi\left(\frac{|\langle \mu, \psi\rangle|}{2\langle \psi, \mathscr{R}\psi\rangle^{1/2}}\right).$$

The task to find the best function $\psi \in L^2(\mathcal{I})$ leads to the maximization of the argument in Φ above. We discuss when this problem can be solved within L^2 (i.e., there is an L^2-function ψ that achieves the best error rate), when it cannot be solved within L^2 (i.e., the best error rate is achieved by a linear functional but it is unbounded, not of the form $\langle X, \psi\rangle$) and what value the optimal error rate can take (remarkably, it may be zero, corresponding to perfect classification). This discussion connects the Hájek–Feldman dichotomy between Gaussian measures, the theory of reproducing kernel Hilbert spaces and constrained convex optimization. The optimization to be solved corresponds to the task to maximize $\langle \mu, \psi\rangle$ subject to $\langle \psi, \mathscr{R}\psi\rangle = 1$, which translates into the unconstrained quadratic optimization problem to minimize $\langle \psi, \mathscr{R}\psi\rangle/2 - \langle \mu, \psi\rangle$, i.e., to the linear inverse problem $\mathscr{R}\psi = \mu$.

This formulation is the starting point for the definition of regularized classifiers. Regardless of whether there is a solution (i.e., whether $\psi = \mathscr{R}^{-1}\mu$ exists in $L^2(\mathcal{I})$), one can consider an approximating, regularized problem that can be solved. Regularization is typically used to solve ill-posed inverse problems, whose solution exists, in a stable way. There, the path of regularized solutions converges to the solution to the problem of interest. Here no solution may exist, but paths of regularized solutions towards the possibly non-existent solution still turn out to be useful, since the misclassification probability converges to the optimal value along these paths. We consider three regularization methods: the principal component method (which solves the optimization in a subspace spanned by leading principal components), the conjugate gradient method (which uses the numerical method of conjugate gradients with early stopping) and the ridge method (which solves the optimization in a ball). In Propositions 1 and 3 in the paper we provide an asymptotic analysis of these methods which shows that as the amount of regularization decreases, the misclassification rate along the regularization path converges to the optimal value. This is true even when there is no bounded solution to the problem (i.e., $\mathscr{R}^{-1}\mu \notin L^2(\mathcal{I})$) and also in the "even more ill-posed" case of perfect classification (i.e., $\mathscr{R}^{-1/2}\mu \notin L^2(\mathcal{I})$). Proposition 2 compares the two methods that use a subspace for regularization, i.e., principal components and conjugate gradients, and shows that the error rate of the former is always higher than or equal to that of the latter when the same dimension is used.

We then present the empirical version with a finite training data set. Motivated by a medical dataset, we do it in the case of incomplete curves. Incompleteness can occur in the training data, with each curve possibly observed on a different domain, and in the new curve we wish to classify. A simple approach would be to consider all curves on the intersection of their observation domains, if it is non-empty, or to discard incomplete curves. However, such restrictions may be too severe and can be avoided. For group j let there be a training sample consisting of n_j independent curves $X_{j1}, \ldots, X_{jn_j}$ that may be observed incompletely with values known only on a subset O_{ji} of the domain. Then, similarly to Paper C, the mean μ_j of group j can be estimated by the cross-sectional average and the covariance kernel $\rho(s,t)$ can be estimated by the empirical covariance using pairwise complete observations of groupwise centred curves. Let the new, independent curve to be classified, X^{new}, be observed on the domain O^{new}. The empirical classifier $\hat{C}_{\hat{\psi}}$ trained on partially observed curves is defined like the theoretical one but with unknown quantities replaced by their estimators. The projection direction $\hat{\psi}$ is constructed by conjugate gradient, principal component or ridge regularization applied to estimates $\hat{\mu}$ and $\hat{\mathscr{R}}$ (defined through the estimated kernel $\hat{\rho}(s,t)$), restricted to the domain of the new curve to be classified (or, possibly, a subset of that domain).

In the theoretical analysis, we study the behaviour of classifiers for incomplete training samples of increasing size with decreasing amount of regularization. We study the conjugate gradient method with increasing number of steps, principal component method with increasing number of eigenfunctions and ridge method with decreasing ridge parameter in Theorems 1, 2 and 3, respectively. The theorems show that under specific regularity conditions they all asymptotically achieve the optimal (Bayes) misclassifica-

tion probability along the empirical regularization path as if there were infinite training data. This holds regardless of whether the theoretical best projection classifier exists as a bounded linear functional and whether the best error rate is positive or zero. Similarly to the problem of function reconstruction in Paper C and the follow-up paper Kraus and Stefanucci (2020), classification is also a prediction rather than an estimation task and we observe a similar interesting phenomenon that involves a possibly non-convergent regularization path along which the predictive performance converges to its optimum.

Further, we propose a domain selection procedure that aims to find the best domain on which the classification is performed. The method searches for the best domain between two extremes, the common domain of all training curves and the domain of the curve to be classified, to capture the location in the domain, where maximum discrimination between the two classes is.

The numerical part of the paper presents a simulation study, which compares the behaviour of the different regularization methods, investigates the performance of cross-validation for the selection of the regularization parameters, studies the impacts of partial observation and demonstrates the usefulness of the domain selection procedure. In a data example, we analyze a set of curves describing the blood vessel morphology in persons with and without aneurysm. The analysis shows an improvement of classification accuracy in comparison with existing methods due to the use of incomplete data and domain selection. Further generalizations and numerical results are contained in the supplementary document.

1.6. Summary of Paper E

Inspired by the data set of heart rate profiles, Paper E (Kraus, 2019) deals with another aspect of partially observed functional data. Although some advanced procedures, such as goodness-of-fit tests, regression, classification and reconstruction methods, have been developed for functional fragments, basic methods of inference about the fundamental characteristics of functional variables were still missing at the time of writing. In particular, the asymptotic distribution of estimators of the mean function and covariance operator, K-sample tests of equal means or covariances, and confidence intervals for eigenvalues and eigenfunctions had not been studied yet in the setting of incomplete functions. Users who wish to perform these basic tasks had the only option: to omit the partially observed functions and apply existing procedures to the complete data only. This approach is not only clearly sub-optimal due to a possibly large loss of information and resulting decay of power and accuracy, but also hardly or totally inapplicable in situations where the data contain few or no complete curves.

In this paper, we address this deficiency of existing methodology and develop essential methods of inference about the mean and covariance structure of incomplete functional data. We find appropriate assumptions on the observation pattern that enable us to establish the asymptotic distribution of estimators of μ and $\mathscr{R}$. We develop tests for comparing the mean functions in K populations of functional data based on samples of fragments. Next, we propose several tests of equal covariance operators in K samples. We also construct confidence intervals for the eigenvalues and eigenfunctions estimated

from incomplete data. The practical implementation of methods for functional fragments is more complicated than for complete curves. The main difficulty is that temporal averaging (e.g., in inner products for dimension reduction) is impossible due to missing values. This leads to asymptotic distributions whose parameters follow rather complicated formulas. More importantly, since dimension reduction is not possible, the asymptotic distributions are, upon discretization, characterized by large objects (matrices or arrays) that are difficult or even impossible to store and manipulate in computer memory. The bootstrap turns out to be a solution to this problem. We provide specific algorithms for resampling functional fragments for mean and covariance testing and for confidence intervals for eigenelements. Our simulation study shows that the proposed methods are superior to the currently only available approach based on omitting incomplete curves.

Let us now describe the contributions of Paper E more specifically.

First, we focus on inference about the mean of functional data. We consider estimation of the mean function μ by the cross-sectional average of available observations as before. In Kraus (2015, Proposition 1) (Paper C) it was shown that under non-restrictive assumptions on the observation pattern such an estimator, $\hat{\mu}$, is consistent. Paper C goes further and provides the asymptotic distribution of the estimator, which is essential in the derivation of the limiting distribution of a test statistics. The paper introduces sets of conditions on the observation pattern. Then it is shown in Theorem 1 that the estimator $\hat{\mu}$ is asymptotically distributed as a Gaussian process and a consistent estimator of the limiting covariance operator is provided. Next, we consider K independent samples of incompletely observed functional data. Our aim is to test the null hypothesis that all K mean functions are equal against the general alternative that the null does not hold. In the literature on complete functional samples there exist two main approaches to comparing mean functions. One is based on the L^2 distance between the means and one uses projections on finite dimensional subspaces. We explore both approaches in the fragmentary setting. Test statistics are constructed and their null asymptotic distributions are obtained under appropriate assumptions.

Next, we develop methods of second-order inference for functional fragments. The covariance function $\rho(s,t)$ can be estimated by the empirical covariance using pairwise complete observations. We previously showed that under certain assumptions on the observation pattern, the operator $\hat{\mathscr{R}}$ with kernel $\hat{\rho}(s,t)$ consistently estimates $\mathscr{R}$. Paper E provides a deeper asymptotic study. We determine conditions on the pattern of missingness that guarantee the weak convergence of the properly normalized difference between $\hat{\mathscr{R}}$ and $\mathscr{R}$ to a Gaussian random operator (Theorem 3). These conditions in particular do not require the existence of any completely observed curves in the data. An estimator of the limiting covariance structure is provided. Then we study the estimators $\hat{\lambda}_m$ and $\hat{\varphi}_m$ of the eigenvalues and eigenfunctions of $\mathscr{R}$. The estimators are obtained by the eigendecomposition of $\hat{\mathscr{R}}$. Theorem 4 establishes their asymptotic distributions with the help of perturbation theory. The theorem generalizes the classic results for completely observed functions. Next, we study tests for equality of covariance operators of several populations. Tests of this null hypothesis can be based on the differences between the estimators $\hat{\mathscr{R}}_j$ and a null estimator $\hat{\mathscr{R}}$. We propose two types of tests measuring the

importance of these contrasts: one approach is based on the Hilbert–Schmidt norm of the contrasts and one is based on their projections on a subspace. We give the asymptotic distribution of the Hilbert–Schmidt and projection statistics in Theorem 5. As an alternative we explore an approach (previously proposed by other authors for complete curves) that takes into account the fact that covariance operators do not form a linear subspace of the Hilbert space of Hilbert–Schmidt operators and uses the square root distance instead of the difference of covariances.

Section 4 of the paper deals with practical issues that arise due to partial observation. Functional data procedures are practically implemented by discretization. Functions then correspond to vectors (possibly with missing values), operators on the function space correspond to matrices and operators on operators correspond to four-way arrays. The direct implementation of the confidence sets and tests using the asymptotic distributions may be excessively demanding in terms of computer memory, especially in the case of covariance inference. Projection covariance tests for complete functions can avoid the computation, storage and manipulation with large arrays by computing principal scores of each function with respect to the required low number d of eigenfunctions (for example, our Paper A here does it). This dimension reduction approach is not applicable in the case of incomplete functions because the principal scores cannot be computed (temporal averaging is precluded by the incompleteness of the curves). Similar problems arise with Hilbert–Schmidt norm tests which involve a large eigenproblem that cannot be reduced due to missingness. To overcome these difficulties we use the bootstrap. We propose algorithms for mean and covariance testing and for the construction of confidence intervals that are based on the resampling of functional fragments.

In the numerical part of the paper, we perform a simulation study whose main goal is to investigate the impact of partial observation on the performance of the different mean and covariance tests and compare the proposed tests using complete and incomplete curves with the simple approach using complete curves only. We also analyze the data set of incomplete heart rate curves. All technical proofs are collected in the appendix. A supplement provides further numerical results.

References

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

Ferraty, F. and Romain, Y., editors (2011). *The Oxford Handbook of Functional Data Analysis.* Oxford University Press, Oxford.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis.* Springer, New York.

Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications.* Springer, New York.

Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators.* Wiley.

Kokoszka, P. and Reimherr, M. (2017). *Introduction to Functional Data Analysis*. CRC Press.

Kraus, D. (2015). Components and completion of partially observed functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(4):777–801.

Kraus, D. (2019). Inferential procedures for partially observed functional data. *Journal of Multivariate Analysis*, 173:583–603.

Kraus, D. and Panaretos, V. M. (2012). Dispersion operators and resistant second-order functional data analysis. *Biometrika*, 99(4):813–832.

Kraus, D. and Stefanucci, M. (2019). Classification of functional fragments by regularized linear classifiers with domain selection. *Biometrika*, 106(1):161–180.

Kraus, D. and Stefanucci, M. (2020). Ridge reconstruction of partially observed functional data is asymptotically optimal. *Statistics & Probability Letters*, 165:108813.

Panaretos, V. M., Kraus, D., and Maddocks, J. H. (2010). Second-order comparison of Gaussian random functions and the geometry of DNA minicircles. *Journal of the American Statistical Association*, 105(490):670–682.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer, New York.

A. Second-order comparison of Gaussian random functions and the geometry of DNA minicircles

By Victor M. Panaretos, David Kraus, and John H. Maddocks

Journal of the American Statistical Association, 105(490):670–682, 2010

DOI: 10.1198/jasa.2010.tm09239

Supplementary materials for this article are available online. Please click the JASA link at http://pubs.amstat.org.

Second-Order Comparison of Gaussian Random Functions and the Geometry of DNA Minicircles

Victor M. PANARETOS, David KRAUS, and John H. MADDOCKS

Given two samples of continuous zero-mean iid Gaussian processes on [0, 1], we consider the problem of testing whether they share the same covariance structure. Our study is motivated by the problem of determining whether the mechanical properties of short strands of DNA are significantly affected by their base-pair sequence; though expected to be true, had so far not been observed in three-dimensional electron microscopy data. The testing problem is seen to involve aspects of ill-posed inverse problems and a test based on a Karhunen–Loève approximation of the Hilbert–Schmidt distance of the empirical covariance operators is proposed and investigated. When applied to a dataset of DNA minicircles obtained through the electron microscope, our test seems to suggest potential sequence effects on DNA shape. Supplemental material available online.

KEY WORDS: Covariance operator; DNA shape; Functional data analysis; Hilbert–Schmidt norm; Karhunen–Loève expansion; Regularization; Spectral truncation; Two-sample testing.

1. INTRODUCTION

The understanding of the mechanical properties of the DNA molecule constitutes a fundamental biophysical task, as important biological processes, such as the packing of DNA in the nucleus or the regulation of genes, can be affected by properties such as stiffness and shape (Vilar and Leibler 2003; Tolstorukov et al. 2005). The study of these properties can focus on different scales, and accordingly involves a variety of mathematical models and techniques. At a coarse-grained level, the behavior of short (of the order of 150 base pairs) strands of DNA is likened to that of a continuous elastic rod. By means of a reaction called *cyclization*, two ends of this elastic rod bend and twist and bind together to form a loop called a *DNA minicircle*. These three-dimensional cyclic structures are an excellent specimen for examining the elastic properties of DNA since a minicircle is in a naturally stressed state without the application of external forces. Furthermore, the short length of these strands will amplify the dependence of the mechanistic behavior on intrinsic factors such as the specific base pair sequence.

Such sequence-dependent shape characteristics are of special interest as they potentially reveal a dual purpose of the DNA base-pair sequence: in addition to holding the genetic code, the sequence may influence the geometric properties of the molecule. While in principle certain particular subsequences are expected to have a strong effect on the mechanical properties of DNA, empirical detection of this effect on stereological data acquired through the electron microscope has been elusive (Hagerman 1988; Amzallag et al. 2006). A specific example is that of a subsequence called the *TATA box*, which promotes gene transcription. It is thought that the mechanical properties of this subseqence are intimately related with its function, and that its presence in a DNA minicircle will enhance its flexibility. Nevertheless, exploratory comparisons between reconstructed minicircles from microscope images containing TATA boxes with reconstructed minicircles with no TATA box did not reveal any effects due to the presence of the sequence (Amzallag et al. 2006).

Motivated by the need of two-sample comparison of loops, as exemplified in DNA minicircle experiments, this article considers the problem of second-order comparison of two samples of random functions, within a functional data analysis framework. In particular, given realisations of n_1 and n_2 independent copies of two continuous zero mean Gaussian processes X and Y on a compact set, we consider the problem of testing the hypothesis $H_0 : \mathscr{R}_X = \mathscr{R}_Y$ against the alternative $H_A : \mathscr{R}_X \neq \mathscr{R}_Y$, where the covariance operators $\mathscr{R}_X, \mathscr{R}_Y$ are not necessarily stationary. The literature on hypothesis testing for functional data is mostly concentrated on tests pertaining to the mean function (Fan and Lin 1998), as encountered, for instance, in functional linear models (Cardot et al. 2003; Cuevas, Febrero, and Fraiman 2004; Shen and Faraway 2004) or functional change detection (Berkes et al. 2009). Hall and Van Keilegom (2007) studied the important issue of the effect that the data smoothing step may have on two-sample testing. Second-order tests for functional data analysis pertaining to serial correlation were also investigated (e.g., Gabrys and Kokoszka 2007; Horváth, Hušková, and Kokoszka 2010). Although the seeds of functional two-sample covariance tests can be found in Grenander (1981), the problem of second-order comparison of functional data has—interestingly—so far received relatively little attention. A related recent article by Benko, Härdle, and Kneip (2009) proposed two-sample bootstrap tests for specific aspects of the spectrum of functional data, such as the equality of a subset of the eigenfunctions, or—assuming that the eigenfunctions are shared—equality of a subset of eigenvalues.

In this article, we consider the difficulties associated with this testing problem, and it is seen that the extension of finite-dimensional procedures can lead to complications, as the infinite-dimensional version of the problem constitutes an ill-posed inverse problem. As an alternative solution, we propose a test based on the approximation of the Hilbert–Schmidt distance of the empirical covariance operators of the two samples of functions based on the Karhunen–Loève expansion. The asymptotic distribution of the test statistic is determined and its

Victor M. Panaretos is Assistant Professor (E-mail: *victor.panaretos@epfl.ch*), David Kraus is Postdoctoral Researcher, and John H. Maddocks is Professor, Section de Mathématiques, Ecole Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland. The authors thank the editor, associate editor, and two referees for providing detailed and constructive comments, and for their fruitful suggestions. The last author wishes to acknowledge support from FN grant 205320-112178.

Journal of the American Statistical Association
June 2010, Vol. 105, No. 490, Theory and Methods
DOI: 10.1198/jasa.2010.tm09239

performance is investigated computationally. The application of our methodology to an electron microscope dataset of two groups of minicirles characterized by the presence or absence of a TATA box suggests the potential existence of significant differences in the two groups, which eluded previous analyses as these focused on the mean (the shape of the minicircle), whereas we detect the differences in the covariance structure (the flexibility/stiffness).

The article is organized as follows. The next section describes the three-dimensional functional dataset of DNA minicircles, from acquisition to registration, and includes a preliminary exploratory analysis. The first part of the third section then provides some functional data analysis background. Section 3.2 introduces our spectral test statistic and develops its asymptotic distribution, while Section 3.3 treats the problem of tuning the amount of regularization. In Section 4 the power and level of the test under various scenarios is investigated by means of simulation. Section 5 presents the results of a two-sample analysis of the DNA minicircles through the spectral test statistics, and the article concludes with a short discussion.

2. DNA MINICIRCLE DATA

The dataset of interest was reconstructed from electron micrographs imaged by Jan Bednar at the Laboratory of Ultrastructural Analysis of the University of Lausanne, Switzerland. A total of 99 DNA minicircles of 158 base-pair length were vitrified and imaged under two different angles, yielding two projected images of the same specimen, which were then used to reconstruct three-dimensional structural models (Jacob et al. 2006). The reconstructed data consist of 99 closed curves (DNA minicircles) in $\mathbb{R}^3$ of two types: both types have identical base pair sequences, except for a 14 base-pair window where 65 curves contain the *TATA sequence*, while the remaining 34 contain a different sequence, called a *CAP sequence*. Biophysical considerations suggest that the presence of a TATA box will have a significant effect on the geometry of the minicircle, and the goal is to compare these two groups to probe for such an effect.

In its reconstructed form, each curve is represented as a combination of periodic B-spline basis functions taking values in $\mathbb{R}^3$. To perform a functional data analysis of the minicircles it is required to register the data. Each curve has thus been centered and scaled, so that the center of mass is at zero and the length of the curve is one. The nature of the experimental setup in single-particle electron microscopy requires that the minicircles be imbedded unconstrained in the aqueous solution, so that the reconstructed curves are not aligned: the original (x, y, z)-coordinates for the different curves are not directly comparable as each curve was subjected to a random unobservable orthogonal transformation. It is thus necessary to align the curves. Landmark alignment methods (e.g., Gasser and Kneip 1995) are not applicable as the exact DNA sequence is not detectable from an electron micrograph. On the other hand, more flexible methods such as warping (e.g., Gervini and Gasser 2004; Tang and Müller 2008) are inappropriate since nonrigid alignment will alter the second-order properties that are of principal interest. As an alternative, we rigidly align curves by their intrinsic characteristics: each curve was individually aligned using the coordinate system induced by its *moments of inertia tensor* (e.g., Arnold 1989), which is described as follows. Consider an object in three dimensions described by a mass distribution μ—for example, for a DNA minicircle, μ will be the uniform measure supported on the curve. Suppose that the object is rotating around an axis, which without loss of generality, is given by $\text{span}(\mathbf{u}) := \{\lambda\mathbf{u} : \lambda \in \mathbb{R}\}$ for some $\mathbf{u} \in \mathbb{S}^2$. Let $r(\mathbf{u}, \mathbf{x}) := \|(\mathbf{I} - \mathbf{u}\mathbf{u}^\top)\mathbf{x}\|$ denote the distance of a point $\mathbf{x}$ from the subspace $\text{span}(\mathbf{u})$. The moment of inertia of the object around the axis $\mathbf{u}$ is given by

$$\mathscr{J}(\mathbf{u}) := \int_{\mathbb{R}^3} r^2(\mathbf{u}, \mathbf{x})\mu(d\mathbf{x}) = \int_{\mathbb{R}^3} \|(\mathbf{I} - \mathbf{u}\mathbf{u}^\top)\mathbf{x}\|^2\mu(d\mathbf{x}).$$

Given a coordinate system defined by an orthonormal basis, say the canonical basis $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$, we can use only these basis vectors to compactly represent the moment of inertia with respect to *any other axis passing by the origin*. Define the inertia matrix as

$$\mathbf{J} := \left\{ \int_{\mathbb{R}^3} \mathbf{x}^\top(\mathbf{e}_i^\top \mathbf{e}_j \mathbf{I} - \mathbf{e}_i \mathbf{e}_j^\top)\mathbf{x}\mu(d\mathbf{x}) \right\}_{i,j}.$$

Notice that the diagonal elements of the above matrix are the moments of inertia with respect to the axes of the coordinate system. The moment of inertia around any unit vector $\mathbf{u}$ can now be recovered as $\mathscr{J}(\mathbf{u}) = \mathbf{u}^\top\mathbf{J}\mathbf{u}$. Since the tensor is symmetric, it possesses real eigenvalues and orthonormal eigenvectors forming a basis, which admit the following interpretation: the first eigenvector, say $\mathbf{w}_1$, determines the axis (first principal axis of inertia, PAI1) around which the curve is most difficult to rotate, in the sense that the corresponding angular moment is maximized: $\mathbf{w}_1^\top\mathbf{J}\mathbf{w}_1 \geq \mathbf{u}^\top\mathbf{J}\mathbf{u}$ for any other $\mathbf{u} \in \mathbb{S}^2$. The projection on the plane orthogonal to $\mathbf{w}_1$ is "most spread" in this sense. The second eigenvector determines the axis within the first principal plane around which the projected curve is most difficult to rotate. That is, within the first principal plane, the projection on the line orthogonal to PAI2 is most spread. Hence, PAI3 carries the most spatial information, whereas PAI1 contains the smallest amount of information. Then, for each curve, the starting point was determined as the point where the projection on the first principal plane intersects the horizontal (PAI2) positive semi-axis and the orientation was chosen as counterclockwise in this plane (i.e., at the beginning the PAI3 coordinate increases from zero and PAI2 is positive).

The projections onto the principal axes of the minicircle curves are depicted in Figures 1 and 2. The data appear to be well aligned, and seem to be elliptical on average within the principal plane of inertia. Deviations from this principal plane, on the other hand, seem to be lacking systematic structure. The effectiveness of this alignment method is of crucial importance, as we will not be able to otherwise proceed with the testing problem (procrustean alignment of the curves will require us to optimize a sum of squares criterion with respect to 99 orthogonal transformations).

A visual inspection reveals five curves (plotted with dashed lines) that appear to be "standing out" of the rest—outliers in a broad sense. Judging whether or not a curve (an infinite dimensional object) is an outlier or not can be far trickier than in the vector case. In particular, it can be that there are further "outlying curves" that do not appear to stick out of the crowd, but are nevertheless intrinsically different from the rest. For this reason, we pursue a robust analysis for the mean curve

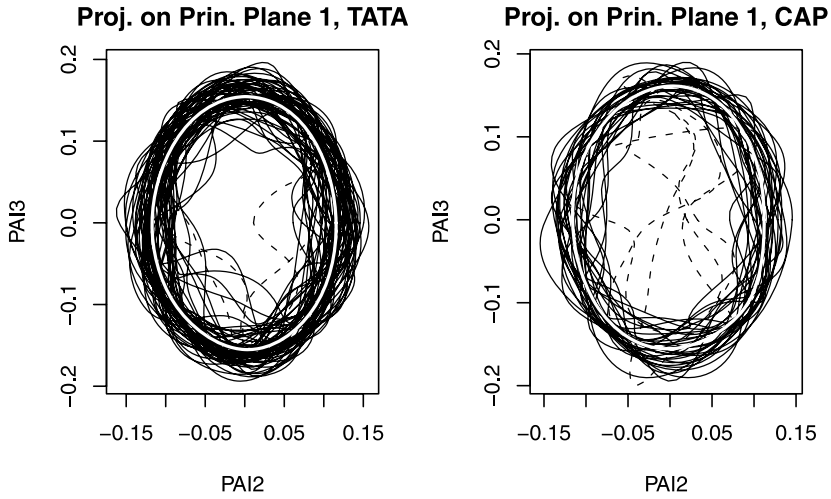

Figure 1. Projection of DNA curves on the first principal plane. Five removed outlying observations plotted in dashed lines. The mean curves (in white) are computed without outlying observations.

using a *functional median* introduced in Gervini (2008). The idea is simple: an iterative robust procedure will assign weights to each curve, and we can then detect outlying curves by looking at small weights. The method confirms our visual intuition, and reveals no further outliers. The outlying observations are removed, and after this preprocessing stage we are left with 94 aligned smooth curves.

3. METHODS

3.1 Background: FDA and Karhunen–Loève Expansions

We adopt a functional data analysis perspective (Ramsay and Silverman 2005; Ferraty and Vieu 2006) and model each curve as the realization of a stochastic process indexed by the closed interval $[0, 1]$ and taking values in $\mathbb{R}^3$ (but every-

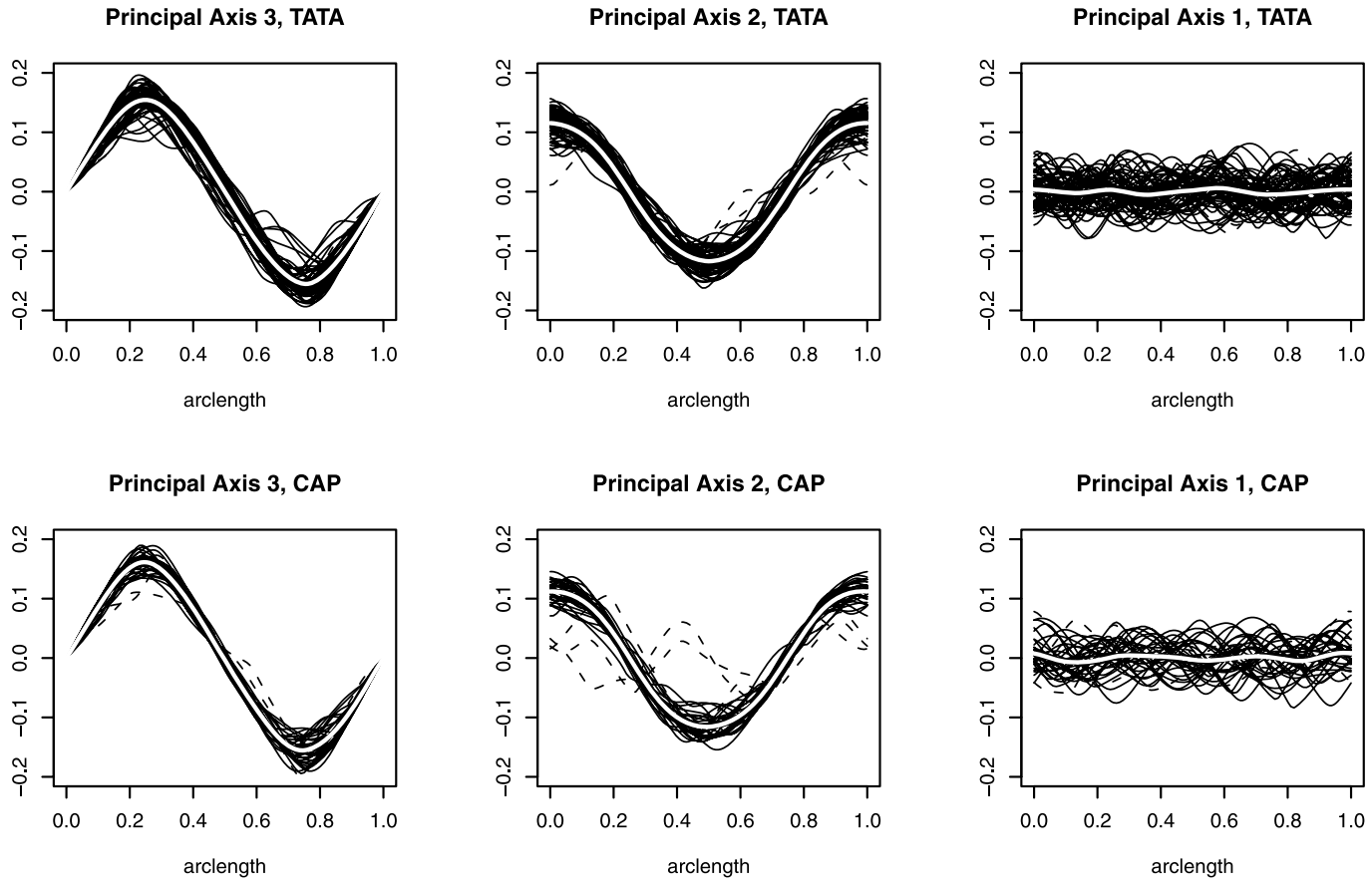

Figure 2. Coordinates of DNA curves on the principal axes of inertia. Five removed outlying observations plotted with dashed lines. Mean curves (in white) are computed without outlying observations.

thing readily extends to the case of $\mathbb{R}^d$). In particular, we assume that we have two independent collections $\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2}$, of iid Gaussian processes on $[0, 1]$, considered as random elements of the Hilbert space $\mathcal{L}^2[0, 1]$ of coordinate-wise square-integrable $\mathbb{R}^3$-valued functions with the inner product $\langle \mathbf{f}, \mathbf{g} \rangle = \int_0^1 \mathbf{f}(t)^\top \mathbf{g}(t)\, dt$. Here, $\mathbf{f}(t)^\top$ represents the transpose of the vector-valued function $\mathbf{f}(t) \in \mathbb{R}^3$. Assuming, without loss of generality, that the mean functions are zero, the processes are characterized by their respective covariance kernels $\mathbf{R}_X(s, t) = \operatorname{cov}(\mathbf{X}_i(s), \mathbf{X}_i(t)) = \mathbb{E}\{\mathbf{X}_i(s)\mathbf{X}_i^\top(t)\}$, and $\mathbf{R}_Y(s, t)$, respectively. Associated with the covariance kernel is the covariance operator $\mathscr{R}_X : \mathcal{L}^2[0, 1] \to \mathcal{L}^2[0, 1]$ defined as $\mathscr{R}_X(\mathbf{f})(t) = \operatorname{cov}(\langle \mathbf{X}_i, \mathbf{f} \rangle, \mathbf{X}_i(t)) = \int_0^1 \mathbf{R}_X(t, s)\mathbf{f}(s)\, ds$. Throughout the article, we will be assuming $\mathbf{R}_X$ to be continuous, so that $\mathscr{R}_X$ is bounded and the X process is continuous (resp. the Y process).

Inference for iid collections of infinite-dimensional random elements is often carried out in practice by an "optimal" reduction to a finite-dimensional setting, using finitely many appropriately chosen contrasts in a *functional principal component analysis* (e.g., Ramsay and Silverman 2002, 2005; Hall and Hosseini-Nasab 2006; also see Dauxois, Pousse, and Romain 1982 for distributional asymptotics). This procedure exploits the Karhunen–Loève theorem (e.g., Adler 1990), which allows for a representation of the process by a stochastic Fourier series with respect to the orthonormal eigenfunctions $\{\boldsymbol{\varphi}_X^{(j)}\}_{j=1}^\infty$ of the operator $\mathscr{R}_X$,

$$\mathbf{X}_i(t) = \sum_{j=1}^{\infty} \sqrt{\lambda_X^{(j)}}\, \xi_{ij} \boldsymbol{\varphi}_X^{(j)}(t),$$

where $\{\lambda_X^{(j)}\}_{j=1}^\infty$ is the nonincreasing sequence of corresponding eigenvalues and $\{\xi_{ij}\}$ is an iid array of standard Gaussian random variables. Convergence of the series is in mean square, uniformly in $t \in [0, 1]$.

Thus, in a practical setting, the empirical covariance kernel may be used to "optimally" reduce infinite-dimensional inferential problems to multivariate ones. Letting $\widehat{\mathbf{R}}_X$ stand for the empirical covariance kernel, $\widehat{\mathbf{R}}_X(s, t) := \frac{1}{n_1} \sum_{i=1}^{n_1} (\mathbf{X}_i(s) - \overline{\mathbf{X}}(s))(\mathbf{X}_i(t) - \overline{\mathbf{X}}(t))^\top$, we denote its eigenvalues (or *principal scores*) by $\{\widehat{\lambda}_X^{k,n_1}\}_{k=1}^{n_1}$ and its eigenfunctions (or *principal components*) by $\{\widehat{\boldsymbol{\varphi}}_X^{k,n_1}\}_{k=1}^{n_1}$. The finite-dimensional reduction is then achieved by retaining a finite number of *principal components* $\{\langle \mathbf{X}_i - \overline{\mathbf{X}}, \widehat{\boldsymbol{\varphi}}_X^{k,n_1} \rangle\}_{k=1}^K$ in lieu of each $\mathbf{X}_i$. These are zero mean and uncorrelated random variables, with corresponding sample variances $\widehat{\lambda}_X^{k,n_1}$. Similarly, for the second sample, the analogous quantities are $\mathbf{R}_Y, \mathscr{R}_Y, \lambda_Y^{(j)}, \boldsymbol{\varphi}_Y^{(j)}$ (and their empirical "hat" counterparts). The dimension reduction afforded by the Karhunen–Loève expansion is the tool we will next employ to construct our test.

3.2 Second-Order Comparison of Gaussian Processes

Let $\{\mathbf{X}_i\}_{i=1}^{n_1}$ and $\{\mathbf{Y}_i\}_{i=1}^{n_2}$ constitute two iid random samples of Gaussian processes indexed by the interval $[0, 1]$ and taking values in $\mathbb{R}^3$ (or indeed $\mathbb{R}^d$). As mentioned in the previous section, these are regarded as random elements of the Hilbert space $\mathcal{L}^2[0, 1]$ of square-integrable $\mathbb{R}^3$-valued functions (where integration is to be understood coordinate-wise). Assuming that the covariance operators $\mathscr{R}_X$ and $\mathscr{R}_Y$ associated with the processes are continuous, we wish to test the hypothesis pair

$$\begin{cases} H_0 : \mathscr{R}_X = \mathscr{R}_Y, \\ H_A : \mathscr{R}_X \neq \mathscr{R}_Y. \end{cases} \tag{1}$$

A natural first approach to developing a test for the hypothesis pair in Equation (1) is to attempt to extend tests developed for the finite-dimensional version of the problem, which was extensively studied. The majority of test statistics for the equality of covariance matrices of Gaussian vectors are based on the determinant, trace, or maximum/minimum eigenvalues of matrices such as: $\mathbf{S}_1\mathbf{S}_2\mathbf{S}^{-1}$, $\mathbf{S}_1\mathbf{S}_2^{-1}$, $\mathbf{S}_2(\mathbf{S}_1 + \mathbf{S}_2)^{-1}$ (Roy 1953; Pillai 1955; Kiefer and Schwartz 1965; Giri 1968); here, $\mathbf{S}_1$ and $\mathbf{S}_2$ are the empirical covariance matrices corresponding to each sample, and $\mathbf{S}$ is the pooled empirical covariance matrix. Evidently, such tests cannot immediately be carried over to the case of Gaussian processes: inversion of an empirical covariance operator will be required, which transforms the construction of the test statistic into an ill-posed inverse problem.

The operator $\widehat{\mathscr{R}}_X^{n_1}$ (resp. $\widehat{\mathscr{R}}_Y^{n_2}$) will be of rank at most n_1 (resp. n_2) as its image is the subspace spanned by $\{\mathbf{X}_i\}_{i=1}^{n_1}$ (resp. $\{\mathbf{Y}_i\}_{i=1}^{n_2}$). Therefore, we cannot talk of its inverse, except if we restrict the operator on $\operatorname{span}\{\mathbf{X}_i\}_{i=1}^{n_1}$ (resp. $\operatorname{span}\{\mathbf{Y}_i\}_{i=1}^{n_1}$), but the two spans will *not* coincide in general and the two empirical operators will *not* be diagonalized by the same basis. Furthermore, since the processes are assumed to be second order, the operators $\mathscr{R}_X$ and $\mathscr{R}_Y$ are necessarily bounded (in fact compact), and it must be the case that $\lambda_X^{(k)}, \lambda_Y^{(k)} \overset{k\to\infty}{\longrightarrow} 0$, the rate of convergence depending on the degree of smoothness of the Gaussian processes (the smoother the process, the faster the rate). Thus, for any finite n_1 and n_2, however large, a test statistic employing an "inverse" of $\widehat{\mathscr{R}}_X$ composed with $\widehat{\mathscr{R}}_Y$ will be unstable to perturbations of the Y-data.

In the infinite-dimesional case, we propose the use of a test statistic based on the norm of the difference of the two empirical covariance operators. Recall that for trace-class operators, one may define the *Hilbert–Schmidt norm*. Consider an integral operator $\mathscr{R} : \mathbf{f} \mapsto \int_0^1 \mathbf{R}(\cdot, s)\mathbf{f}(s)\, ds$ such that $\int_0^1 \int_0^1 \operatorname{trace}\{\mathbf{R}(s, t)^\top \mathbf{R}(s, t)\}\, ds\, dt < \infty$. The Hilbert–Schmidt norm of the operator $\mathscr{R}$ is defined as

$$\|\mathscr{R}\|_{\mathrm{HS}} := \sqrt{\int_0^1 \int_0^1 \operatorname{trace}\{\mathbf{R}(s, t)^\top \mathbf{R}(s, t)\}\, ds\, dt}.$$

Assuming that the covariance operators in question are Hilbert–Schmidt, a test may be based on the squared Hilbert–Schmidt distance $\|\widehat{\mathscr{R}}_X^N - \widehat{\mathscr{R}}_Y^N\|_{\mathrm{HS}}^2$ of their empirical counterparts. Of course, the sampling distribution of this latter quantity will depend on the unknown covariance operators even asymptotically. To be able to "normalize" the test statistic, we employ a very useful property of the Hilbert–Schmidt norm: for any orthonormal system $\{\mathbf{e}_i\}_{i=1}^\infty$ of $\mathcal{L}^2[0, 1]$, we have

$$\|\mathscr{R}\|_{\mathrm{HS}}^2 = \sum_{i=1}^{\infty} \|\mathscr{R}\mathbf{e}_i\|_{\mathcal{L}^2}^2. \tag{2}$$

Therefore, we may use a basis to obtain a countable expression for $\|\widehat{\mathscr{R}}_X^N - \widehat{\mathscr{R}}_Y^N\|_{\mathrm{HS}}^2$. In practice, one will need to truncate a series such as the above to obtain an "optimal" finite-dimensional

reduction, that is, the choice of contrasts $\{\mathbf{e}_i\}$ should be such that the truncated version of Equation (2) retains the bulk of the norm.

For each of the two empirical operators, the optimal contrasts will coincide with their eigenfunctions, as dictated by the Karhunen–Loève expansion, but to use the relation in Equation (2) we need to use a common basis. As a compromise, we thus choose the eigenfunctions $\{\widehat{\boldsymbol{\varphi}}_{XY}^{k,N}\}$ corresponding to the empirical covariance operator of the pooled sample of $N = n_1 + n_2$ curves and base our test on

$$\sum_{k=1}^{K} \|(\widehat{\mathscr{R}}_X^N - \widehat{\mathscr{R}}_Y^N)\widehat{\boldsymbol{\varphi}}_{XY}^{k,N}\|_{\mathcal{L}^2}^2,$$

which by Parseval's theorem, may be further approximated by

$$\sum_{i=1}^{K}\sum_{j=1}^{K} \langle(\widehat{\mathscr{R}}_X^N - \widehat{\mathscr{R}}_Y^N)\widehat{\boldsymbol{\varphi}}_{XY}^{i,N}, \widehat{\boldsymbol{\varphi}}_{XY}^{j,N}\rangle^2. \tag{3}$$

With this quantity in mind, the following theorem, whose proof may be found in the Appendix, provides the basis for our test:

Theorem 1. Let $\{\mathbf{X}_n\}_{n=1}^{n_1}$ and $\{\mathbf{Y}_n\}_{n=1}^{n_2}$ be two collections of zero mean iid continuous Gaussian random functions indexed by the interval $[0, 1]$ and taking values in $\mathbb{R}^d$, possessing covariance operators $\mathscr{R}_X$ and $\mathscr{R}_Y$ with distinct eigenvalues. Let $\widehat{\mathscr{R}}_X^{n_1}$ and $\widehat{\mathscr{R}}_Y^{n_2}$ denote the empirical covariance operators based on $\{\mathbf{X}_n\}_{n=1}^{n_1}$ and $\{\mathbf{Y}_n\}_{n=1}^{n_2}$. For $N = n_1 + n_2$, let $\widehat{\mathscr{R}}_{XY}^N$ denote the empirical covariance operator of the pooled collection, and $\{\widehat{\boldsymbol{\varphi}}_{XY}^{k,N}\}_{k=1}^N$ the corresponding eigenfunctions. Finally, let $\widehat{\lambda}_{X,XY}^{k,n_1}$, $\widehat{\lambda}_{Y,XY}^{k,n_2}$ denote the empirical variance of the kth Fourier coefficient of $\{\mathbf{X}_n\}_{n=1}^{n_1}$ and $\{\mathbf{Y}_n\}_{n=1}^{n_2}$, respectively, with respect to the eigenfunctions $\{\widehat{\boldsymbol{\varphi}}_{XY}^{n,K}\}_{n=1}^N$. Assuming that $\mathbb{E}[\|\mathbf{X}_1\|_{L^2}^4] < \infty$, $\mathbb{E}[\|\mathbf{Y}_1\|_{L^2}^4] < \infty$, and $n_1/N \to \theta \in (0, 1)$ as $N = n_1 + n_2 \to \infty$, it follows that, under the hypothesis $H_0 : \mathscr{R}_X = \mathscr{R}_Y$,

$$\begin{aligned} T_N(K) \;:=\; & \frac{n_1 n_2}{2N}\sum_{i=1}^{K}\sum_{j=1}^{K}\big(\langle(\widehat{\mathscr{R}}_X^{n_1} - \widehat{\mathscr{R}}_Y^{n_2})\widehat{\boldsymbol{\varphi}}_{XY}^{i,N}, \widehat{\boldsymbol{\varphi}}_{XY}^{j,N}\rangle^2\big) \\ & \Big/\left(\left(\frac{n_1}{N}\widehat{\lambda}_{X,XY}^{i,n_1} + \frac{n_2}{N}\widehat{\lambda}_{Y,XY}^{i,n_2}\right)\right. \\ & \left.\times\left(\frac{n_1}{N}\widehat{\lambda}_{X,XY}^{j,n_1} + \frac{n_2}{N}\widehat{\lambda}_{Y,XY}^{j,n_2}\right)\right) \\ \xrightarrow{w}\; & \chi_{K(K+1)/2}^2 \end{aligned}$$

as $N \to \infty$, for any finite $K \le \operatorname{rank}(\mathscr{R}_X) = \operatorname{rank}(\mathscr{R}_Y) \le \infty$.

Under the alternative hypothesis, the test statistic will converge to a sum of $K(K+1)/2$ dependent shifted chi square random variables.

Our proposed test procedure is thus to *reject* the hypothesis $H_0 : \mathscr{R}_X = \mathscr{R}_Y$ at level α, whenever the test statistic exceeds the corresponding critical value,

$$T_N(K) \ge \chi_{K(K+1)/2,1-\alpha}^2.$$

Of course, conducting the test requires the selection of a spectral truncation level, K. This choice must be made judiciously, as it has a direct bearing on the power of the test:

1. Conservative choices of K [i.e., choosing $K \ll \operatorname{rank}(\mathscr{R}_X) \wedge \operatorname{rank}(\mathscr{R}_Y)$] may result in Type II error due to differences in the higher frequency covariance structure, especially in situations where the two covariances share the same eigenfunctions, but have different eigenvalues at higher frequencies.
2. Greedy choices of K [choosing $K > \operatorname{rank}(\mathscr{R}_X) \wedge \operatorname{rank}(\mathscr{R}_Y)$] will inflate the variance of the test statistic since an element of ill-posedness will enter when dividing with the empirical eigenvalues of higher order terms.

In the latter sense, the test can also be thought of as an $\mathcal{L}^2$-regularized test. These aspects are further considered quantitatively in Section 4. It should be noted that the problem of choosing K is directly analogous to the choice of a cutoff point in principal component analysis and the choice of a bandwidth in a nonparametric problem; thus we deal with it using empirical eigenvalue scree-plots as well as penalized goodness-of-fit criteria (see Sections 3.3 and 5.1).

A more user-friendly expression for the test statistic T can be given if we introduce some additional notation. Let $\widehat{\lambda}_{X,XY}^{ij,N} := \langle\widehat{\mathscr{R}}_X^{n_1}\widehat{\boldsymbol{\varphi}}_{XY}^{i,N}, \widehat{\boldsymbol{\varphi}}_{XY}^{j,N}\rangle = n_1^{-1}\sum_i \langle\mathbf{X}_i - \overline{\mathbf{X}}, \widehat{\boldsymbol{\varphi}}_{XY}^{i,N}\rangle\langle\mathbf{X}_i - \overline{\mathbf{X}}, \widehat{\boldsymbol{\varphi}}_{XY}^{j,N}\rangle$ be the empirical covariance of the ith and jth Fourier coefficients of the X-curves, with respect to the basis $\{\widehat{\boldsymbol{\varphi}}_{XY}^{k,N}\}_{k\ge1}$ (resp. $\widehat{\lambda}_{Y,XY}^{ij,N}$). For simplicity, we also write $\widehat{\lambda}_{X,XY}^{jj,N} \equiv \widehat{\lambda}_{X,XY}^{j,N}$ (resp. $\widehat{\lambda}_{Y,XY}^{jj,N}$). Then we may re-express the test statistic as

$$\begin{aligned} T_N(K) := \frac{n_1 n_2}{2N}\sum_{i=1}^{K}\sum_{j=1}^{K}\big((\widehat{\lambda}_{X,XY}^{ij,N} - \widehat{\lambda}_{Y,XY}^{ij,N})^2\big) \\ \Big/\left(\left(\frac{n_1}{N}\widehat{\lambda}_{X,XY}^{i,n_1} + \frac{n_2}{N}\widehat{\lambda}_{Y,XY}^{i,n_2}\right)\right. \\ \left.\times\left(\frac{n_1}{N}\widehat{\lambda}_{X,XY}^{j,n_1} + \frac{n_2}{N}\widehat{\lambda}_{Y,XY}^{j,n_2}\right)\right). \end{aligned}$$

If for some reason, we a priori know the eigenfunctions of $\mathscr{R}_X$ and $\mathscr{R}_Y$ to be equal, then the following test statistic may be used instead of T:

$$T_1 = \sum_{k=1}^{K}\frac{n_1 n_2}{N}\frac{(\widehat{\lambda}_{X,XY}^{k,N} - \widehat{\lambda}_{Y,XY}^{k,N})^2}{2((n_1/N)\widehat{\lambda}_X^{k,N} + (n_2/N)\widehat{\lambda}_Y^{k,N})^2}.$$

The motivation for this statistic is that when the eigenfunctions coincide, then

$$\sum_{k=1}^{K}\|(\widehat{\mathscr{R}}_X^{n_1} - \widehat{\mathscr{R}}_Y^{n_2})\widehat{\boldsymbol{\varphi}}_{XY}^{k,N}\|_{\mathcal{L}^2}^2 \approx \sum_{k=1}^{K}(\widehat{\lambda}_{X,XY}^{k,N} - \widehat{\lambda}_{Y,XY}^{k,N})^2.$$

It follows as an immediate corollary to Theorem 1 that, under H_0, the statistic T_1 is asymptotically chi-square distributed with K degrees of freedom [assuming $n_1/N \to \theta \in (0, 1)$]. One may also wish to consider modified versions of the test statistics T and T_1, obtained via suitable variance-stabilizing transformations. In the case of the test statistic T, we apply a log transformation to the diagonal terms of the sum in Equation (3), and Fisher's z-transformation to the off-diagonal terms to obtain a

test statistic with the same asymptotic distribution as T (an immediate corollary to Theorem 1),

$$T^* = \sum_{k=1}^{K} \frac{n_1 n_2}{N} \frac{(\log \widehat{\lambda}^{k,N}_{X,XY} - \log \widehat{\lambda}^{k,N}_{Y,XY})^2}{2} + \sum_{1 \le j < k \le K} \frac{n_1 n_2}{N} \left(\frac{1}{2} \log \frac{\sqrt{\widehat{\lambda}^{j,N}_{XY} \widehat{\lambda}^{k,N}_{XY}} + \widehat{\lambda}^{jk,N}_{X,XY}}{\sqrt{\widehat{\lambda}^{j,N}_{XY} \widehat{\lambda}^{k,N}_{XY}} - \widehat{\lambda}^{jk,N}_{X,XY}} - \frac{1}{2} \log \frac{\sqrt{\widehat{\lambda}^{j,N}_{XY} \widehat{\lambda}^{k,N}_{XY}} + \widehat{\lambda}^{jk,N}_{Y,XY}}{\sqrt{\widehat{\lambda}^{j,N}_{XY} \widehat{\lambda}^{k,N}_{XY}} - \widehat{\lambda}^{jk,N}_{Y,XY}} \right)^2.$$

A variance-stabilized alternative to T_1 may also be similarly constructed by retaining only the first component of T^* (the diagonal terms), yielding

$$T_1^* = \sum_{j=1}^{K} \frac{n_1 n_2}{N} \frac{(\log \widehat{\lambda}^{j,N}_{X,XY} - \log \widehat{\lambda}^{j,N}_{Y,XY})^2}{2}.$$

The latter statistic is approximately χ^2-distributed with K degrees of freedom. Simulations conducted in Section 4 seem to suggest that the modified tests achieve a level closer to the nominal level, and consequently, may provide higher power.

In the infinite rank case, one might wish to let K to grow along with N, allowing for the comparison of progressively finer and finer differences (located at the extreme tails of the operator spectra) as sample size increases. As noted previously, any such attempt will necessarily lead to instabilities: due to the fast decay of the eigenvalues, we are attempting to compare extremely small quantities, based on the empirical tails of the spectra, which are highly unstable. This instability will manifest itself through the very large integrated mean squared errors involved when estimating higher order eigenfunctions, whose available bounds grow for fixed N depending inversely on the rate of decay of the spectrum (see also Bosq 2000, lemma 4.3); the ill-posedness is especially severe for smooth processes. Controlling the rate of growth of K with respect to both N and the rate of decay of the true eigenvalues will thus be necessary—decreasing the amount of regularization requires an increase in sample size, depending also on the spectral decay properties. Modifying the test statistic to obtain a central limit theorem as $K_N \to \infty$ will require a very slow rate of growth of K_N with respect to N since:

1. Although the truncation level grows as K_N, the number of summands in the test statistic grows like K_N^2.
2. While these K_N^2 summation terms do become independent as N grows (allowing for a CLT phenomenon), no *mixing concept* applies. In effect this means that one has to look at the convergence in distribution to independence of a random vector of increasing dimension ($= K_N^2$). For any fixed dimension the required weak convergence will be at a rate of $N^{-1/2}$—therefore K_N must grow slow enough to allow the $N^{-1/2}$ rate to compensate for the K_N^2 rate of increase of the dimension.
3. The required global convergence to independence is regulated by the convergence of the empirical eigenfunctions to the true ones; this in turn depends on the spacings between the true eigenvalues. For K components, the rate of convergence of the Kth empirical eigenfunction decays like $N^{-1/2} \max\{(\lambda_{K-1} - \lambda_K)^{-1}, (\lambda_K - \lambda_{K+1})^{-1}\}$. Therefore, when we let K_N grow, it has to be at a rate slow enough to annihilate the blow-up of the inverse spacing of order K_N.

The study of these intricacies is rather technical, and further development is contained in the supplement.

3.3 On the Selection of Truncation Level

By analogy to finite-dimensional principal component analysis (PCA), the choice of a truncation parameter K can be made on the basis of scree plots and cumulative variance plots. A visual inspection of the scree plots can be employed to identify inflection points, which combined with the information provided by the cumulative variance plots, can suggest an appropriate truncation level K for use in testing. Note that the decrease of the scores $\widehat{\lambda}^{k,N}_{X,XY}$ and $\widehat{\lambda}^{k,N}_{Y,XY}$ is not monotone, since the basis $\{\widehat{\boldsymbol{\varphi}}^{k,N}_{XY}\}$ does not correspond to the eigenbasis of either of the two groups of curves. Therefore, a little more care needs to be taken, although the basic idea still holds.

The truncation of the Hilbert–Schmidt norm expansion effectively induces smoothing upon the curves, and can be regarded as a choice of a *regularization tuning parameter*. Consequently, potentially more automatic criteria can be based on tuning the amount of smoothing so as to minimize a penalized goodness-of-fit error. Concentrating on the X-curves, a natural definition of goodness-of-fit error is,

$$\mathrm{PE}_X(K) := \sum_{n=1}^{n_1} \left\| \sum_{k=1}^{K} \langle \mathbf{X}_n^*, \widehat{\boldsymbol{\varphi}}^{k,N}_{XY} \rangle \widehat{\boldsymbol{\varphi}}^{k,N}_{XY} - \mathbf{X}_n^* \right\|^2_{\mathcal{L}^2} = \sum_{n=1}^{n_1} \| \widetilde{\mathbf{X}}_n(K) - \mathbf{X}_n^* \|^2_{\mathcal{L}^2},$$

where $\mathbf{X}_i^*$ is the ith mean-corrected curve. Of course, the above criterion is nonincreasing in K since it accounts only for the fit, and there is no penalty for the "complexity" of $\widetilde{\mathbf{X}}_n(K)$. Such a penalty is often based on the norm of the image of $\widetilde{\mathbf{X}}_n(K)$ through a suitably chosen differential operator (in the spirit of Ramsay and Silverman 2005, section 5.3.3). The choice of penalty reflects the qualitative specification of what "parsimonious" is in a given context. In the present scenario, a sample of curves is available, and so the penalty can be made to be data-dependent, by penalizing deviations from the average smoothness properties of the observed curves. These smoothness properties are naturally reflected by the norm of the *reproducing kernel Hilbert space* (RKHS) generated by the empirical covariance operator of the X-sample, $\widehat{\mathscr{R}}_X$, yielding the penalized

fit criterion,

$$\mathrm{PFC}_X(K) = \underbrace{\sum_{n=1}^{n_1} \|\widetilde{\mathbf{X}}_n(K) - \mathbf{X}_n^*\|^2_{\mathcal{L}^2}}_{\mathrm{GOF}_X(K)} + \underbrace{\frac{2\sum_{j=1}^{N}\widehat{\lambda}_{XY}^{j,N}}{n_1}\sum_{n=1}^{n_1}\sum_{j=1}^{n_1}\frac{1}{\widehat{\lambda}_X^{j,N}}\langle\widetilde{\mathbf{X}}_n(K),\widehat{\boldsymbol{\varphi}}_X^{j,N}\rangle^2}_{\mathrm{PEN}_X(K)}. \quad (4)$$

When the null hypothesis is true, we expect to have $\widehat{\boldsymbol{\varphi}}_X^{j,N} \approx \widehat{\boldsymbol{\varphi}}_{XY}^{j,N}$; this essentially reduces $\mathrm{PFC}_X(K)$ to the Gaussian pseudo-likelihood-based Akaike information criterion (AIC) employed by Yao, Müller, and Wang (2005a) (see also Yao, Müller, and Wang 2005b). The analogous quantity $\mathrm{PFC}_Y(K)$ can similarly be defined for the Y-curves. Since the sample size for the two groups are not equal, the natural choice of K is then given by minimizing the sum of goodness-of-fit terms [$\mathrm{GOF}_X(K)$ and $\mathrm{GOF}_Y(K)$] plus the convex combination of the smoothness penalties [$\mathrm{PEN}_X(K)$ and $\mathrm{PEN}_Y(K)$]:

$$\arg\min_K \Big\{\mathrm{GOF}_X(K) + \mathrm{GOF}_Y(K) + \frac{n_1}{N}\mathrm{PEN}_X(K) + \frac{n_2}{N}\mathrm{PEN}_Y(K)\Big\}.$$

In practice, the number of terms taken in the sum comprising the penalty may be less than n_i, to avoid dividing by terms that are numerically zero. A variant of this selection criterion can be based on the leave-one-out cross-validated prediction error, where one whole curve is left out at a time (Rice and Silverman 1991). The performance of the selection criterion is investigated in simulations presented in the next section.

4. A SIMULATION STUDY

To assess the behavior of the proposed tests under the null hypothesis and under various alternatives we carry out a number of simulations. We consider one situation with equal covariance functions (simulation scenario A) and several alternative configurations (scenarios B–I). The two test statistics T and T^* introduced in the previous section are considered under various choices of K, the truncation level, and for the automatic selection K^* given by the penalized fit criterion. The number of observations in each sample is 50. The tests are replicated 5000 times under H_0 and 1000 times under H_A, respectively, at the 5% nominal level of significance using the asymptotic χ^2 approximation.

In the first eight scenarios, the Gaussian processes in both samples are of the form

$$\sum_{j=1}^{3}\xi_j\sqrt{2}\sin(2\pi j(t+\delta_j)) + \sum_{j=1}^{3}\zeta_j\sqrt{2}\cos(2\pi j(t+\eta_j)), \qquad t\in[0,1],$$

where the coefficients ξ_j, ζ_j are independent Gaussian random variables with mean zero and $\mathrm{var}(\xi_j) = v_j$, $\mathrm{var}(\zeta_j) = w_j$ (the variance terms where chosen so as to induce "elbow" effects as one expects to see in practice). Various values of v_j, w_j, δ_j, η_j used in A–H are reported together with the corresponding results in Table 1 (the shift parameters δ_j, η_j are reported only for F, the only case where they are nonzero). The last scenario deals with rough processes (infinitely many components).

Results for scenario A show that the true level for all variants of the test is close to the nominal level, provided the number of

Table 1. Empirical rejection probabilities on the nominal level 5%, sample size $n_1 = n_2 = 50$, number of replications 5000 for A, 1000 for B–I. Here, $\mathbf{u}^X = (\mathbf{v}^X, \mathbf{w}^X)$ (resp. $\mathbf{u}^Y$) and K^* is the automatic truncation choice given by the penalised fit criterion

			K				
	Parameters	Test	1	2	3	4	K^*
A	$\mathbf{u}^X = (12, 7, 0.5, 9, 5, 0.3)$	T	0.045	0.049	0.044	0.044	0.047
	$\mathbf{u}^Y = (12, 7, 0.5, 9, 5, 0.3)$	T^*	0.051	0.056	0.057	0.056	0.059
B	$\mathbf{u}^X = (14, 7, 0.5, 6, 5, 0.3)$	T	0.422	0.264	0.185	0.150	0.148
	$\mathbf{u}^Y = (8, 7, 0.5, 6, 5, 0.3)$	T^*	0.443	0.315	0.223	0.174	0.175
C	$\mathbf{u}^X = (15, 10, 0.5, 4, 3, 0.3)$	T	0.186	0.331	0.218	0.169	0.167
	$\mathbf{u}^Y = (11, 6, 0.5, 4, 3, 0.3)$	T^*	0.201	0.366	0.269	0.207	0.208
D	$\mathbf{u}^X = (12, 7, 0.5, 9, 3, 0.3)$	T	0.040	0.204	0.836	0.973	0.962
	$\mathbf{u}^Y = (12, 7, 0.5, 2, 5, 0.3)$	T^*	0.047	0.221	0.848	0.984	0.980
E	$\mathbf{u}^X = (12, 7, 0.5, 9, 3, 0.3)$	T	0.047	0.246	0.644	0.964	0.962
	$\mathbf{u}^Y = (12, 7, 0.5, 3, 9, 0.3)$	T^*	0.055	0.267	0.686	0.976	0.975
F	$\mathbf{u}^X = \mathbf{u}^Y = (12, 7, 4, 0.5, 0.3, 0.1)$	T	0.257	0.693	0.909	1.000	1.000
	$\boldsymbol{\delta}^X = (0.15, 0.15, 0.15)$	T^*	0.273	0.706	0.916	1.000	1.000
G	$\mathbf{u}^X = (12, 7, 0.5, 8, 6, 0.3)$	T	0.042	0.040	0.054	1.000	1.000
	$\mathbf{u}^Y = (12, 7, 0.5, 8, 0, 0.3)$	T^*	0.047	0.048	0.068	1.000	1.000
H	$\mathbf{u}^X = (12, 7, 0.5, 9, 5, 0.3)$	T	0.044	0.140	0.500	1.000	1.000
	$\mathbf{u}^Y = (12, 7, 0.5, 0, 5, 0.3)$	T^*	0.049	0.154	0.520	1.000	1.000
I	Brownian motion versus	T	0.719	0.608	0.483	0.377	0.493
	Ornstein–Uhlenbeck process	T^*	0.731	0.644	0.532	0.443	0.546

components K does not exceed the effective complexity of the covariance operator (which is 4 in this case). The slight conservatism of T is removed by variance stabilizing transformations used in T^*. Indeed, the stabilized statistics seem to be preferable because they also provide slightly higher power (as is seen in the remaining simulations).

Under scenario B, both covariance operators are of effective complexity 4 and possess the same sequence of eigenfunctions (the same set with the same order), but the sequences of eigenvalues differ (the largest eigenvalue is different). Not surprisingly, the power decreases as K increases because there is no difference in the components other than in the first one, so adding them increases the degrees of freedom without any significant contribution to the test statistic. Configuration C is similar to B, but with the two largest eigenvalues being different. The highest power is achieved with $K = 2$, as expected. When compared to the next few scenarios, where there are differences associated with the eigenfunctions also, the power in B and C is clearly lower. This is due to the fact that the test statistic takes the comparison of the eigenfunctions—where there are no differences—into account, and thus is not as powerful in detecting differences that lie only on the eigenvalues (the diagonal form of the tests T_1 and T_1^* will be more powerful in this case).

In scenario D, the effective complexity of the operators is the same in Equation (4), the operators have the same set of eigenfunctions (in different order) and different sequences of eigenvalues. The difference of the covariance operators is not detected by tests with one component because the largest eigenvalue and the corresponding eigenfunction are the same in both samples. When the choice of K is close to the true effective complexity, the power of the tests is very high (this includes the automatic choice). The same is true for the next four scenarios as well.

Under scenario E, both operators (of effective rank 4) have the same sequence of eigenvalues, and the same set of eigenfunctions, but the latter are permuted to correspond to different eigenvalues. This scenario illustrates a situation where the diagonal form of the test statistics (T_1 and T_1^*) will be inapplicable. It is interesting to make the comparison with scenario D, where the sets of eigenfunctions are the same for both samples as well. In D the sequences of eigenvalues differ also, hence more information is on the diagonal.

Scenario F differs from the previous configurations in that the sets of eigenfunctions are completely different (sines versus shifted sines). The eigenvalues are the same, and the effective operator rank is 3 in both cases.

In the next configuration, scenario G, the first three eigenvalues and eigenfunctions are the same in both samples. The covariance operators have different effective ranks: 4 in the first sample, 3 in the second sample. Therefore, it is not surprising that the departure from H_0 is not detected by tests with less than 4 components while it is clearly detected by four-component tests. Note that with the automatic choice K^*, the alternative is always detected.

Configuration H is again a situation with different effective ranks of operators (4 versus 3) but unlike the previous situation, only the first eigenfunction and eigenvalue coincide in both samples. The next two eigenvalues are different and the corresponding eigenfunctions differ as well. Thus, as of $K = 2$, the tests start detecting the alternative, with highest power for $K = 4$.

Under scenario I, curves in both samples come from distributions with covariance operators with infinite rank, namely the standard Brownian motion $W(t)$ and the Ornstein–Uhlenbeck process $U(t)$ satisfying $dU(t) = -\theta U(t)\,dt + dW(t)$ with $\theta = 1$. The covariance operators of the two processes differ in all components. The major portion of the difference is captured by tests with one component, then the power slowly decays.

A general observation when focusing on the behavior of the tests when the number of components K was selected using the selection criterion introduced in the previous section is that the power and level are comparable with those when employing the true effective rank. Under scenario A, the selection criterion chose $K = 4$ in 96.3% of simulations and $K = 5$ in 3.7% of simulations. Doing the same for the alternative configurations, it turned out that the power is similar to the power of tests with fixed values of K close to the values most frequently selected by the selection criterion. Hence this automatic dimension reduction technique appears to be useful in practice.

It should be mentioned that the role of the selection criterion is to probe the effective complexity of the data and not the complexity of the difference between the two samples. The selection rule is not related to the null hypothesis or the alternative and does not reflect validity or invalidity of either of them. This explains the reliability of the post-selection test. Note that a completely different approach can be based on the selection of the "most different" components (the most likely alternative) using a criterion involving the test statistic in the spirit of data-driven smooth tests (e.g., Ledwina 1994).

5. ANALYSIS OF DNA MINICIRCLES

5.1 Finite-Dimensional Approximation

Figure 3 shows the empirical variance of the scores with respect to the basis $\{\boldsymbol{\varphi}_{XY}^{k,N}\}$ separately for the TATA and CAP groups ($\widehat{\lambda}_{X,XY}^{k,N}$ and $\widehat{\lambda}_{Y,XY}^{k,N}$, respectively, in the notation used previously) as well as for the pooled sample ($\widehat{\lambda}_{XY}^{j,N}$). The plots also display cumulative proportions of the total variance explained by the corresponding components. Separate plots are constructed for the analysis carried out marginally on each principal axis and jointly on the principal plane.

When inspecting the marginal plots for the projections on each axis of inertia, we observe that four or at most five principal components should constitute an adequate choice. When looking at the marginal plot for the projection onto the principal plane of inertia, it seems that setting $K = 6$ or $K = 7$ is more than adequate (accounting for at least 85% and 90% of the variance, respectively, and with a clear "elbow" effect).

The reason for placing special emphasis on the principal plane is that, as one can observe from Figure 2, the DNA minicircle curves tend to be planar on average, and the more interesting signal is not to be found in the deviations from the planar aspect of the structure, but within the planar structure itself (see the discussion at the end of the next section). The penalized prediction error criterion introduced in Section 3.3 yields $K = 7$ components in the principal plane.

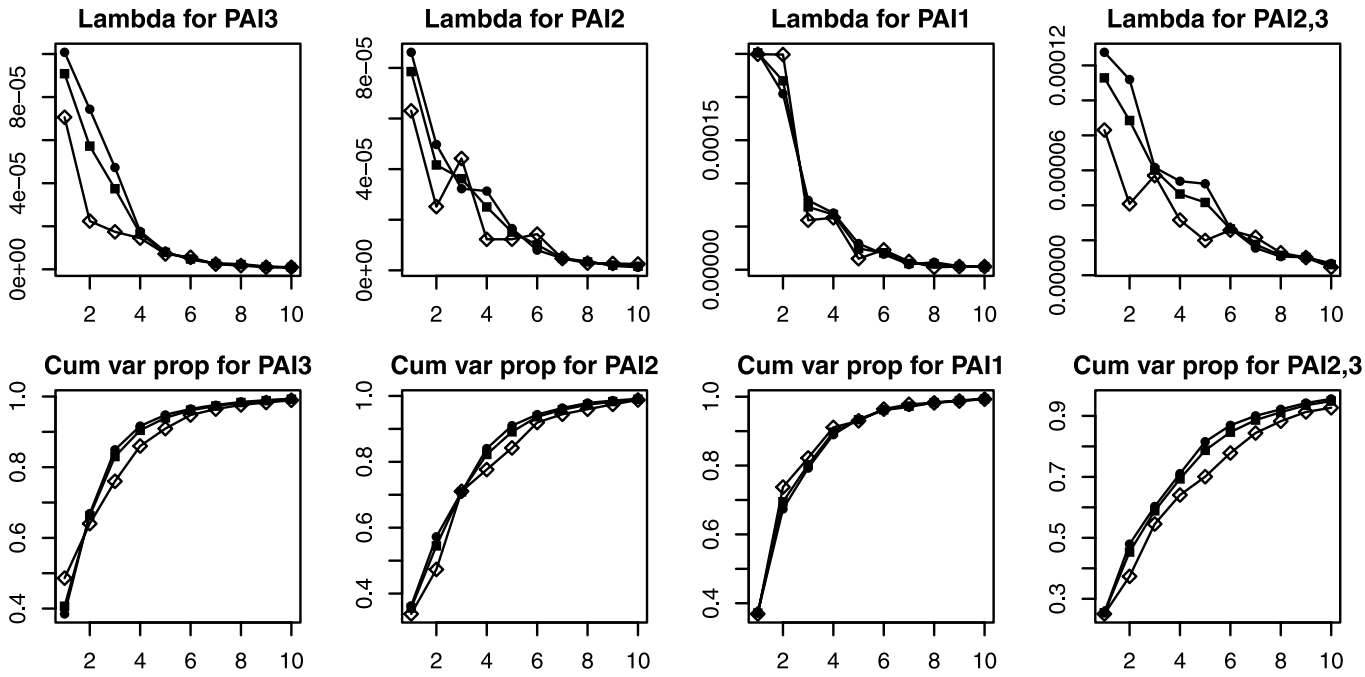

Figure 3. Empirical variances (scree plot) and cumulative proportions of variance explained by components for the TATA (circles) and CAP (diamonds) group and for both groups together (squares).

5.2 First-Order Inference

As was mentioned in the Introduction, a previous exploratory analysis of the data (Amzallag et al. 2006) that used clustering of the minicircles with respect to a Procrustean metric did not reveal any observable differences between the geometry of the two groups. The clustering distance used (a mean-square-based pairwise Procrustean distance) induces clustering with respect to the mean shape of the minicircles, which can be seen to be essentially identical between the two groups (Figure 2). To probe this finding more formally, we test the hypothesis of equal mean curves versus a general alternative, based on a variant of the test proposed by Berkes et al. (2009). We reject the hypothesis of equal mean curves when the value of the statistic

$$\sum_{j=1}^{K} \frac{n_1 n_2}{N} \frac{(\langle \overline{\mathbf{X}}, \widehat{\boldsymbol{\varphi}}_{XY}^{j,N} \rangle - \langle \overline{\mathbf{Y}}, \widehat{\boldsymbol{\varphi}}_{XY}^{j,N} \rangle)^2}{\widehat{\lambda}_{XY}^{j,N}}$$

is large compared to a χ^2_K distribution (the approximation employs results in Dauxois, Pousse, and Romain 1982). The results of this comparison are displayed in Table 2. The corresponding values of the test statistic are insignificant and one cannot reject the null hypothesis; indeed, the results of the test do not vary much with K.

Table 2. p-values for comparison of mean functions in the TATA and CAP group for various truncation levels K, for the full three-dimensional curves, and their projections onto the prinipal plane of inertia

K	PAI1, 2, 3	PAI2, 3
1	0.40	0.64
2	0.68	0.69
3	0.85	0.64
4	0.60	0.55
5	0.34	0.58
6	0.46	0.61

As discussed in the previous section, it seems, in fact, that the interesting "signal" of the minicircles is effectively planar (see Figure 2). It is, therefore, interesting to test the hypothesis that the mean function of the PAI1 coordinate is zero—for this will suggest that our analysis should concentrate on the prinicipal inertia plane (the projection of the Gaussian processes on this plane is obviously a Gaussian process). To this aim, we use the one-sample version of the test statistic used for mean comparison (which in the one-sample situation, is in fact an approximate likelihood-ratio statistic; Grenander 1981). For the TATA group the p-value of the test with $K = 4$ components is 0.29. For the CAP curves the p-value is 0.30 (also using four components). Hence the tests show no significant systematic deviation of the curves from the first principal plane, and their three-dimensional nature seems to only be due to random variation around a planar mean shape. For this reason, in the next section we concentrate on the comparison of the curves projected onto the principal plane of inertia.

5.3 Second-Order Inference

As the first-order comparison of the two minicircle groups did not reveal any significant differences, we turn our attention to the detection of second-order differences. Indeed, since the scientific hypothesis is that one type of curve (TATA) is more flexible, it may be intuitively expected that a detectable difference will lie in the covariance structure rather than the mean structure.

We test the hypothesis that both groups of curves share the same covariance operator by employing the test statistic T^*. The results are summarized in Table 3. Marginal tests on each

Table 3. p-values for the comparison of covariance functions in the TATA and CAP group on different principal inertia axes using the test statistic T^* under various truncation levels K

	p-value			
K	PAI3	PAI2	PAI1	PAI2, 3
1	0.252	0.313	0.976	0.167
2	0.001	0.118	0.823	0.005
3	0.000	0.087	0.782	0.025
4	0.001	0.022	0.886	0.051
5	0.001	0.053	0.555	0.009
6	0.010	0.087	0.327	0.005
7	0.019	0.098	0.360	0.023
8	0.046	0.173	0.148	0.094

inertia axis show that the covariance functions of the projections onto PAI3 seem significantly different for the two groups (with either the empirical selection $K = 4$ or the automatic choice $K = 5$). Differences of projections onto PAI2 appear marginally insignificant depending on the choice of K (the empirical choice is $K = 5$ and the automatic choice is $K = 7$). No significant difference is observed for PAI1, indicating that random deviations from the first principal plane may have the same covariance structure in the two groups (which is in keeping with our previous finding that the deviations from the principal plane can be thought to be residual). Since the curves appear to be planar on average, it is the covariance of their planar components where most structure is to be found. Indeed, when our test is carried out for the projection of the curves onto the principal plane of inertia using $K = 6$ (empirical) or $K = 7$ (automatic), it *rejects the null hypothesis of no flexibility differences, at the 1% and 3% significance levels, respectively*. In fact, the test based on T_1^* gives even more significant results, yielding a p-value that is numerically zero.

In the frequency domain, these differences can already be seen in the scree plots (Figure 3), where the TATA curves are seen to be more flexible in the sense that the variances of their Fourier coefficients are more inflated when compared to the CAP curves. Since the covariance kernels associated with the two operators under comparison are matrix-valued functions, there is no easy way to visualize the detected differences in the time domain. Figure 4 contains surface and contour plots of the empirical covariance kernels restricted to the third principal axis—the axis where the most significant differences were detected. The plot reveals differences both in terms of the norm as well as in terms of the structure.

6. CONCLUDING REMARKS

Motivated by the problem of comparison of groups of DNA minicircles, we introduce and study a testing procedure for two sample-comparison of Gaussian processes with respect to their covariance structure.

The proposed test function is based on an approximation of the Hilbert–Schmidt distance between the empirical covariance operators of the two groups, by means of the Karhunen–Loève representation of the pooled sample. The approximation

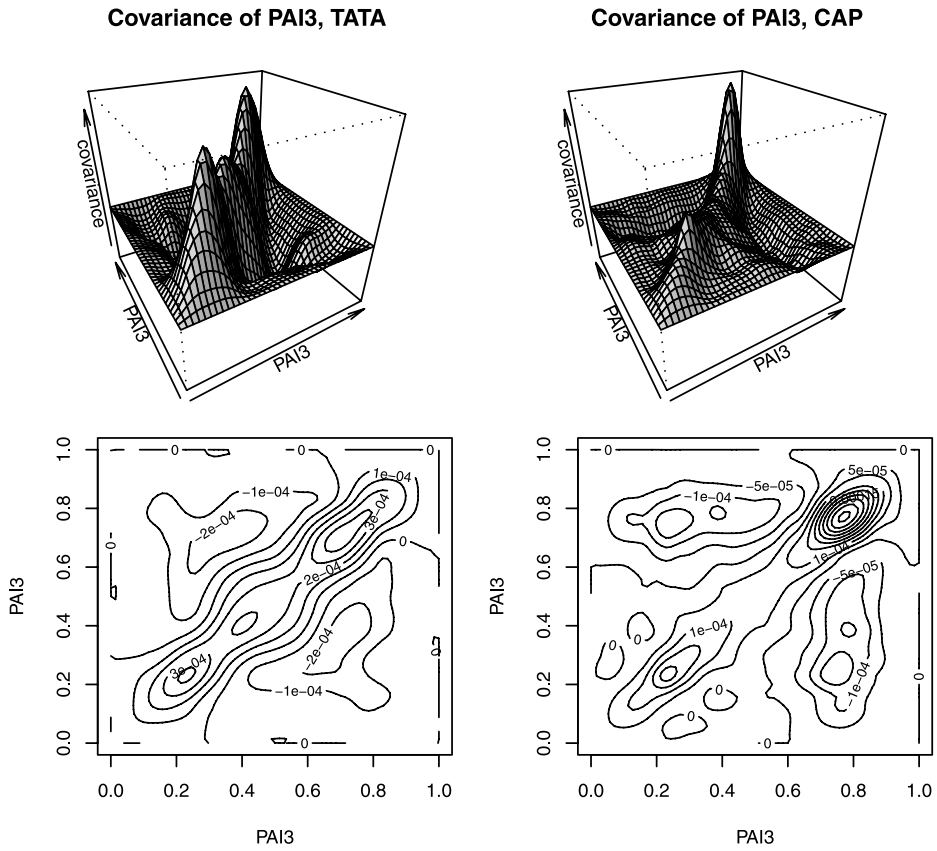

Figure 4. Surface and contour plots of the empirical covariance kernels corresponding to the TATA and CAP projections onto the third axis of inertia.

was seen to admit a *regularization* interpretation, the problem of testing presenting aspects of ill-posedness. The asymptotic distribution of the test function was established, and variance-stabilized variants with similar asymptotic properties were proposed. Finite-sample simulations under the null and various alternatives were used to investigate the performance of the proposed test. It should be noted that the results obtained readily extend to random functions defined over arbitrary compact Euclidean domains, and taking values in Euclidean spaces of arbitrary dimension (i.e., random fields).

The test was then carried out for a sample of 94 DNA minicircles of two different types. One type is believed to possess higher flexibility than the other, but this eluded empirical confirmation via electron microscopy. Our test rejected the hypothesis that the curves share the same covariance structure on their principal plane of inertia (the signals are essentially planar), providing support for the potential existence of differences between the geometry of the two groups. Interestingly, the difference was detected in the second-order characteristics, whereas previous analyses focused on first-order characteristics.

An important aspect of our testing procedure, as is the case with any spectral truncation regularization procedure, is the choice of truncation level K for the series representation of the Hilbert–Schmidt norm. A careless choice of truncation can affect the power of the test procedure. Our proposed approach for the choice of K was through visual inspection of functional PCA scree plots, combined with penalized prediction error minimization. Interesting further work will be to investigate LASSO-type component selection. Yet a further approach will be to consider *adaptive* modifications of the proposed tests that will automatically choose the level K based on the data; for example, tests based on statistics of the form $\max_K(T_N(K) - \beta K \log N)$, for some tuning parameter $\beta > 0$.

The asymptotic approximations for the distributions of the test statistics investigated hold for Gaussian processes. Departures from this assumption will affect the limiting law of the statistics. In simulations we observed that the test derived under the Gaussian assumption used in a non-Gaussian case becomes conservative when the scores have lighter tails than the normal distribution and anticonservative in the opposite case. Our tests are based on sums of squares of components which are asymptotically normal independent variables. When the data are not Gaussian, these components have asymptotically a multivariate normal distribution with unknown covariance structure. The limiting covariance matrix can be estimated and a chi-square test statistic can be based on the corresponding quadratic form (see also Horváth, Hušková, and Kokoszka 2010 for a similar approach in a different context). Some simulations showed that the convergence to the limiting distribution might be slow and one has to use only a small value of K, especially for the off-diagonal test.

Of course, testing whether a process is Gaussian is a research project in itself, but informal qq-plots constructed for the Karhunen–Loève coefficients of the minicircle data did not reveal any noteworthy departures from normality. For the benefit of the doubt, however, we also employed permutation tests based on our test statistics, with similar results—but with slightly more inflated p-values (Panaretos and Kraus 2009).

APPENDIX

Proof of Theorem 1

Introduce the notation $\mathscr{X}_i\mathbf{f} := \langle \mathbf{X}_i, \mathbf{f}\rangle \mathbf{X}_i$ and $\mathscr{Y}_i\mathbf{f} = \langle \mathbf{Y}_i, \mathbf{f}\rangle \mathbf{Y}_i$, so that $\widehat{\mathscr{R}}_X^n = n^{-1}\sum_i \mathscr{X}_i$ and $\widehat{\mathscr{R}}_Y^n = n^{-1}\sum_i \mathscr{Y}_i$. These are viewed as random elements of the Hilbert space of Hilbert–Schmidt operators acting on $\mathcal{L}^2[0,1]$. Under the hypothesis $H_0 : \mathscr{R}_X = \mathscr{R}_Y$, the collections $\{\mathscr{X}_i\}$ and $\{\mathscr{Y}_i\}$ are iid random operators with mean $\mathscr{R}_X = \mathscr{R}_Y$ and common covariance $\mathfrak{S} := \mathbb{E}[\mathscr{X}_i \otimes \mathscr{X}_i] - \mathscr{R}_X \otimes \mathscr{R}_X = \mathbb{E}[\mathscr{Y}_i \otimes \mathscr{Y}_i] - \mathscr{R}_Y \otimes \mathscr{R}_Y$, where $\otimes$ denotes the tensor product, $(u \otimes v)w = \langle v, w\rangle_{\mathcal{H}} u$ for any elements v, w, u of a Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$. In addition, our moment assumptions imply that $\mathbb{E}\|\mathscr{X}_i\|_{\mathrm{HS}}^2 < \infty$. We may, therefore, apply the Hilbert space central limit theorem (e.g., Bosq 2000, theorem 2.7) to conclude that

$$\sqrt{n_1}(\widehat{\mathscr{R}}_X^{n_1} - \mathscr{R}_X) \xrightarrow{w} \mathscr{Z}_1 \qquad \text{and}$$

$$\sqrt{n_2}(\widehat{\mathscr{R}}_Y^{n_2} - \mathscr{R}_Y) \xrightarrow{w} \mathscr{Z}_2 \quad \text{as } n_1, n_2 \to \infty,$$

where $\mathscr{Z}_1$ and $\mathscr{Z}_2$ are independent Gaussian random operators with mean 0 and covariance operator $\mathfrak{S}$. Now, given i, j, consider the sequence of random variables

$$W_N^{i,j} = \big\langle \sqrt{n_1 n_2/N}(\widehat{\mathscr{R}}_X^{n_1} - \widehat{\mathscr{R}}_Y^{n_2})\,\mathrm{sgn}[\langle \widehat{\boldsymbol{\varphi}}_{XY}^{i,N}, \boldsymbol{\varphi}_i\rangle]\widehat{\boldsymbol{\varphi}}_{XY}^{i,N}, \mathrm{sgn}[\langle \widehat{\boldsymbol{\varphi}}_{XY}^{j,N}, \boldsymbol{\varphi}_j\rangle]\widehat{\boldsymbol{\varphi}}_{XY}^{j,N}\big\rangle.$$

On the one hand, the strong law in Hilbert space implies that $\|\widehat{\mathscr{R}}_{XY}^N - \mathscr{R}_X\|_{\mathrm{HS}} \xrightarrow{\text{a.s.}} 0$ under the hypothesis H_0. Consequently, convergence also occurs with probability 1 in the strong operator topology, so that by Bosq (2000, lemma 4.3)

$$\big\| \mathrm{sgn}[\langle \widehat{\boldsymbol{\varphi}}_{XY}^{k,N}, \boldsymbol{\varphi}_k\rangle]\widehat{\boldsymbol{\varphi}}_{XY}^{k,N} - \widetilde{\boldsymbol{\varphi}}_k \big\|_{\mathcal{L}^2} \xrightarrow{\text{a.s.}} 0 \qquad \forall k \geq 1. \tag{A.1}$$

On the other hand, as $N \to \infty$ with $n_1/N \to \theta \in (0,1)$ we will have

$$\sqrt{\frac{n_2}{N}}\sqrt{n_1}\widehat{\mathscr{R}}_X^{n_1} - \sqrt{\frac{n_1}{N}}\sqrt{n_2}\widehat{\mathscr{R}}_Y^{n_2} \xrightarrow{w} \sqrt{1-\theta}\,\mathscr{Z}_1 - \sqrt{\theta}\,\mathscr{Z}_2 = \mathscr{Z}, \tag{A.2}$$

with $\mathscr{Z}$ a zero-mean Gaussian random operator with covariance $\mathfrak{S}$. Combining Equations (A.1) and (A.2) with the Hilbert space Slutsky lemma establishes that, for all $i, j \in \{1, \ldots, K\}$,

$$W_N^{i,j} \xrightarrow{w} \langle \mathscr{Z}\boldsymbol{\varphi}_i, \boldsymbol{\varphi}_j\rangle.$$

For the next step, we note that $\mathscr{Z}$, being a Gaussian process itself, also admits a Karhunen–Loève decomposition, with respect to the eigenfunctions of $\mathfrak{S}$. These eigenfunctions can be retrieved directly from the definition of $\mathfrak{S}$ and the Karhunen–Loève expansion of the typical X process, $\mathbf{X} = \sum_i \sqrt{\lambda_i}\xi_i\boldsymbol{\varphi}_i$. Defining the operator $\Phi_{ij}\mathbf{f} := \langle \boldsymbol{\varphi}_i, \mathbf{f}\rangle\boldsymbol{\varphi}_j$, we immediately see that $\mathscr{X} = \sum_{i,j}\sqrt{\lambda_i\lambda_j}\xi_i\xi_j\Phi_{ij}$ and $\mathscr{R}_X = \sum_j \lambda_j\Phi_{jj}$. Hence, upon recalling that the $\{\xi_i\}$ are an iid standard Gaussian array we may write

$$\begin{aligned}
\mathfrak{S} &= \mathbb{E}[\mathscr{X} \otimes \mathscr{X}] - \mathscr{R}_X \otimes \mathscr{R}_X \\
&= \sum_{i,j,q,p} \sqrt{\lambda_i\lambda_j\lambda_p\lambda_q}\,\mathbb{E}[\xi_i\xi_j\xi_p\xi_q]\Phi_{ij} \otimes \Phi_{qp} - \sum_{i,j}\lambda_i\lambda_j\Phi_{ii} \otimes \Phi_{jj} \\
&= \sum_{i\neq j}\lambda_i\lambda_j\Phi_{ii} \otimes \Phi_{jj} + \sum_{i\neq j}\lambda_i\lambda_j\Phi_{ij} \otimes \Phi_{ji} + \sum_{i\neq j}\lambda_i\lambda_j\Phi_{ij} \otimes \Phi_{ij} \\
&\quad + \sum_i 3\lambda_i^2\Phi_{ii} \otimes \Phi_{ii} - \sum_i \lambda_i^2\Phi_{ii} \otimes \Phi_{ii} - \sum_{i\neq j}\lambda_i\lambda_j\Phi_{ii} \otimes \Phi_{jj} \\
&= 2\sum_i \lambda_i^2\Phi_{ii} \otimes \Phi_{ii} + \sum_{i\neq j}\lambda_i\lambda_j(\Phi_{ij} \otimes \Phi_{ji} + \Phi_{ij} \otimes \Phi_{ij}),
\end{aligned}$$

since $\mathbb{E}[\xi_i\xi_j\xi_p\xi_q]$ is 1 whenever pairs of indices are equal but not all indices are totally coincident, 3 when all indices are equal, and zero

otherwise. Regrouping the summation by adding the terms that are symmetric with respect to their indices, we further obtain

$$\begin{aligned}\mathfrak{S} &= 2\sum_i \lambda_i^2 \Phi_{ii}\otimes\Phi_{ii} \\ &\quad + \sum_{i<j}\lambda_i\lambda_j(\Phi_{ij}\otimes\Phi_{ji}+\Phi_{ij}\otimes\Phi_{ij}+\Phi_{ji}\otimes\Phi_{ij}+\Phi_{ji}\otimes\Phi_{ji}) \\ &= 2\sum_i \lambda_i^2 \Phi_{ii}\otimes\Phi_{ii} \\ &\quad + \sum_{i<j}\lambda_i\lambda_j\{\Phi_{ij}\otimes(\Phi_{ij}+\Phi_{ji})+\Phi_{ji}\otimes(\Phi_{ij}+\Phi_{ji})\} \\ &= \sum_i(\sqrt{2}\lambda_i)^2\Phi_{ii}\otimes\Phi_{ii}+\sum_{i<j}\lambda_i\lambda_j(\Phi_{ij}+\Phi_{ji})\otimes(\Phi_{ij}+\Phi_{ji}).\end{aligned}$$

It is straightforward to verify that $\{\Phi_{ij}+\Phi_{ji}\}_{i<j}\cup\{\Phi_{ii}\}_{i\geq 1}$ constitutes a complete orthogonal system of operators for the Hilbert space of Hilbert–Schmidt operators acting on $\mathcal{L}^2[0,1]$. We may, therefore, represent $\mathscr{Z}$ in a Karhunen–Loève expansion as

$$\mathscr{Z} = \sqrt{2}\sum_i \lambda_i\zeta_{ii}\Phi_{ii}+\sum_{i<j}\lambda_i^{1/2}\lambda_j^{1/2}\zeta_{ij}(\Phi_{ij}+\Phi_{ji})$$

for $\{\zeta_{ij}\}_{i,j=1}^{\infty}$ an iid array of standard Gaussian variables. Consequently, we may express the Gaussian process $\mathscr{Z}\boldsymbol{\varphi}_k$ as

$$\begin{aligned}\mathscr{Z}\boldsymbol{\varphi}_k &= \sqrt{2}\sum_{i=1}^{\infty}\lambda_i\zeta_{ii}\langle\boldsymbol{\varphi}_i,\boldsymbol{\varphi}_k\rangle\boldsymbol{\varphi}_i \\ &\quad + \sum_{i<j}\lambda_i^{1/2}\lambda_j^{1/2}\zeta_{ij}(\langle\boldsymbol{\varphi}_i,\boldsymbol{\varphi}_k\rangle\boldsymbol{\varphi}_j+\langle\boldsymbol{\varphi}_j,\boldsymbol{\varphi}_k\rangle\boldsymbol{\varphi}_i) \\ &= \sqrt{2}\lambda_k\zeta_{kk}\boldsymbol{\varphi}_k+\sum_{i<j}\lambda_i^{1/2}\lambda_j^{1/2}\zeta_{ij}\langle\boldsymbol{\varphi}_i,\boldsymbol{\varphi}_k\rangle\boldsymbol{\varphi}_j \\ &\quad + \sum_{i<j}\lambda_i^{1/2}\lambda_j^{1/2}\zeta_{ij}\langle\boldsymbol{\varphi}_j,\boldsymbol{\varphi}_k\rangle\boldsymbol{\varphi}_i \\ &= \sqrt{2}\lambda_k\zeta_{kk}\boldsymbol{\varphi}_k+\sum_{k<j}\lambda_k^{1/2}\lambda_j^{1/2}\zeta_{kj}\boldsymbol{\varphi}_j+\sum_{i<k}\lambda_i^{1/2}\lambda_k^{1/2}\zeta_{ik}\boldsymbol{\varphi}_i,\end{aligned}$$

where we used the fact that $\{\boldsymbol{\varphi}_i\}$ is an orthonormal system. It follows that for arbitrary $k,n\in\{1,\ldots,K\}$, the random variable $\langle\mathscr{Z}\boldsymbol{\varphi}_k,\boldsymbol{\varphi}_n\rangle$ admits the representation

$$\begin{aligned}\langle\mathscr{Z}\boldsymbol{\varphi}_k,\boldsymbol{\varphi}_n\rangle &= \sqrt{2}\lambda_k\zeta_{kk}\langle\boldsymbol{\varphi}_k,\boldsymbol{\varphi}_n\rangle+\sum_{k<j}\lambda_k^{1/2}\lambda_j^{1/2}\zeta_{kj}\langle\boldsymbol{\varphi}_j,\boldsymbol{\varphi}_n\rangle \\ &\quad + \sum_{i<k}\lambda_i^{1/2}\lambda_k^{1/2}\zeta_{ik}\langle\boldsymbol{\varphi}_i,\boldsymbol{\varphi}_n\rangle \\ &= \begin{cases}\sqrt{2}\lambda_k\zeta_{kk} & \text{if } k=n\\ \lambda_k^{1/2}\lambda_n^{1/2}\zeta_{kn} & \text{if } k<n\\ \lambda_k^{1/2}\lambda_n^{1/2}\zeta_{nk} & \text{if } k>n.\end{cases}\end{aligned}$$

It follows that $\langle\mathscr{Z}\boldsymbol{\varphi}_k,\boldsymbol{\varphi}_k\rangle\overset{\text{iid}}{\sim}\mathcal{N}(0,2\lambda_k^2)$ independently of $\langle\mathscr{Z}\boldsymbol{\varphi}_m,\boldsymbol{\varphi}_n\rangle\overset{\text{iid}}{\sim}\mathcal{N}(0,\lambda_m\lambda_n)$, $m\neq n$. Consequently, we have

$$\frac{1}{2}\frac{\langle\mathscr{Z}\boldsymbol{\varphi}_k,\boldsymbol{\varphi}_k\rangle^2}{\lambda_k^2}\overset{\text{iid}}{\sim}\chi_1^2,$$

independently of

$$\frac{1}{2}\frac{\langle\mathscr{Z}\boldsymbol{\varphi}_m,\boldsymbol{\varphi}_n\rangle^2+\langle\mathscr{Z}\boldsymbol{\varphi}_n,\boldsymbol{\varphi}_m\rangle^2}{\lambda_m\lambda_n}=\frac{\langle\mathscr{Z}\boldsymbol{\varphi}_m,\boldsymbol{\varphi}_n\rangle^2}{\lambda_m\lambda_n}\sim\chi_1^2.$$

The continuous mapping theorem now implies that

$$\begin{aligned}\frac{1}{2}\frac{(W_N^{ij})^2+(W_N^{ji})^2}{\lambda_i\lambda_j} &= \frac{n_1n_2}{2N}\sum_{i=1}^{K}\sum_{j=1}^{K}\frac{\langle(\widehat{\mathscr{R}}_X^{n_1}-\widehat{\mathscr{R}}_Y^{n_2})\widehat{\boldsymbol{\varphi}}_{XY}^{i,N},\widehat{\boldsymbol{\varphi}}_{XY}^{j,N}\rangle^2}{\lambda_i\lambda_j} \\ &\xrightarrow{w}\chi^2_{K(K+1)/2}.\end{aligned}$$

To complete the proof, we note that

$$\frac{n_1}{N}\widehat{\lambda}_{X,XY}^{k,n_1}+\frac{n_2}{N}\widehat{\lambda}_{Y,XY}^{k,n_2}\xrightarrow{p}\theta\lambda_k+(1-\theta)\lambda_k=\lambda_k\quad\forall k\in\{1,\ldots,K\},$$

so that the result follows from the application of Slutsky's lemma.

SUPPLEMENTAL MATERIALS

Additional plots and tables and detailed study: Additional plots and tables are available in a supplementary file. In addition, the supplementary file contains a more detailed study of the problem of comparing the complete spectrum, extending the discussion in the last part of Section 3.2. (Supplement.pdf)

[Received April 2009. Revised December 2009.]

REFERENCES

Adler, R. J. (1990), *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes. Lecture Notes and Monographs Series*, Hayward: Institute of Mathematical Statistics. [673]

Amzallag, A., Vaillant, C., Jacob, M., Unser, M., Bednar, M., Kahn, J. D., Dubochet, J., Stasiak, A., and Maddocks, J. H. (2006), "3D Reconstruction and Comparison of Shapes of DNA Minicircles Observed by Cryo-Electron Microscopy," *Nucleic Acids Research*, 34 (18), e125. [670,678]

Arnold, V. I. (1989), *Mathematical Methods of Classical Mechanics*, New York: Springer. [671]

Benko, M., Härdle, W., and Kneip, A. (2009), "Common Functional Principal Components," *The Annals of Statistics*, 37, 1–34. [670]

Berkes, I., Gabrys, R., Horváth, L., and Kokoszka, P. (2009), "Detecting Changes in the Mean of Functional Observations," *Journal of the Royal Statistical Society, Ser. B*, 71, 927–946. [670,678]

Bosq, D. (2000), *Linear Processes in Function Spaces*, New York: Springer. [675,680]

Cardot, H., Ferraty, F., Mas, A., and Sarda, P. (2003), "Testing Hypotheses in the Functional Linear Model," *Scandinavian Journal of Statistics*, 30 (1), 241–255. [670]

Cuevas, A., Febrero, M., and Fraiman, R. (2004), "An ANOVA Test for Functional Data," *Computational Statistics and Data Analysis*, 47, 111–122. [670]

Dauxois, J., Pousse, A., and Romain, Y. (1982), "Asymptotic Theory for the Principal Component Analysis of a Random Vector Function: Some Applications to Statistical Inference," *Journal of Multivariate Analysis*, 12, 136–154. [673,678]

Fan, J., and Lin, S.-K. (1998), "Tests of Significance When the Data Are Curves," *Journal of the American Statistical Association*, 93, 1007–1021. [670]

Ferraty, F., and Vieu, P. (2006), *Nonparametric Functional Data Analysis*, New York: Springer. [672]

Gabrys, R., and Kokoszka, P. (2007), "Portmanteau Test of Independence for Functional Observations," *Journal of the American Statistical Association*, 102, 1338–1348. [670]

Gasser, T., and Kneip, A. (1995), "Searching for Structure in Curve Samples," *Journal of the American Statistical Association*, 90, 1179–1188. [671]

Gervini, D. (2008), "Robust Functional Estimation Using the Median and Spherical Principal Components," *Biometrika*, 95 (3), 587–600. [672]

Gervini, D., and Gasser, T. (2004), "Self-Modelling Warping Functions," *Journal of the Royal Statistical Society, Ser. B*, 66 (4), 959–971. [671]

Giri, N. (1968), "On Tests of the Equality of Two Covariance Matrices," *The Annals of Mathematical Statistics*, 39, 275–277. [673]

Grenander, U. (1981), *Abstract Inference*, New York: Wiley. [670,678]

Hagerman, P. J. (1988), "Flexibility of DNA," *Annual Review Biophysics and Biophysical Chemistry*, 17, 265–286. [670]

Hall, P., and Hosseini-Nassab, M. (2006), "On Properties of Functional Principal Components Analysis," *Journal of the Royal Statistical Society, Ser. B*, 68 (1), 109–126. [673]

Hall, P., and Van Keilegom, I. (2007), "Two Sample Tests in Functional Data Analysis Starting From Discrete Data," *Statistica Sinica*, 17, 1511–1531. [670]

Horváth, L., Hušková, M., and Kokoszka, P. (2010), "Testing the Stability of the Functional Autoregressive Process," *Journal of Multivariate Analysis*, 101 (2), 352–367. [670,680]

Jacob, M., Blu, T., Vaillaint, C., Maddocks, J. H., and Unser, M. (2006), "3-D Shape Estimation of DNA Molecules From Stereo Cryo-Electron Micrographs Using a Projection Steerable Snake," *IEEE Transations on Image Processing*, 15 (1), 214–227. [671]

Kiefer, J., and Schwartz, R. (1965), "Admissible Bayes Character of T^2-Test, R^2-Test, and Other Fully Invariant Tests for Classical Multivariate Normal Problems," *The Annals of Mathematical Statistics*, 36, 747–770. [673]

Ledwina, T. (1994), "Data-Driven Version of Neyman's Smooth Test of Fit," *Journal of the American Statistical Association*, 89, 1000–1005. [677]

Panaretos, V. M., and Kraus, D. (2009), "Second Order Comparison of Gaussian Processes With Applications to DNA Shape Analysis," Technical Report 01-09, Chair of Mathematical Statistics, EPFL. [680]

Pillai, K. C. S. (1955), "Some New Test Criteria in Multivariate Analysis," *The Annals of Mathematical Statistics*, 26, 117–121. [673]

Ramsay, J. O., and Silverman, B. W. (2002), *Applied Functional Data Analysis: Methods and Case Studies*, New York: Springer. [673]

——— (2005), *Functional Data Analysis*, New York: Springer. [672,673,675]

Rice, J., and Silverman, B. W. (1991), "Estimating the Mean and Covariance Structure Nonparametrically When the Data Are Curves," *Journal of the Royal Statistical Society, Ser. B*, 53, 233–243. [676]

Roy, S. N. (1953), "On a Heuristic Method of Test Construction and Its Use in Multivariate Analysis," *The Annals of Mathematical Statistics*, 24, 220–238. [673]

Shen, Q., and Faraway, J. (2004), "An *F* Test for Linear Models With Functional Responses," *Statistica Sinica*, 14, 1239–1257. [670]

Tang, R., and Müller, H. G. (2008), "Pairwise Curve Synchronization for Functional Data," *Biometrika*, 95 (4), 875–889. [671]

Tolstorukov, M. Y., Virnik, K. M., Adhya, S., and Zhurkin, V. B. (2005), "A-Tract Clusters May Facilitate DNA Packaging in Bacterial Nucleoid," *Nucleic Acids Research*, 33 (12), 3907–3918. [670]

Vilar, J. M. G., and Leibler, S. (2003), "DNA Looping and Physical Constraints on Transcription Regulation," *Journal of Molecular Biology*, 331 (5), 981–989. [670]

Yao, F., Müller, H. G., and Wang, J. L. (2005a), "Functional Data Analysis of Sparse Longitudinal Data," *Journal of the American Statistical Association*, 100, 577–590. [676]

——— (2005b), "Functional Linear Regression Analysis for Longitudinal Data," *The Annals of Statistics*, 33, 2873–2903. [676]

Supplemental File: Second–Order Comparison of Gaussian Random Functions and the Geometry of DNA Minicircles

This supplementary note contains additional plots and tables in Section 1. In addition, Section 2 contains a more detailed study of the problem of comparing the complete spectrum, extending the discussion in the last part of Section 3.2 in the main body of the paper.

1 Supplementary Figures and Tables

This section contains figures and a table not presented in the main body of the paper. The first two figures contain plots of the projected aligned curves onto their principal axes of inertia, including their superimposition. The third figure contains scree plots with respect to the mixed eigenbasis for the two groups separately, as well as jointly. The last figure depicts the Normal QQ plots of the Karhunen-Loève residuals, as described in the discussion section of the paper.

Finally, a complete table containing the results of the simulations for level and power corresponding to Section 4 is also given. In addition to the main test statistic proposed in the paper, the complete table also presents simulations for the diagonal form of the statistic (which compares only the eigenvalues). It is observed that when the difference lies only in the eigenvalues, this test statistic performs more powerfully, as would be expected. However, in the cases where differences also lie in the eigenfunctions, it is outperformed by the full version of the test statistic.

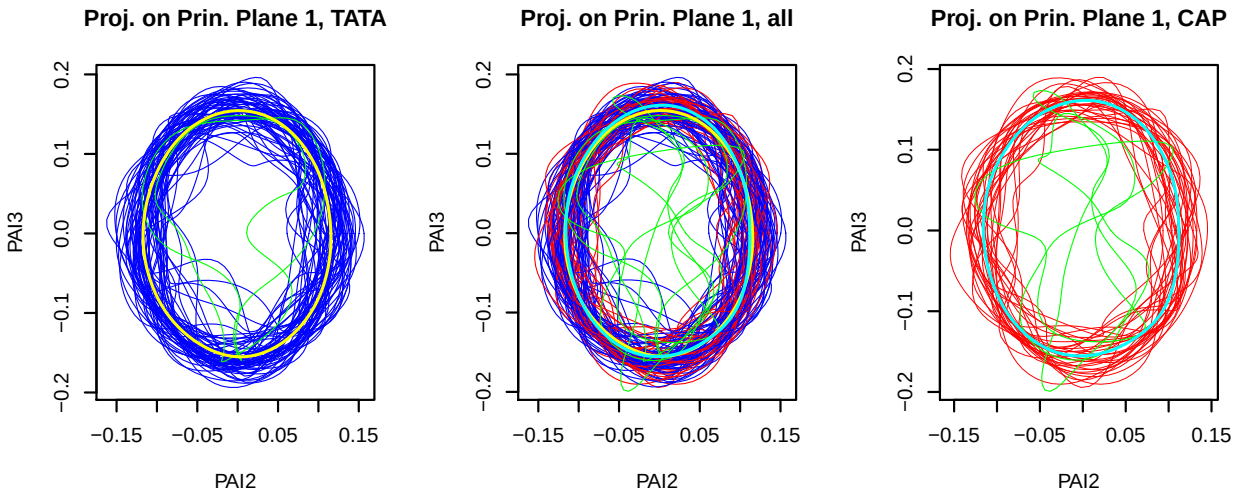

Figure 1: Projection of DNA curves on the first principal plane. Five removed outlying observations plotted in green. Mean curves (yellow and cyan) computed without outlying observations.

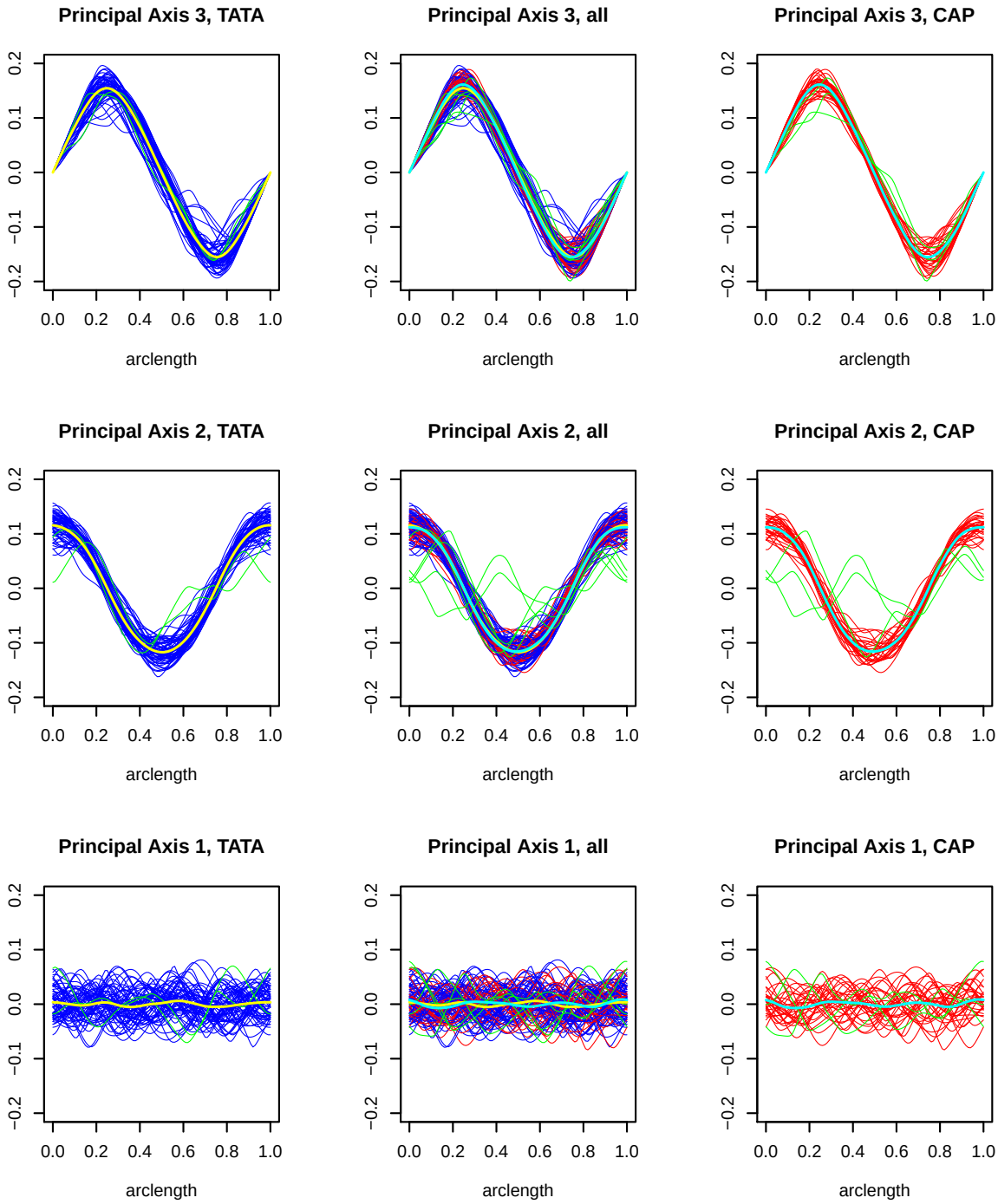

Figure 2: Coordinates of DNA curves on the principal axes of inertia. Five removed outlying observations plotted in green. Mean curves (yellow and cyan) computed without outlying observations.

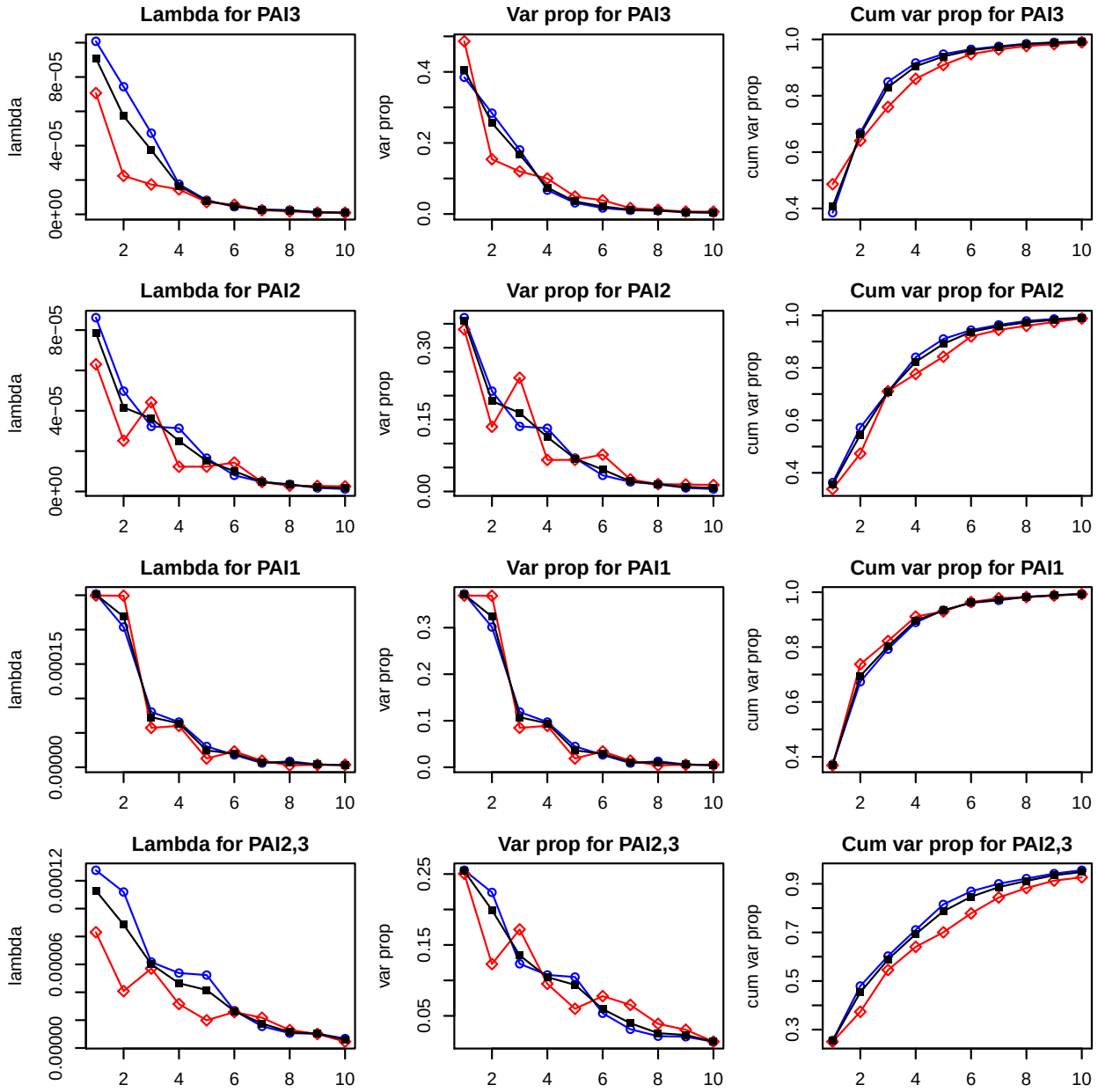

Figure 3: Empirical variances (scree plot), proportions and cumulative proportions of variance explained by components for the TATA (blue lines with circles) and CAP (red with diamonds) group and for both groups together (black with squares).

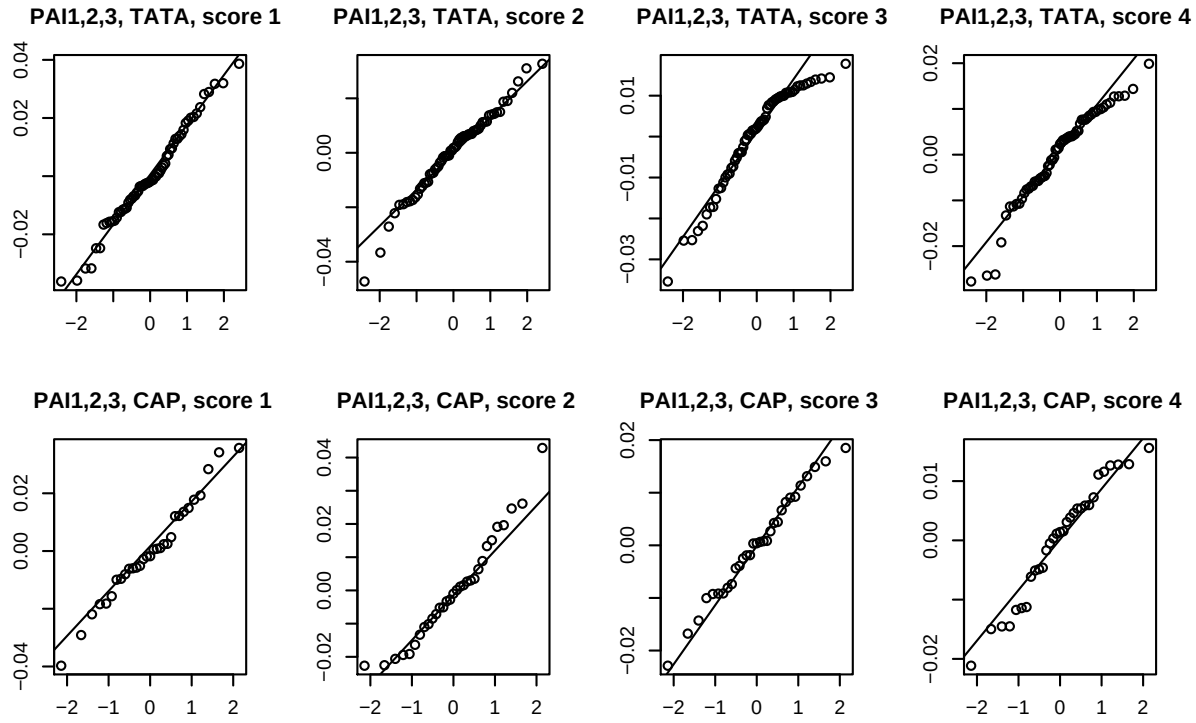

Figure 4: QQ plots corresponding to the centred Fourier coefficients when projecting onto the first four empirical eigenfunctions for each sample of curves, respectively. The exact distribution of these quantities will not be Gaussian, even if the processes are Gaussian. However, asymptotically, their distribution will be Gaussian. There do not appear systematic deviations, except for the plot corresponding to the third Fourier coefficient in the TATA group, which seems to suggest lighter upper tails as compared to the Gaussian.

Table 1: Empirical rejection probabilities on the nominal level 5 %, sample size $n_1 = n_2 = 50$, number of replications 5000 for A, 1000 for B–I. Here, $\boldsymbol{u}^X = (\boldsymbol{v}^X, \boldsymbol{w}^X)$ (resp. $\boldsymbol{u}^Y$) and K^* is the automatic truncation choice given by the penalised fit criterion.

			K				
	Parameters	Test	1	2	3	4	K^*
A	$\boldsymbol{u}^X = (12, 7, 0.5, 9, 5, 0.3)$	T	0.045	0.049	0.044	0.044	0.047
	$\boldsymbol{u}^Y = (12, 7, 0.5, 9, 5, 0.3)$	T^*	0.051	0.056	0.057	0.056	0.059
		T_1	0.045	0.046	0.045	0.047	0.047
		T_1^*	0.051	0.054	0.056	0.061	0.061
B	$\boldsymbol{u}^X = (14, 7, 0.5, 6, 5, 0.3)$	T	0.422	0.264	0.185	0.150	0.148
	$\boldsymbol{u}^Y = (8, 7, 0.5, 6, 5, 0.3)$	T^*	0.443	0.315	0.223	0.174	0.175
		T_1	0.422	0.317	0.265	0.219	0.222
		T_1^*	0.443	0.350	0.306	0.267	0.267
C	$\boldsymbol{u}^X = (15, 10, 0.5, 4, 3, 0.3)$	T	0.186	0.331	0.218	0.169	0.167
	$\boldsymbol{u}^Y = (11, 6, 0.5, 4, 3, 0.3)$	T^*	0.201	0.366	0.269	0.207	0.208
		T_1	0.186	0.380	0.312	0.279	0.273
		T_1^*	0.201	0.420	0.358	0.317	0.314
D	$\boldsymbol{u}^X = (12, 7, 0.5, 9, 3, 0.3)$	T	0.040	0.204	0.836	0.973	0.962
	$\boldsymbol{u}^Y = (12, 7, 0.5, 2, 5, 0.3)$	T^*	0.047	0.221	0.848	0.984	0.980
		T_1	0.040	0.202	0.766	0.803	0.799
		T_1^*	0.047	0.217	0.783	0.822	0.820
E	$\boldsymbol{u}^X = (12, 7, 0.5, 9, 3, 0.3)$	T	0.047	0.246	0.644	0.964	0.962
	$\boldsymbol{u}^Y = (12, 7, 0.5, 3, 9, 0.3)$	T^*	0.055	0.267	0.686	0.976	0.975
		T_1	0.047	0.227	0.477	0.597	0.594
		T_1^*	0.055	0.250	0.509	0.620	0.617
F	$\boldsymbol{u}^X = \boldsymbol{u}^Y = (12, 7, 4, 0.5, 0.3, 0.1)$	T	0.257	0.693	0.909	1.000	1.000
	$\boldsymbol{\delta}^X = (0.15, 0.15, 0.15)$	T^*	0.273	0.706	0.916	1.000	1.000
		T_1	0.257	0.474	0.521	0.567	0.637
		T_1^*	0.273	0.496	0.544	0.594	0.655
G	$\boldsymbol{u}^X = (12, 7, 0.5, 8, 6, 0.3)$	T	0.042	0.040	0.054	1.000	1.000
	$\boldsymbol{u}^Y = (12, 7, 0.5, 8, 0, 0.3)$	T^*	0.047	0.048	0.068	1.000	1.000
		T_1	0.042	0.047	0.051	1.000	1.000
		T_1^*	0.047	0.061	0.062	1.000	1.000
H	$\boldsymbol{u}^X = (12, 7, 0.5, 9, 5, 0.3)$	T	0.044	0.140	0.500	1.000	1.000
	$\boldsymbol{u}^Y = (12, 7, 0.5, 0, 5, 0.3)$	T^*	0.049	0.154	0.520	1.000	1.000
		T_1	0.044	0.139	0.478	0.992	0.992
		T_1^*	0.049	0.155	0.497	0.993	0.993
I	Brownian motion versus	T	0.719	0.608	0.483	0.377	0.493
	Ornstein–Uhlenbeck process	T^*	0.731	0.644	0.532	0.443	0.546
		T_1	0.719	0.627	0.547	0.476	0.551
		T_1^*	0.731	0.666	0.596	0.542	0.595

2 Comparing the Full Spectrum

The test procedure developed in the paper employs an optimal finite dimensional reduction in order to regularise the problem of testing. This is motivated by a Parseval decomposition of the Hilbert-Schmidt distance between the two operators,

$$\|\mathscr{R}_X - \mathscr{R}_Y\|_{HS}^2 = \sum_{k=1}^{K} \| \left(\mathscr{R}_X - \mathscr{R}_Y\right) \boldsymbol{\varphi}_{XY}^k \|_{\mathcal{L}^2}^2 + \epsilon,$$

where ϵ can be made arbitrarily small by appropriate choice of K. By making such a choice, the statistic will be (eventually) able to detect departures from the null hypothesis unless one operator is contained within a ball of small radius centred at the other operator; in this latter case, the test will still be able to detect the difference (eventually), except if this small difference lies completely at the high frequency end of the spectrum (in which case, for all practical purposes, the difference is irrelevant).

We are willing to tolerate this small level of "bias", in order to control the overall type II error of the problem. Comparison of the higher order terms of the operator spectrum on the basis of a finite sample is an ill-defined estimation problem: the fast decay of the spectrum means that we are attempting to compare extremely small quantities that have variance roughly proportional to their magnitude. In addition, the estimators of higher order eigenfunction will be characterised by very large integrated mean squared errors (available bounds grow for fixed N depending inversely on the rate of decay of the spectrum). Therefore, by trying to increase K in order to eliminate the small type II error introduced by the truncation, we are in effect causing an overall blow-up of the type II error.

If one nevertheless wishes to compare even the finest differences in the spectrum, then one needs to let K grow to infinity along with N, $K = K_N$ and modify the test statistic so as to obtain a Gaussian limit. Regularisation now manifests itself by the imposition of an allowed rate of growth of K_N. That is, a rate of growth of K relative to N that does not

allow overwhelming instabilities due to the growing K. As one might expect, this growth will depend inversely on the rate of decay of the true eigenvalues (a lot of data is required to compare the finest details of the two procsses). Inevitably, in fact, this rate will be rather slow due to the following:

(a) Although the truncation level will grow as K_N, the number of terms being compared is K_N^2.

(b) While these K^2 summation terms do become independent as N grows (allowing for a CLT phenomenon) no mixing concept applies. In effect, this means that one has to look at the convergence in distribution to independence of a random vector of increasing dimension ($= K_N^2$). For any fixed dimension, the weak convergence will be at a rate of $N^{-1/2}$. Therefore, if one wishes to use L^p norms in order to use the Hilbert structure of the problem, K_N must grow slow enough to allow the $N^{-1/2}$ rate to compensate for the K_N^2 rate of increase in dimension.

(c) This required "global convergence" to independence is regulated by the convergence of the empirical eigenfunctions to the true ones; this in turn depends on the spacings between the true eigenvalues: the rate of convergence of the Kth empirical eigenfunction behaves like $N^{-1/2}\lambda_K^{-1}$. Therefore, when we let K grow, it has to be at rate slow enough, to allow $N^{-1/2}$ to annihilate the blow-up of the inverse eigenvalues.

The above heuristics are made precise in the proof of the next theorem, which provides a sufficient *regularisation rate* for asymptotically comparing the whole spectrum of infinite rank processes.

Theorem 1. *Let* $\{\boldsymbol{X}_n\}_{n=1}^{n_1}$ *and* $\{\boldsymbol{Y}_n\}_{n=1}^{n_2}$ *be two collections of zero mean iid continuous Gaussian random functions indexed by the interval* $[0,1]$ *and taking values in* $\mathbb{R}^d$*, possessing covariance operators* $\mathscr{R}_X$ *and* $\mathscr{R}_Y$*. Suppose that both operators are of infinite rank and have distinct eigenvalues. Let* $\widehat{\mathscr{R}}_X^{n_1}$ *and* $\widehat{\mathscr{R}}_Y^{n_2}$ *denote the empirical covariance operators based on*

$\{\boldsymbol{X}_n\}_{n=1}^{n_1}$ *and* $\{\boldsymbol{Y}_n\}_{n=1}^{n_2}$. *For* $N = n_1 + n_2$, *let* $\widehat{\mathscr{R}}_{XY}^N$ *denote the empirical covariance operator of the pooled collection, and* $\{\hat{\boldsymbol{\varphi}}_{XY}^{k,N}\}_{k=1}^N$ *the corresponding eigenfunctions. Finally, let* $\hat{\lambda}_{X,XY}^{k,n_1}$, $\hat{\lambda}_{Y,XY}^{k,n_2}$ *denote the empirical variance of the* k*th Fourier coefficient of* $\{\boldsymbol{X}_n\}_{n=1}^{n_1}$ *and* $\{\boldsymbol{Y}_n\}_{n=1}^{n_2}$, *respectively, with respect to the eigenfunctions* $\{\hat{\boldsymbol{\varphi}}_{XY}^{n,K}\}_{n=1}^N$. *Assuming that* $\mathbb{E}[\|\boldsymbol{X}_1\|_{L^2}^4] < \infty$, $\mathbb{E}[\|\boldsymbol{Y}_1\|_{L^2}^4] < \infty$, *and* $n_1/N \to \theta \in (0,1)$ *as* $N = n_1 + n_2 \to \infty$, *it follows that, under the hypothesis* $H_0 : \mathscr{R}_X = \mathscr{R}_Y$,

$$S_N := \frac{n_1 n_2}{2N\sqrt{K_N(K_N+1)/2}} \sum_{i=1}^{K_N}\sum_{j=1}^{K_N} \left\langle (\widehat{\mathscr{R}}_X^{n_1} - \widehat{\mathscr{R}}_Y^{n_2})\check{\boldsymbol{\varphi}}_{XY}^{i,N}, \check{\boldsymbol{\varphi}}_{XY}^{j,N} \right\rangle^2 - \sqrt{\frac{K_N(K_N+1)}{2}} \xrightarrow{w} \mathcal{N}(0,1),$$

as $N \to \infty$, *for any* $K_N \uparrow \infty$ *such that* $K_N^7 \lambda_{3K_N/2}^{-3/2} = o(\sqrt{N})$, *where*

$$\check{\boldsymbol{\varphi}}_{XY}^{k,N} = \frac{\hat{\boldsymbol{\varphi}}_{XY}^{k,N}}{\sqrt{\frac{n_1}{N}\hat{\lambda}_{X,XY}^{k,n_1} + \frac{n_2}{N}\hat{\lambda}_{Y,XY}^{k,n_2}}}.$$

Proof of Theorem 2. Let $\{Z_{Nk}\}$ denote the triangular array of random variables defined as

$$Z_{Nk} := \frac{1}{\sqrt{K_N(K_N+1)/2}} \left(\frac{n_1 n_2}{N} \left\langle (\widehat{\mathscr{R}}_X^{n_1} - \widehat{\mathscr{R}}_Y^{n_2})\check{\boldsymbol{\varphi}}_{XY}^{i(k),N}, \check{\boldsymbol{\varphi}}_{XY}^{j(k),N} \right\rangle^2 - 1 \right), \qquad i(k) \neq j(k)$$

and

$$Z_{Nk} := \frac{1}{\sqrt{K_N(K_N+1)/2}} \left(\frac{n_1 n_2}{2N} \left\langle (\widehat{\mathscr{R}}_X^{n_1} - \widehat{\mathscr{R}}_Y^{n_2})\check{\boldsymbol{\varphi}}_{XY}^{i(k),N}, \check{\boldsymbol{\varphi}}_{XY}^{i(k),N} \right\rangle^2 - 1 \right), \qquad \text{otherwise,}$$

where $(i(k), j(k))$ is the the kth element of the index array $\{(i,j) : i \leq j \leq K_N\}$, when enumerating row-wise. Clearly, for $\kappa_N = K_N(K_N+1)/2$,

$$S_N = \sum_{k=1}^{\kappa_N} Z_{Nk}.$$

Write $\mathscr{Z}_N := (n_1 n_2/N)^{1/2}(\widehat{\mathscr{R}}_X^{n_1} - \widehat{\mathscr{R}}_Y^{n_2})$ and define

$$\widetilde{Z}_{Nk} := \sqrt{\frac{n_1 n_2}{N}} \left\langle (\widehat{\mathscr{R}}_X^{n_1} - \widehat{\mathscr{R}}_Y^{n_2}) \mathrm{sgn}[\langle \check{\boldsymbol{\varphi}}_{XY}^{i(k),N}, \check{\boldsymbol{\varphi}}_{i(k)} \rangle] \check{\boldsymbol{\varphi}}_{XY}^{i(k),N}, \mathrm{sgn}[\langle \check{\boldsymbol{\varphi}}_{XY}^{j(k),N}, \check{\boldsymbol{\varphi}}_{j(k)} \rangle] \check{\boldsymbol{\varphi}}_{XY}^{j(k),N} \right\rangle, \quad i(k) \neq j(k)$$

and

$$\widetilde{Z}_{Nk} := \sqrt{\frac{n_1 n_2}{2N}} \left\langle (\widehat{\mathscr{R}}_X^{n_1} - \widehat{\mathscr{R}}_Y^{n_2}) \mathrm{sgn}[\langle \check{\boldsymbol{\varphi}}_{XY}^{i(k),N}, \check{\boldsymbol{\varphi}}_{i(k)} \rangle] \check{\boldsymbol{\varphi}}_{XY}^{i(k),N}, \mathrm{sgn}[\langle \check{\boldsymbol{\varphi}}_{XY}^{i(k),N}, \check{\boldsymbol{\varphi}}_{i(k)} \rangle] \check{\boldsymbol{\varphi}}_{XY}^{i(k),N} \right\rangle, \quad \text{otherwise,}$$

where we use the notation $\check{\boldsymbol{\varphi}}_k := \lambda_k^{-\frac{1}{2}} \boldsymbol{\varphi}_k$. The corresponding natural filtration is denoted by $\mathcal{F}_{N,k} := \sigma(\widetilde{Z}_{Nm}; 1 \leq m \leq k)$, and notice that $\{Z_{Nk}\}$ is also adapted to the filtration $\{\mathcal{F}_{N,k}\}$. Finally, we will write $\boldsymbol{Z}_{Nj} := (Z_{N1}, \ldots, Z_{Nj})^\top$ (resp. $\widetilde{\boldsymbol{Z}}_{Nj}$). We will show that

(A) $\sum_{k=1}^{\kappa_N} \mathbb{E}\left[Z_{Nk} \mathbf{1}_{\{|Z_{Nk}| \leq 1\}} | \mathcal{F}_{N,k-1}\right] \xrightarrow{\mathbb{P}} 0.$

(B) $\sum_{k=1}^{\kappa_N} \mathrm{Var}\left[Z_{Nk} \mathbf{1}_{\{|Z_{Nk}| \leq 1\}} | \mathcal{F}_{N,k-1}\right] \xrightarrow{\mathbb{P}} 1.$

(C) $\sum_{k=1}^{\kappa_N} \mathbb{P}[|Z_{Nk}| > \epsilon | \mathcal{F}_{N,k-1}] \xrightarrow{\mathbb{P}} 0, \; \forall\, \epsilon > 0.$

The conclusion will then follow from an "almost-martingale" central limit theorem for triangular arrays, Shorack (5, Thm. 12.2). Fix some N, let $d = \kappa_N$, and let $\boldsymbol{\zeta} \sim \mathcal{N}_d(\mathbf{0}, I)$. Letting d_∞ denote the Kolmogorov metric, we obtain

$$\begin{aligned} d_\infty\left(\widetilde{\boldsymbol{Z}}_{Nd}, \boldsymbol{\zeta}\right) \;\leq\; & d_\infty\left(\widetilde{\boldsymbol{Z}}_{Nd}, \left\{\sqrt{\frac{n_1 n_2}{2N}} \left\langle (\widehat{\mathscr{R}}_X^{n_1} - \widehat{\mathscr{R}}_Y^{n_2}) \check{\boldsymbol{\varphi}}_{i(m)}, \check{\boldsymbol{\varphi}}_{j(m)} \right\rangle\right\}_{m=1}^{d}\right) \\ & + d_\infty\left(\left\{\sqrt{\frac{n_1 n_2}{2N}} \left\langle (\widehat{\mathscr{R}}_X^{n_1} - \widehat{\mathscr{R}}_Y^{n_2}) \check{\boldsymbol{\varphi}}_{i(m)}, \check{\boldsymbol{\varphi}}_{j(m)} \right\rangle\right\}_{m=1}^{d}, \boldsymbol{\zeta}\right) \end{aligned}$$

First we concentrate on the second term of the right hand side. From the proof of Theorem 1 and Pólya's theorem we know that this term converges to zero. In fact, recalling that $\widehat{\mathscr{R}}_X^{n_1} = n_1^{-1} \sum_{i=1}^{n_i} \mathscr{X}_i$ (resp. $\widehat{\mathscr{R}}_Y^{n_2}$) and that the $\boldsymbol{\varphi}_k$ are the eigenfunctions of the common covariance operator, the convergence can be seen to be due to the standard multidimensional

central limit theorem. We therefore have the following Berry-Esseen upper bound (e.g. DasGupta (2, Cor. 11.1)),

$$d_\infty\left(\left\{\sqrt{\frac{n_1n_2}{2N}}\left\langle(\widehat{\mathscr{R}}_X^{n_1}-\widehat{\mathscr{R}}_Y^{n_2})\check{\boldsymbol{\varphi}}_{i(m)},\check{\boldsymbol{\varphi}}_{j(m)}\right\rangle\right\}_{m=1}^{d},\boldsymbol{\zeta}\right)\leq\frac{Cd^{\frac{1}{4}}}{\sqrt{N}}.$$

Turning our attention to the first term in our triangle inequality, and letting $\nu_{i(k)} := \operatorname{sgn}[\langle\check{\boldsymbol{\varphi}}_{XY}^{i(k),N},\check{\boldsymbol{\varphi}}_{i(k)}\rangle]$, we note that

$$\mathbb{E}\left\|\widetilde{\boldsymbol{Z}}_{Nd}-\left\{\sqrt{\frac{n_1n_2}{2N}}\left\langle(\widehat{\mathscr{R}}_X^{n_1}-\widehat{\mathscr{R}}_Y^{n_2})\check{\boldsymbol{\varphi}}_{i(m)},\check{\boldsymbol{\varphi}}_{j(m)}\right\rangle\right\}_{m=1}^{d}\right\|_1=$$

$$=\sum_{k=1}^{d}\mathbb{E}\left|\widetilde{Z}_{Nk}-\sqrt{\frac{n_1n_2}{2N}}\left\langle(\widehat{\mathscr{R}}_X^{n_1}-\widehat{\mathscr{R}}_Y^{n_2})\check{\boldsymbol{\varphi}}_{i(k)},\check{\boldsymbol{\varphi}}_{j(k)}\right\rangle\right|$$

where, for every $1\leq k\leq d$ we have

$$\begin{aligned}
&\left|\widetilde{Z}_{Nk}-\sqrt{\frac{n_1n_2}{2N}}\left\langle(\widehat{\mathscr{R}}_X^{n_1}-\widehat{\mathscr{R}}_Y^{n_2})\check{\boldsymbol{\varphi}}_{i(k)},\check{\boldsymbol{\varphi}}_{j(k)}\right\rangle\right|\\
&\quad=\left|\left\langle\mathscr{Z}_N\nu_{i(k)}\check{\boldsymbol{\varphi}}_{XY}^{i(k),N},\nu_{j(k)}\check{\boldsymbol{\varphi}}_{XY}^{j(k),N}\right\rangle-\left\langle\mathscr{Z}_N\check{\boldsymbol{\varphi}}_{i(k)},\check{\boldsymbol{\varphi}}_{j(k)}\right\rangle\right|\\
&=\left|\left\langle\mathscr{Z}_N\nu_{i(k)}\check{\boldsymbol{\varphi}}_{XY}^{i(k),N},\nu_{j(k)}\check{\boldsymbol{\varphi}}_{XY}^{j(k),N}\right\rangle-\left\langle\mathscr{Z}_N\nu_{i(k)}\check{\boldsymbol{\varphi}}_{XY}^{i(k),N},\check{\boldsymbol{\varphi}}_{j(k)}\right\rangle+\left\langle\mathscr{Z}_N\nu_{i(k)}\check{\boldsymbol{\varphi}}_{XY}^{i(k),N},\check{\boldsymbol{\varphi}}_{j(k)}\right\rangle-\left\langle\mathscr{Z}_N\check{\boldsymbol{\varphi}}_{i(k)},\check{\boldsymbol{\varphi}}_{j(k)}\right\rangle\right|\\
&=\left|\left\langle\mathscr{Z}_N\nu_{i(k)}\check{\boldsymbol{\varphi}}_{XY}^{i(k),N},\nu_{j(k)}\check{\boldsymbol{\varphi}}_{XY}^{j(k),N}-\check{\boldsymbol{\varphi}}_{j(k)}\right\rangle+\left\langle\mathscr{Z}_N\left(\nu_{i(k)}\check{\boldsymbol{\varphi}}_{XY}^{i(k),N}-\check{\boldsymbol{\varphi}}_{i(k)}\right),\check{\boldsymbol{\varphi}}_{j(k)}\right\rangle\right|\\
&=\left|\left\langle\mathscr{Z}_N\nu_{i(k)}\check{\boldsymbol{\varphi}}_{XY}^{i(k),N},\nu_{j(k)}\check{\boldsymbol{\varphi}}_{XY}^{j(k),N}-\check{\boldsymbol{\varphi}}_{j(k)}\right\rangle+\left\langle\mathscr{Z}_N\check{\boldsymbol{\varphi}}_{j(k)},\nu_{i(k)}\check{\boldsymbol{\varphi}}_{XY}^{i(k),N}-\check{\boldsymbol{\varphi}}_{i(k)}\right\rangle\right|\\
&\leq\left\|\mathscr{Z}_N\nu_{i(k)}\check{\boldsymbol{\varphi}}_{XY}^{i(k),N}\right\|_{\mathcal{L}^2}\left\|\nu_{j(k)}\check{\boldsymbol{\varphi}}_{XY}^{j(k),N}-\check{\boldsymbol{\varphi}}_{j(k)}\right\|_{\mathcal{L}^2}+\left\|\mathscr{Z}_N\check{\boldsymbol{\varphi}}_{j(k)}\right\|_{\mathcal{L}^2}\left\|\nu_{i(k)}\check{\boldsymbol{\varphi}}_{XY}^{i(k),N}-\check{\boldsymbol{\varphi}}_{i(k)}\right\|_{\mathcal{L}^2}\\
&\leq\left\|\mathscr{Z}_N\right\|_{HS}\left\|\nu_{i(k)}\check{\boldsymbol{\varphi}}_{XY}^{i(k),N}\right\|_{\mathcal{L}^2}\left\|\nu_{j(k)}\check{\boldsymbol{\varphi}}_{XY}^{j(k),N}-\check{\boldsymbol{\varphi}}_{j(k)}\right\|_{\mathcal{L}^2}+\left\|\mathscr{Z}_N\right\|_{HS}\left\|\check{\boldsymbol{\varphi}}_{j(k)}\right\|_{\mathcal{L}^2}\left\|\nu_{i(k)}\check{\boldsymbol{\varphi}}_{XY}^{i(k),N}-\check{\boldsymbol{\varphi}}_{i(k)}\right\|_{\mathcal{L}^2}\\
&=\left\|\mathscr{Z}_N\right\|_{HS}\left(\left\|\nu_{j(k)}\check{\boldsymbol{\varphi}}_{XY}^{j(k),N}-\check{\boldsymbol{\varphi}}_{j(k)}\right\|_{\mathcal{L}^2}+\left\|\nu_{i(k)}\check{\boldsymbol{\varphi}}_{XY}^{i(k),N}-\check{\boldsymbol{\varphi}}_{i(k)}\right\|_{\mathcal{L}^2}\right)
\end{aligned}$$

Here we have used the Cauchy-Schwartz inequality and the fact that $\mathscr{Z}_N$ is a bounded

operator. By the triangle inequality we now obtain

$$\begin{aligned}
&\|\mathscr{Z}_N\|_{HS}\left(\left\|\nu_{j(k)}\check{\boldsymbol{\varphi}}_{XY}^{j(k),N}-\check{\boldsymbol{\varphi}}_{j(k)}\right\|_{\mathcal{L}^2}+\left\|\nu_{i(k)}\check{\boldsymbol{\varphi}}_{XY}^{i(k),N}-\check{\boldsymbol{\varphi}}_{i(k)}\right\|_{\mathcal{L}^2}\right)\\
&\quad\leq\|\mathscr{Z}_N\|_{HS}\left(\left\|\nu_{j(k)}\check{\boldsymbol{\varphi}}_{XY}^{j(k),N}-\nu_{j(k)}\lambda_{j(k)}^{-1/2}\hat{\boldsymbol{\varphi}}_{XY}^{j(k),N}\right\|_{\mathcal{L}^2}+\left\|\nu_{j(k)}\lambda_{j(k)}^{-1/2}\hat{\boldsymbol{\varphi}}_{XY}^{j(k),N}-\check{\boldsymbol{\varphi}}_{j(k)}\right\|_{\mathcal{L}^2}\right.\\
&\qquad\qquad\left.+\left\|\nu_{i(k)}\check{\boldsymbol{\varphi}}_{XY}^{i(k),N}-\nu_{i(k)}\lambda_{i(k)}^{-1/2}\hat{\boldsymbol{\varphi}}_{XY}^{i(k),N}\right\|_{\mathcal{L}^2}+\left\|\nu_{i(k)}\lambda_{i(k)}^{-1/2}\hat{\boldsymbol{\varphi}}_{XY}^{i(k),N}-\check{\boldsymbol{\varphi}}_{i(k)}\right\|_{\mathcal{L}^2}\right)\\
&\quad=\|\mathscr{Z}_N\|_{HS}\left((\hat{\lambda}_{j(k)}^{-1/2}-\lambda_{j(k)}^{-1/2})+\lambda_{j(k)}^{-1/2}\left\|\nu_{j(k)}\hat{\boldsymbol{\varphi}}_{XY}^{j(k),N}-\boldsymbol{\varphi}_{j(k)}\right\|_{\mathcal{L}^2}\right.\\
&\qquad\qquad\left.+(\hat{\lambda}_{i(k)}^{-1/2}-\lambda_{i(k)}^{-1/2})+\lambda_{i(k)}^{-1/2}\left\|\nu_{i(k)}\hat{\boldsymbol{\varphi}}_{XY}^{i(k),N}-\boldsymbol{\varphi}_{i(k)}\right\|_{\mathcal{L}^2}\right)
\end{aligned}$$

where we have used the simplified notation

$$\hat{\lambda}_{i(k)}=\sqrt{\frac{n_1}{N}\hat{\lambda}_{X,XY}^{i(k),n_1}+\frac{n_2}{N}\hat{\lambda}_{Y,XY}^{i(k),n_2}}.$$

We now apply the inequality given in Bosq (1, Lem. 4.3) and obtain

$$\begin{aligned}
&\|\mathscr{Z}_N\|_{HS}\left((\hat{\lambda}_{j(k)}^{-1/2}-\lambda_{j(k)}^{-1/2})+(\hat{\lambda}_{i(k)}^{-1/2}-\lambda_{i(k)}^{-1/2})+\lambda_{j(k)}^{-1/2}\left\|\nu_{j(k)}\hat{\boldsymbol{\varphi}}_{XY}^{j(k),N}-\boldsymbol{\varphi}_{j(k)}\right\|_{\mathcal{L}^2}\right.\\
&\qquad\qquad\left.+\lambda_{i(k)}^{-1/2}\left\|\nu_{i(k)}\hat{\boldsymbol{\varphi}}_{XY}^{i(k),N}-\boldsymbol{\varphi}_{i(k)}\right\|_{\mathcal{L}^2}\right)\\
&\quad\leq\|\mathscr{Z}_N\|_{HS}\left((\hat{\lambda}_{j(k)}^{-1/2}-\lambda_{j(k)}^{-1/2})+(\hat{\lambda}_{i(k)}^{-1/2}-\lambda_{i(k)}^{-1/2})\right.\\
&\qquad\qquad\lambda_{j(k)}^{-1/2}2\sqrt{2}\max\left\{(\lambda_{j(k)-1}-\lambda_{j(k)})^{-1},(\lambda_{j(k)}-\lambda_{j(k)+1})^{-1}\right\}\|\widehat{\mathscr{R}}_{XY}^{N}-\mathscr{R}_X\|_{HS}\\
&\qquad\qquad\left.+\lambda_{i(k)}^{-1/2}2\sqrt{2}\max\left\{(\lambda_{i(k)-1}-\lambda_{i(k)})^{-1},(\lambda_{i(k)}-\lambda_{i(k)+1})^{-1}\right\}\|\widehat{\mathscr{R}}_{XY}^{N}-\mathscr{R}_X\|_{HS}\right)
\end{aligned}$$

Recapitulating, we have obtained

$$\begin{aligned}
&\left|\widetilde{Z}_{Nk}-\sqrt{\frac{n_1n_2}{2N}}\left\langle(\widehat{\mathscr{R}}_X^{n_1}-\widehat{\mathscr{R}}_Y^{n_2})\check{\boldsymbol{\varphi}}_{i(k)},\check{\boldsymbol{\varphi}}_{j(k)}\right\rangle\right|\\
&\quad\leq\|\mathscr{Z}_N\|_{HS}\left((\hat{\lambda}_{j(k)}^{-1/2}-\lambda_{j(k)}^{-1/2})+(\hat{\lambda}_{i(k)}^{-1/2}-\lambda_{i(k)}^{-1/2})\right.\\
&\qquad\qquad\left.\lambda_{j(k)}^{-1/2}2\sqrt{2}\max\left\{(\lambda_{j(k)-1}-\lambda_{j(k)})^{-1},(\lambda_{j(k)}-\lambda_{j(k)+1})^{-1}\right\}\|\widehat{\mathscr{R}}_{XY}^{N}-\mathscr{R}_X\|_{HS}\right.
\end{aligned}$$

$$+ \lambda_{i(k)}^{-1/2} 2\sqrt{2} \max\left\{(\lambda_{i(k)-1} - \lambda_{i(k)})^{-1}, (\lambda_{i(k)} - \lambda_{i(k)+1})^{-1}\right\} \|\widehat{\mathscr{R}}_{XY}^{N} - \mathscr{R}_X\|_{HS}\Big)$$

Now we take expectations on both sides, expand the right hand side, and repeatedly apply the Cauchy-Schwartz inequality (with respect to the mean-square norm) to obtain

$$\begin{aligned}
&\mathbb{E}\left|\widetilde{Z}_{Nk} - \sqrt{\frac{n_1 n_2}{2N}}\left\langle(\widehat{\mathscr{R}}_X^{n_1} - \widehat{\mathscr{R}}_Y^{n_2})\check{\boldsymbol{\varphi}}_{i(k)}, \check{\boldsymbol{\varphi}}_{j(k)}\right\rangle\right| \\
&\quad \leq \sqrt{\mathbb{E}\,\|\mathscr{Z}_N\|_{HS}^2}\sqrt{\mathbb{E}(\hat{\lambda}_{j(k)}^{-1/2} - \lambda_{j(k)}^{-1/2})^2} + \sqrt{\mathbb{E}\,\|\mathscr{Z}_N\|_{HS}^2}\sqrt{\mathbb{E}(\hat{\lambda}_{i(k)}^{-1/2} - \lambda_{i(k)}^{-1/2})^2} \\
&\quad + \lambda_{j(k)}^{-1/2} 2\sqrt{2} \max\left\{(\lambda_{j(k)-1} - \lambda_{j(k)})^{-1}, (\lambda_{j(k)} - \lambda_{j(k)+1})^{-1}\right\} \sqrt{\mathbb{E}\,\|\mathscr{Z}_N\|_{HS}^2}\sqrt{\mathbb{E}\|\widehat{\mathscr{R}}_{XY}^{N} - \mathscr{R}_X\|_{HS}^2} \\
&\quad + \lambda_{i(k)}^{-1/2} 2\sqrt{2} \max\left\{(\lambda_{i(k)-1} - \lambda_{i(k)})^{-1}, (\lambda_{i(k)} - \lambda_{i(k)+1})^{-1}\right\} \sqrt{\mathbb{E}\,\|\mathscr{Z}_N\|_{HS}^2}\sqrt{\mathbb{E}\|\widehat{\mathscr{R}}_{XY}^{N} - \mathscr{R}_X\|_{HS}^2}
\end{aligned}$$

We note first that, by Minkowski's inequality, $\sqrt{\mathbb{E}\,\|\mathscr{Z}_N\|_{HS}^2}$ is bounded above for all N, by definition of the random operator $\mathscr{Z}_N$. Next, $\sqrt{\mathbb{E}(\hat{\lambda}_{i(k)}^{-1} - \lambda_{i(k)}^{-1})^2}$ and $\sqrt{\mathbb{E}(\hat{\lambda}_{i(k)}^{-1} - \lambda_{i(k)}^{-1})^2}$ are, asymptotically in N, of the order of $O(\lambda_{i(k)}^{-1/2} N^{-1/2})$ and so are also of the order of $O(\lambda_{i(d)}^{-1/2} N^{-1/2})$, when $k \leq d$. This can be seen by applying the Delta method to the CLT given in Dauxois et. al (3, Prop. 8). Finally, $\sqrt{\mathbb{E}\|\widehat{\mathscr{R}}_{XY}^{N} - \mathscr{R}_X\|_{HS}^2}$ is asymptotically of the order of $O(N^{-1/2})$ by the CLT in Hilbert Space (Bosq (1, Thm 2.7)).

Now by definition of $i(k)$ and $j(k)$, we have that $i(d)[i(d)+1]/2 = j(d)[j(d)+1]/2 = d$, so that it holds that

$$\lambda_{i(k)} = \lambda_{\frac{\sqrt{8d+1}-1}{2}} \geq \lambda_{\frac{3\sqrt{d}}{2}}.$$

Combining all the above, we arrive at

$$\mathbb{E}\left|\widetilde{Z}_{Nk} - \sqrt{\frac{n_1 n_2}{2N}}\left\langle(\widehat{\mathscr{R}}_X^{n_1} - \widehat{\mathscr{R}}_Y^{n_2})\check{\boldsymbol{\varphi}}_{i(k)}, \check{\boldsymbol{\varphi}}_{j(k)}\right\rangle\right| = O\left(\lambda_{3\sqrt{d}/2}^{-3/2} N^{-1/2}\right).$$

so that

$$\mathbb{E}\left\|\widetilde{\boldsymbol{Z}}_{Nd} - \left\{\sqrt{\frac{n_1 n_2}{2N}}\left\langle(\widehat{\mathscr{R}}_X^{n_1} - \widehat{\mathscr{R}}_Y^{n_2})\check{\boldsymbol{\varphi}}_{i(m)}, \check{\boldsymbol{\varphi}}_{j(m)}\right\rangle\right\}_{m=1}^{d}\right\|_1 = O\left(\lambda_{3\sqrt{d}/2}^{-3/2} N^{-1/2} d\right).$$

Letting d_W denote the L_1-Wasserstein distance between two probability measures, we have (e.g. Gibbs & Su (4)),

$$\begin{aligned} d_\infty(G_{N,d}, H_{N,d}) &\leq (1 + \|h_{N,d}\|_\infty)\sqrt{d_W(G_{N,d}, H_{N,d})} \\ &\leq (1 + \|h_{N,d}\|_\infty)\sqrt{\mathbb{E}\left\|\widetilde{\boldsymbol{Z}}_{Nd} - \left\{\sqrt{\frac{n_1 n_2}{2N}}\left\langle(\widehat{\mathscr{R}}_X^{n_1} - \widehat{\mathscr{R}}_Y^{n_2})\check{\boldsymbol{\varphi}}_{i(m)}, \check{\boldsymbol{\varphi}}_{j(m)}\right\rangle\right\}_{m=1}^{d}\right\|_1} \\ &= (1 + \|h_{N,d}\|_\infty) O\left(\lambda_{3\sqrt{d}/2}^{-3/4} N^{-1/4} d^{1/2}\right). \end{aligned}$$

where $H_{N,d}$ is the distribution function of $\left\{\sqrt{\frac{n_1 n_2}{2N}}\left\langle(\widehat{\mathscr{R}}_X^{n_1} - \widehat{\mathscr{R}}_Y^{n_2})\check{\boldsymbol{\varphi}}_{i(m)}, \check{\boldsymbol{\varphi}}_{j(m)}\right\rangle\right\}_{m=1}^{d}$, $G_{N,d}$ is the distribution function of $\widetilde{\boldsymbol{Z}}_{Nd}$, and $h_{N,d}$ is the density function of $H_{N,d}$. But $h_{N,d}$ is the density of a difference of two independent random vectors, each of which is in turn the sum of n_1 and n_2 iid random vectors, respectively. Thus, letting $h_d^{[1]}$ and $h_d^{[2]}$ be the respective densities, and by symmetry, we have,

$$\begin{aligned} \|h_{N,d}\|_\infty = \|\underbrace{h_{d,n_1}^{[1]} * \ldots * h_{d,n_1}^{[1]}}_{n_1 \text{ times}} * \underbrace{h_{d,n_2}^{[2]} * \ldots * h_{d,n_2}^{[2]}}_{n_2 \text{ times}}\|_\infty &\leq \|\underbrace{h_{d,n_1}^{[1]} * \ldots * h_{d,n_1}^{[1]}}_{n_1 \text{ times}}\|_1 \|\underbrace{h_{d,n_2}^{[2]} * \ldots * h_{d,n_2}^{[2]}}_{n_2 \text{ times}}\|_\infty \\ &= \|\underbrace{h_{d,n_2}^{[2]} * \ldots * h_{d,n_2}^{[2]}}_{n_2 \text{ times}}\|_\infty \end{aligned}$$

Now it is immediate that

$$\|h_{d,n_2}^{[2]} * \ldots * h_{d,n_2}^{[2]}\|_\infty \leq \|h_{n_2}^{[2]} * \ldots * h_{n_2}^{[2]}\|_\infty,$$

where $h_{n_2}^{[2]}$ is the marginal density of $\sqrt{\frac{n_1 n_2}{2N}}\left\langle(\frac{1}{n_2}\mathscr{X}_1)\check{\boldsymbol{\varphi}}_{i(1)}, \check{\boldsymbol{\varphi}}_{j(1)}\right\rangle$. But it must the case that $\|h_{n_2}^{[2]} * \ldots * h_{n_2}^{[2]}\|_\infty$ be bounded above, since $\sum_{i=1}^{n_2}\sqrt{\frac{n_1 n_2}{2N}}\left\langle(\frac{1}{n_2}\mathscr{X}_i)\check{\boldsymbol{\varphi}}_{i(1)}, \check{\boldsymbol{\varphi}}_{j(1)}\right\rangle$ is a sequence of variables with diffuse laws converging weakly to a non-degenerate Gaussian.

We are thus in a position to conclude that

$$d_\infty\left(\widetilde{\boldsymbol{Z}}_{Nd}, \boldsymbol{\zeta}\right) = O\left(\lambda_{3\sqrt{d}/2}^{-3/4} N^{-1/4} d^{1/2}\right). \tag{1}$$

Now recall that, with probability one,

$$\mathbb{E}\left(Z_{Nk}\mathbf{1}_{\{|Z_{Nk}|\leq 1\}}|\mathcal{F}_{N,k-1}\right) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{\kappa_N}}\left(x^2-1\right)\mathbf{1}_{\left\{|x^2-1|\leq\sqrt{2\kappa_N}\right\}} F_{\widetilde{Z}_{Nk}|\widetilde{\boldsymbol{Z}}_{N,k-1}}(dx|\widetilde{\boldsymbol{Z}}_{N,k-1})$$

where he have used standard notation for conditional distribution functions. It follows that, given ζ a standard Gaussian random variable,

$$\begin{aligned}
&\mathbb{E}\left(Z_{Nk}\mathbf{1}_{\{|Z_{Nk}|\leq 1\}}|\mathcal{F}_{N,k-1}\right) - \mathbb{E}\left(\frac{1}{\sqrt{\kappa_N}}(\zeta^2-1)\mathbf{1}_{\left\{|\zeta^2-1|\leq\sqrt{\kappa_N}\right\}}\right) \\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{\kappa_N}}\left(x^2-1\right)\mathbf{1}_{\left\{|x^2-1|\leq\sqrt{2\kappa_N}\right\}} F_{\widetilde{Z}_{Nk}|\widetilde{\boldsymbol{Z}}_{N,k-1}}(dx|\widetilde{\boldsymbol{Z}}_{N,k-1}) \\
&\quad - \int_{-\infty}^{+\infty} \frac{1}{\sqrt{\kappa_N}}\left(x^2-1\right)\mathbf{1}_{\left\{|x^2-1|\leq\sqrt{2\kappa_N}\right\}} F_{\zeta}(dx) \\
&\quad = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{\kappa_N}}\left(x^2-1\right)\mathbf{1}_{\left\{|x^2-1|\leq\sqrt{2\kappa_N}\right\}} \left[F^{\widetilde{\boldsymbol{Z}}_{N,k-1}}_{\widetilde{Z}_{Nk}|\widetilde{\boldsymbol{Z}}_{N,k-1}} - F_{\zeta}\right](dx)
\end{aligned}$$

with the alternative notation $F^{\widetilde{\boldsymbol{Z}}_{N,k-1}}_{\widetilde{Z}_{Nk}|\widetilde{\boldsymbol{Z}}_{N,k-1}}(x) \equiv F_{\widetilde{Z}_{Nk}|\widetilde{\boldsymbol{Z}}_{N,k-1}}(x|\widetilde{\boldsymbol{Z}}_{N,k-1})$. From (1) we have that for $\boldsymbol{\zeta}\sim\mathcal{N}_k(0,I)$, $d_\infty(\widetilde{\boldsymbol{Z}}_{Nk},\boldsymbol{\zeta}) = O\left(\lambda^{-1/3}_{3\sqrt{d}/2}N^{-1/4}k^{1/2}\right)$, so by Lemma 1 (see below), given any $\boldsymbol{z}\in\mathbb{R}^{k-1}$,

$$\sup_{x\in\mathbb{R}}\left|F^{\boldsymbol{z}}_{\widetilde{Z}_{Nk}|\widetilde{\boldsymbol{Z}}_{N,k-1}}(x) - F_{\zeta}(x)\right| = O\left(\lambda^{-3/4}_{3\sqrt{d}/2}N^{-1/4}k^{1/2}\right)$$

and so given $z\in\mathbb{R}^{k-1}$

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{\kappa_N}}\left(x^2-1\right)\mathbf{1}_{\left\{|x^2-1|\leq\sqrt{2\kappa_N}\right\}} \left[F^{\boldsymbol{z}}_{\widetilde{Z}_{Nk}|\widetilde{\boldsymbol{Z}}_{N,k-1}} - F_{\zeta}\right](dx) = O\left(\lambda^{-3/4}_{3\sqrt{\kappa_N}/2}N^{-1/4}k^{1/2}\kappa_N^{1/4}\right).$$

Consequently, for $\{\zeta_k\}$ an iid sequence of standard Gaussian variables, and for all $\omega\in\Omega$,

$$\sum_{k=1}^{\kappa_N}\left[\mathbb{E}\left[Z_{Nk}\mathbf{1}_{\{|Z_{Nk}|\leq 1\}}|\mathcal{F}_{N,k-1}\right] - \mathbb{E}\left[\frac{1}{\sqrt{\kappa_N}}(\zeta_k^2-1)\mathbf{1}_{\{|\zeta_k|\leq\sqrt{\kappa_N}\}}\right]\right] = O\left(\frac{\kappa_N^{7/4}}{N^{1/4}\lambda^{3/4}_{3\sqrt{\kappa_N}/2}}\right) = O\left(\frac{K_N^{7/2}}{N^{1/4}\lambda^{3/4}_{3\sqrt{\kappa_N}/2}}\right)$$

And, since

$$K_N^7 \lambda^{-3/2}_{\frac{3\sqrt{2K_N(K_N+1)}}{2}} \leq K_N^7 \lambda^{-3/2}_{\frac{3K_N}{2}} = o\left(\sqrt{N}\right),$$

it follows from our assumptions that the quantity above converges to zero almost certainly. But, on the other hand,

$$\begin{aligned}
&\left|\sum_{k=1}^{\kappa_N} \mathbb{E}\left[Z_{Nk}\mathbf{1}_{\{|Z_{Nk}|\leq 1\}}|\mathcal{F}_{N,k-1}\right]\right| \\
&\quad\leq \left|\sum_{k=1}^{\kappa_N}\left[\mathbb{E}\left[Z_{Nk}\mathbf{1}_{\{|Z_{Nk}|\leq 1\}}|\mathcal{F}_{N,k-1}\right] - \mathbb{E}\left[\frac{1}{\sqrt{\kappa_N}}(\zeta_k^2-1)\mathbf{1}_{\{|\zeta_k|\leq\sqrt{\kappa_N}\}}\right]\right]\right| \\
&\quad\quad + \left|\sum_{k=1}^{\kappa_N}\mathbb{E}\left[\frac{1}{\sqrt{\kappa_N}}(\zeta_k^2-1)\mathbf{1}_{\{|\zeta_k|\leq\sqrt{\kappa_N}\}}\right]\right|
\end{aligned}$$

with the last term obviously converging to zero as $N \to \infty$ so that condition (A) is fulfilled.

We now turn our attention to condition (B). By definition:

$$\sum_{k=1}^{\kappa_N}\mathrm{Var}\left[Z_{Nk}\mathbf{1}_{\{|Z_{Nk}|\leq 1\}}|\mathcal{F}_{N,k-1}\right] = \sum_{k=1}^{\kappa_N}\mathbb{E}\left[Z_{Nk}^2\mathbf{1}_{\{|Z_{Nk}|\leq 1\}}|\mathcal{F}_{N,k-1}\right] - \sum_{k=1}^{\kappa_N}\mathbb{E}^2\left[Z_{Nk}\mathbf{1}_{\{|Z_{Nk}|\leq 1\}}|\mathcal{F}_{N,k-1}\right]$$

That the second term converges to zero almost surely follows from our proof of condition (A). Hence, it suffices to concentrate on the first term. Following the same steps as with (A), we may write

$$\int_{-\infty}^{+\infty}\frac{(x^2-1)^2}{2\kappa_N}\mathbf{1}_{\left\{|x^2-1|\leq\sqrt{2\kappa_N}\right\}}\left[F^{\boldsymbol{z}}_{\widetilde{Z}_{Nk}|\widetilde{\boldsymbol{Z}}_{N,k-1}} - F_\zeta\right](dx) = O\left(\frac{K_N^{3/2}}{N^{1/4}\lambda^{3/4}_{3\sqrt{\kappa_N}/2}}\right)$$

This in turn imples that, with probability one,

$$\sum_{k=1}^{\kappa_N}\left[\mathbb{E}\left[Z_{Nk}\mathbf{1}_{\{|Z_{Nk}|\leq 1\}}|\mathcal{F}_{N,k-1}\right] - \mathbb{E}\left[\frac{1}{\sqrt{\kappa_N}}(\zeta_k^2-1)\mathbf{1}_{\{|\zeta_k|\leq\sqrt{\kappa_N}\}}\right]\right] \overset{N\to\infty}{\longrightarrow} 0.$$

Finally, we see that

$$\begin{aligned}
&\sum_{k=1}^{\kappa_N} \mathbb{E}\left[Z_{Nk}^2 \mathbf{1}_{\{|Z_{Nk}|\leq 1\}} | \mathcal{F}_{N,k-1}\right] \\
&\quad = \sum_{k=1}^{\kappa_N} \left[\mathbb{E}\left[Z_{Nk}^2 \mathbf{1}_{\{|Z_{Nk}|\leq 1\}} | \mathcal{F}_{N,k-1}\right] - \mathbb{E}\left[\frac{1}{2\kappa_N}(\zeta_k^2 - 1)^2 \mathbf{1}_{\{|\zeta_k|\leq\sqrt{\kappa_N}\}}\right]\right] \\
&\quad\quad + \sum_{k=1}^{\kappa_N} \mathbb{E}\left[\frac{1}{2\kappa_N}(\zeta_k^2 - 1)^2 \mathbf{1}_{\{|\zeta_k|\leq\sqrt{\kappa_N}\}}\right]
\end{aligned}$$

with the last term clearly converging to 1 almost certainly. This establishes condition (B).

Finally, we concentrate on condition (C). By definition,

$$\begin{aligned}
\mathbb{P}[|Z_{Nk}| > \epsilon | \mathcal{F}_{N,k-1}] &= 1 - \mathbb{E}\left[\mathbf{1}\left\{|Z_{Nk}| < \epsilon\right\} | \mathcal{F}_{N,k-1}\right] \\
&= 1 + \left(\mathbb{E}\left[\mathbf{1}\left\{|\zeta^2 - 1| < \epsilon\sqrt{\kappa_N}\right\}\right] - \mathbb{E}\left[\mathbf{1}\left\{|Z_{Nk}| < \epsilon\right\} | \mathcal{F}_{N,k-1}\right]\right) \\
&\quad -\mathbb{E}\left[\mathbf{1}\left\{|\zeta^2 - 1| < \epsilon\sqrt{\kappa_N}\right\}\right] \\
&= \left(\mathbb{E}\left[\mathbf{1}\left\{|\zeta^2 - 1| < \epsilon\sqrt{\kappa_N}\right\}\right] - \mathbb{E}\left[\mathbf{1}\left\{|Z_{Nk}| < \epsilon\right\} | \mathcal{F}_{N,k-1}\right]\right) + \mathbb{P}[|\zeta^2 - 1| > \epsilon\sqrt{\kappa_N}]
\end{aligned}$$

It is clear from our analysis of (A) and (B) that

$$\sum_{k=1}^{\kappa_N} \left(\mathbb{E}\left[\mathbf{1}\left\{|\zeta^2 - 1| < \epsilon\sqrt{\kappa_N}\right\}\right] - \mathbb{E}\left[\mathbf{1}\left\{|Z_{Nk}| < \epsilon\right\} | \mathcal{F}_{N,k-1}\right]\right) \xrightarrow{a.s.} 0.$$

Finally, we have

$$\sum_{k=1}^{\kappa_N} \mathbb{P}[|\zeta^2 - 1| > \epsilon\sqrt{\kappa_N}] = \kappa_N \mathbb{P}[|\zeta^2 - 1| > \epsilon\sqrt{\kappa_N}] = O\left(\frac{\kappa_N e^{-\left(1+\epsilon\sqrt{\kappa_N}\right)^{1/2}}}{\left(1 + \epsilon\sqrt{\kappa_N}\right)^{1/4}}\right) \overset{N\to\infty}{\longrightarrow} 0$$

by the tail decay properties of the Gaussian distribution. This completes the proof. □

Lemma 1. *Assume that F_n is a sequence of distribution functions on $\mathbb{R}^d$ converging weakly*

to a standard Gaussian distribution function Φ^d, at a rate ϵ_n in the Kolmogorov distance,

$$\sup_{\boldsymbol{x}\in\mathbb{R}^d} |F_n(\boldsymbol{x}) - \Phi^d(\boldsymbol{x})| = O(\epsilon_n).$$

Letting $d = p + q$, and given $\boldsymbol{y} \in \mathbb{R}^q$, we have

$$\sup_{x\in\mathbb{R}^p} |F_n(\boldsymbol{x}|\boldsymbol{y}) - \Phi^q(\boldsymbol{x})| = O(\epsilon_n).$$

Proof. By definition, and by our uniform bound, given any $\boldsymbol{y} \in \mathbb{R}^q$ we have that

$$\sup_{\boldsymbol{x}\in\mathbb{R}^p} |F_n(\boldsymbol{x}|\boldsymbol{y})F_n(\boldsymbol{y}) - \Phi^p(\boldsymbol{x})\Phi^q(\boldsymbol{y})| = \sup_{\boldsymbol{x}\in\mathbb{R}^p} |F_n(\boldsymbol{x},\boldsymbol{y}) - \Phi^d(\boldsymbol{x},\boldsymbol{y})| = O(\epsilon_n).$$

Now divide across by $\Phi^q(\boldsymbol{y})$, and obtain

$$\sup_{\boldsymbol{x}\in\mathbb{R}^p} \left| F_n(\boldsymbol{x}|\boldsymbol{y})\frac{F_n(\boldsymbol{y})}{\Phi^q(\boldsymbol{y})} - \Phi^p(\boldsymbol{x}) \right| = O(\epsilon_n) \tag{2}$$

By assumption of the theorem, it must also be that

$$|F_n(\boldsymbol{y}) - \Phi^q(\boldsymbol{y})| = O(\epsilon_n).$$

In turn, this implies that

$$\left| \frac{F_n(\boldsymbol{y})}{\Phi^q(\boldsymbol{y})} - 1 \right| = O(\epsilon_n), \tag{3}$$

for if this were not the case, for every $\alpha > 0$ and $M \geq 1$, there would exist and $m \geq M$ such that

$$\left| \frac{F_m(\boldsymbol{y})}{\Phi^q(\boldsymbol{y})} - 1 \right| > \frac{\alpha}{\Phi^q(\boldsymbol{y})}|\epsilon_m|,$$

or equivalently, for every $\alpha > 0$ and $M \geq 1$, there would exist and $m \geq M$ such that

$$|F_m(\boldsymbol{y}) - \Phi^q(\boldsymbol{y})| > \alpha|\epsilon_m|,$$

which would contradict the fact that $\sup_{\boldsymbol{u}} |F_n(\boldsymbol{u}) - \Phi^q(\boldsymbol{u})| \in O(\epsilon_n)$.

Now conditions (2) and (3) allow us to complete the proof by applying the triangle inequality:

$$d_\infty \left(F_n(\cdot|\boldsymbol{y}), \Phi_p\right) \leq d_\infty \left(F_n(\cdot|\boldsymbol{y}), \frac{F_n(\boldsymbol{y})}{\Phi_q(\boldsymbol{y})} F_n(\cdot|\boldsymbol{y})\right) + d_\infty \left(\frac{F_n(\boldsymbol{y})}{\Phi_q(\boldsymbol{y})} F_n(\cdot|\boldsymbol{y}), \Phi_p\right)$$

since

$$\begin{aligned} d_\infty \left(F_n(\cdot|\boldsymbol{y}), \frac{F_n(\boldsymbol{y})}{\Phi_q(\boldsymbol{y})} F_n(\cdot|\boldsymbol{y})\right) &= \sup_{\boldsymbol{x}\in\mathbb{R}^p} \left| F_n(\boldsymbol{x}|\boldsymbol{y}) - \frac{F_n(\boldsymbol{y})}{\Phi_q(\boldsymbol{y})} F_n(\boldsymbol{x}|\boldsymbol{y}) \right| \\ &= \left| 1 - \frac{F_n(\boldsymbol{y})}{\Phi_q(\boldsymbol{y})} \right| \sup_{\boldsymbol{x}\in\mathbb{R}^p} |F_n(\boldsymbol{x}|\boldsymbol{y})| \\ &= \left| 1 - \frac{F_n(\boldsymbol{y})}{\Phi_q(\boldsymbol{y})} \right| = O(\epsilon_n) \end{aligned}$$

□

References

[1] Bosq, D.(2000). *Linear processes in function spaces.* Springer.

[2] DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability.* Springer.

[3] Dauxois, J. Pousse, A. & Romain, Y. (1982). Asymptotic theory for the principal component analysis of a random vector function: some applications to statistical inference. *Journal of Multivariate Analysis, **12**: 136–154.*

[4] Gibbs, A.L. & Su, F.E. (2002). On choosing and bounding probability metrics. *International Statistical Review, **70**(3): 419–435.*

[5] Shorack, G. R. (2000). *Probability for Statisticians.* Springer.

B. Dispersion operators and resistant second-order functional data analysis

By David Kraus and Victor M. Panaretos

Biometrika, 99(4):813–832, 2012

DOI: 10.1093/biomet/ass037

Biometrika (2012), **99**, 4, *pp.* 813–832

Printed in Great Britain

doi: 10.1093/biomet/ass037
Advance Access publication 26 August 2012

Dispersion operators and resistant second-order functional data analysis

By DAVID KRAUS and VICTOR M. PANARETOS

Institute of Mathematics, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

david.kraus@epfl.ch victor.panaretos@epfl.ch

Summary

Inferences related to the second-order properties of functional data, as expressed by covariance structure, can become unreliable when the data are non-Gaussian or contain unusual observations. In the functional setting, it is often difficult to identify atypical observations, as their distinguishing characteristics can be manifold but subtle. In this paper, we introduce the notion of a dispersion operator, investigate its use in probing the second-order structure of functional data, and develop a test for comparing the second-order characteristics of two functional samples that is resistant to atypical observations and departures from normality. The proposed test is a regularized M-test based on a spectrally truncated version of the Hilbert–Schmidt norm of a score operator defined via the dispersion operator. We derive the asymptotic distribution of the test statistic, investigate the behaviour of the test in a simulation study and illustrate the method on a structural biology dataset.

Some key words: Covariance operator; Karhunen–Loève expansion; M-estimation; Resistant test; Spectral truncation; Two-sample testing.

1. Introduction

The second-order structure of a random function is key to understanding the nature of the functional observations that it induces, as it is inextricably linked with the smoothness properties of the stochastic fluctuations of the function. Given a suitable random function in a separable Hilbert space, e.g., $L^2[0, 1]$, these second-order properties are encapsulated in the covariance operator. The link with the smoothness properties of the random function is then given by the Karhunen–Loève expansion (e.g., Adler, 1990), which provides an optimal Fourier representation of the random function, using a basis comprised by the eigenfunctions of this operator. Consequently, a significant part of functional data analysis has concentrated on estimating the covariance operator, and employing its spectral decomposition in order to probe the smoothness properties of the functional data; see Bosq (2000), Dauxois et al. (1982), Hall & Hosseini-Nasab (2006), Ramsay & Silverman (2005), Gervini (2006), Hall et al. (2006) and Yao & Lee (2006), to name but a few. A natural inference problem is that of comparing the covariance structures of two samples of functional data, in order to decide whether they share the same fluctuation properties. Aspects of this problem were considered in Benko et al. (2009), who employed a bootstrap procedure to compare subsets of eigenfunctions or eigenvalues of the two samples in a financial context. The more global problem of testing whether two samples share the same covariance operator was investigated in the Gaussian case by Panaretos et al. (2010), motivated by the study of mechanical properties of DNA, and subsequently by Boente et al. (2011) through

a simulation-based approach. In a slightly different setting, Gabrys & Kokoszka (2007) and Horváth et al. (2010) investigated second-order tests to detect the presence or change of serial correlation in functional data. The goal of this paper is to study the problem of second-order inference in a more general setting. We focus on situations where the data are not Gaussian, and indeed may be characterized by the presence of influential observations. That we do not use the word outlier is deliberate: in the functional case, observations can significantly impact the empirical covariance operator, though they may not be outlying. The infinite-dimensional nature of the data means that an observation can be atypical in many ways, the deviation from the mean being only one; observations close to the mean may contain unusual frequency components. Detection of such observations via exploratory techniques may be nontrivial (Sun & Genton, 2011).

Such influential observations might significantly influence the estimation of the covariance, and, even more profoundly, the quality of the estimators of its spectrum. For these reasons, robustified estimates of the spectrum have been proposed, based on the spectra of robust estimators of the covariance operator. Locantore et al. (1999) proposed the use of the spectrum of the so-called spherical covariance operator in a discretized setting (Boente & Fraiman, 1999). Gervini (2008) introduced the functional median and further studied the properties of the spherical covariance spectrum for functional data concentrated on an unknown finite-dimensional hyperplane. Bali et al. (2012) adapted the projection-pursuit method of Li & Chen (1985) in the functional case. The sensitivity of the empirical covariance operator and its spectrum to the presence of influential observations can have an impact on testing procedures for the covariance operator. This is already observed in the finite-dimensional case (Layard, 1974; Olson, 1974), where deviations from a Gaussian assumption, or the presence of influential observations, can completely ruin a testing procedure even in one dimension (Box, 1953; Hampel et al., 1986). Finite-dimensional robust or resistant tests for covariance matrices cannot be directly extended to the functional case, as they often depend on the assumption of an invertible empirical covariance, which will by default be violated in the functional case for all sample sizes (Tiku & Balakrishnan, 1985; O'Brien, 1992; Zhang et al., 1991; Anderson, 2006). Even if a pseudo-inverse operator is employed, one immediately runs into the problem of ill-posedness.

To cope with these issues, this paper introduces a class of operators that we term dispersion operators that are implicitly defined through a variational problem, motivated by M-estimators of location for the tensor product of the centred functional observations. It is then proposed that these operators be used as proxies for the covariance operator, when inferences on the second-order structure are to be drawn for non-Gaussian and potentially contaminated functional samples. The implicit definition of a dispersion operator gives rise to a score equation, as the dispersion operator is a zero of the Fréchet derivative of the variational problem with respect to the operator argument. This functional score equation is then used as a basis to construct a test for the second-order comparison of two functional samples. The test is based on the distance of the functional score equation under the null hypothesis from zero, measured by an appropriately renormalized Hilbert–Schmidt distance.

2. SECOND-ORDER INFERENCE BASED ON THE DISPERSION OPERATOR

2·1. *Covariance operators*

To describe the second-order properties of a random element X in a separable Hilbert space of functions $\mathcal{H}$, often taken to be $L^2[0, 1]$, with norm $\|\cdot\|$ and inner product $\langle\cdot, \cdot\rangle$, one typically considers the covariance operator of X, $\mathscr{C} : \mathcal{H} \to \mathcal{H}$, defined as

$$\mathscr{C}(f) = E\{\langle f, X - \mu\rangle(X - \mu)\};$$

here $\mu = E(X)$ represents the mean of the function X. For example, in the case $\mathcal{H} \equiv L^2[0, 1]$, with inner product $\langle f, g \rangle = \int_0^1 f(t)g(t)\,\mathrm{d}t$, the covariance operator is represented as an integral operator

$$\mathscr{C}(f) = \int_0^1 r(\cdot, s) f(s)\,\mathrm{d}s,$$

where $r(s, t) = E[\{X(s) - \mu(s)\}\{X(t) - \mu(t)\}]$ stands for the covariance kernel of the process X. For the purposes of this paper, it will be more fruitful to think of the covariance operator as an operator related to tensor products on $\mathcal{H}$, rather than through the sample path perspective based on the covariance kernel. In particular, we will think of the covariance operator as

$$\mathscr{C} = E\{(X - \mu) \otimes (X - \mu)\},$$

where $\otimes$ stands for the tensor product on $\mathcal{H}$: for $f, g \in \mathcal{H}$, $f \otimes g$ defines an operator on $\mathcal{H}$ through $(f \otimes g)(h) = \langle g, h \rangle f$, where $h \in \mathcal{H}$. In this setting, and provided that $E(\|X\|^2) < \infty$, the covariance operator $\mathscr{C}$ can itself be thought of as an element of a Hilbert space, the space $\mathrm{HS}(\mathcal{H}, \mathcal{H})$ of Hilbert–Schmidt operators acting on $\mathcal{H}$. This is the space of linear operators $\mathscr{R}$ on $\mathcal{H}$ such that

$$\|\mathscr{R}\|_{\mathrm{HS}} = \left(\sum_{k=1}^{\infty} \|\mathscr{R} e_k\|^2 \right)^{1/2} < \infty,$$

where $\{e_k\}$ is any orthonormal basis of $\mathcal{H}$. Here, $\|\cdot\|_{\mathrm{HS}}$ defines a norm on $\mathrm{HS}(\mathcal{H}, \mathcal{H})$, corresponding to the inner product $\langle \mathscr{R}_1, \mathscr{R}_2 \rangle_{\mathrm{HS}} = \sum_{k=1}^{\infty} \langle \mathscr{R}_1 e_k, \mathscr{R}_2 e_k \rangle$. In what follows, we will usually omit the subscript HS, as the nature of the norm or inner product employed, whether it is an operator or an element norm, will be clearly implied from the space where its argument belongs.

In this Hilbert–Schmidt setting, the covariance operator can be seen as the operator $\mathscr{C} \in \mathrm{HS}(\mathcal{H}, \mathcal{H})$ that solves the variational problem

$$\min_{\mathscr{R} \in \mathrm{HS}(\mathcal{H}, \mathcal{H})} E\{\|(X - \mu) \otimes (X - \mu) - \mathscr{R}\|^2\}.$$

The sample counterpart of the covariance operator, the empirical covariance operator,

$$\hat{\mathscr{C}}_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}) \otimes (X_i - \bar{X}),$$

can be represented as the solution to the problem

$$\min_{\mathscr{R} \in \mathrm{HS}(\mathcal{H}, \mathcal{H})} \frac{1}{n} \sum_{i=1}^{n} \|(X_i - \bar{X}) \otimes (X_i - \bar{X}) - \mathscr{R}\|^2,$$

where $X_1, \ldots, X_n$ is a collection of independent and identically distributed copies of X, and $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$ stands for their empirical mean. This being essentially a least squares problem, both the empirical covariance operator and methods based on it will be sensitive to the presence of atypical observations in the dataset $X_1, \ldots, X_n$. In fact, it can also be seen that the empirical covariance operator admits a Gaussian maximum likelihood estimator interpretation, in a Cramér–Wold sense: if X is assumed Gaussian, then $\hat{\mathscr{C}}_n$ is the unique element of $\mathrm{HS}(\mathcal{H}, \mathcal{H})$

such that, for every $f \in \mathcal{H}$, $\langle f, \hat{\mathscr{C}}_n f \rangle$ is the unique maximum likelihood estimator of the variance of $\langle f, X \rangle$. The law of X is completely determined by the laws of the collection $\{\langle f, X \rangle : f \in \mathcal{H}\}$, and of course $\langle f, X \rangle$ is Gaussian with mean $\langle f, \mu \rangle$ and variance $\langle f, \mathscr{C} f \rangle$.

The basic strategy of this paper will be to obtain procedures pertaining to the second-order structure of X that are more resistant to departures from normality and to the presence of influential observations by replacing the squared norm in the variational problem defining the covariance by a less sensitive loss function. This gives rise to a new class of second-order characteristics, which we call dispersion operators.

2·2. *Dispersion operators*

Let P be a distribution on the separable Hilbert space $\mathcal{H}$ and let X be a random element with this distribution. The usual covariance is the integral of the operator

$$\mathscr{P}(x; \mu) = (x - \mu) \otimes (x - \mu), \quad x \in \mathcal{H},$$

with respect to P. This suggests that a dispersion operator could be defined as an M-estimator of the location of $\mathscr{P}(X; \mu)$. Let ρ be a nonnegative, differentiable, strictly increasing and convex function on $\mathbb{R}_0^+$ with $\rho(0) = 0$. We define the ρ-dispersion operator of the distribution P as

$$\mathscr{R}(\mathsf{P}) = \underset{\mathscr{R} \in \mathrm{HS}(\mathcal{H}, \mathcal{H})}{\arg\min} \; M(\mathsf{P}; \mathscr{R}, \mu), \tag{1}$$

where

$$\begin{aligned} M(\mathsf{P}; \mathscr{R}, \mu) &= E_{\mathsf{P}}[\rho\{\|\mathscr{P}(X; \mu) - \mathscr{R}\|\} - \rho\{\|\mathscr{P}(X; \mu)\|\}] \\ &= \int [\rho\{\|\mathscr{P}(x; \mu) - \mathscr{R}\|\} - \rho\{\|\mathscr{P}(x; \mu)\|\}] \, \mathrm{d}\mathsf{P}(x). \end{aligned} \tag{2}$$

In the definition of the dispersion operator, μ is chosen to be some suitable element of $\mathcal{H}$ with the interpretation of a location parameter. It is natural to use μ equal to the ρ-centre

$$\mu(\mathsf{P}) = \underset{\mu \in \mathcal{H}}{\arg\min} \, L(\mathsf{P}; \mu),$$

where

$$L(\mathsf{P}; \mu) = E_{\mathsf{P}}\{\rho(\|X - \mu\|) - \rho(\|X\|)\} = \int \{\rho(\|x - \mu\|) - \rho(\|x\|)\} \, \mathrm{d}\mathsf{P}(x).$$

Equivalently, one may define $\mu(\mathsf{P})$ and $\mathscr{R}(\mathsf{P})$ as solutions to score equations. The objective functionals $L(\mathsf{P}; \mu)$ and $M(\mathsf{P}; \mathscr{R}, \mu)$ are real-valued functionals defined on the Hilbert spaces $\mathcal{H}$ and $\mathrm{HS}(\mathcal{H}, \mathcal{H})$, respectively. The corresponding scores are their Fréchet derivatives, that is, linear functionals on the corresponding Hilbert space that can be uniquely identified with an element of that Hilbert space. Specifically, the centre $\mu(\mathsf{P})$ is the solution to the functional equation

$$G(\mathsf{P}; \mu) = 0,$$

where the element

$$G(\mathsf{P}; \mu) = \frac{\partial}{\partial \mu} L(\mathsf{P}; \mu) = E_{\mathsf{P}} \left\{ \frac{\rho'(\|X - \mu\|)}{\|X - \mu\|} (\mu - X) \right\} = \int \frac{\rho'(\|x - \mu\|)}{\|x - \mu\|} (\mu - x) \; \mathrm{d}\mathsf{P}(x)$$

of $\mathcal{H}$ determines the Fréchet derivative of L with respect to μ. The dispersion operator is defined as the solution to the operator equation

$$\mathscr{G}(\mathsf{P}; \mathscr{R}, \mu) = \mathscr{O}, \tag{3}$$

where $\mathscr{O}$ is the zero operator on $\mathcal{H}$ and the operator

$$\begin{aligned}\mathscr{G}(\mathsf{P}; \mathscr{R}, \mu) = \frac{\partial}{\partial \mathscr{R}} M(\mathsf{P}; \mathscr{R}, \mu) &= E_{\mathsf{P}} \left[\frac{\rho'\{\|\mathscr{P}(X; \mu) - \mathscr{R}\|\}}{\|\mathscr{P}(X; \mu) - \mathscr{R}\|} \{\mathscr{R} - \mathscr{P}(X; \mu)\} \right] \\ &= \int \frac{\rho'\{\|\mathscr{P}(x; \mu) - \mathscr{R}\|\}}{\|\mathscr{P}(x; \mu) - \mathscr{R}\|} \{\mathscr{R} - \mathscr{P}(x; \mu)\} \, \mathrm{d}\mathsf{P}(x)\end{aligned}$$

determines the Fréchet derivative of M with respect to $\mathscr{R}$.

The empirical dispersion operator based on the sample $X_1, \ldots, X_n$ is the dispersion operator of the empirical distribution $\hat{\mathsf{P}}$ of the sample, that is, $\mathscr{R}(\hat{\mathsf{P}})$. The empirical dispersion operator can be in general computed around any element $\mu \in \mathcal{H}$; in practice, one naturally uses the empirical centre $\mu(\hat{\mathsf{P}})$, i.e., the centre of the empirical distribution.

PROPOSITION 1. *Let* P *be a distribution on the separable Hilbert space* $\mathcal{H}$ *that is not concentrated on a line in* $\mathcal{H}$ *or on four points of* $\mathcal{H}$. *Assume that* ρ *is nonnegative, strictly increasing on* $[0, \infty)$ *and convex. Then, the objective function* $M(\mathsf{P}; \mathscr{R}, \mu)$ *as a functional of* $\mathscr{R}$ *is strictly convex for any* $\mu \in \mathcal{H}$ *and thus the* ρ*-dispersion operator around* μ *exists and is unique.*

Proposition 1 holds without any moment assumptions because the subtraction of $\rho\{\|\mathscr{P}(X; \mu)\|\}$ and $\rho(\|X\|)$ in the definition of $M(\mathsf{P}; \mathscr{R}, \mu)$ and $L(\mathsf{P}; \mu)$, respectively, guarantees the existence and finiteness of the objective functions. Under fairly weak further assumptions, we may also deduce that the empirical dispersion operator is well defined and consistent.

COROLLARY 1. *Let* $X_1, \ldots, X_n$ *be independent random elements with law* P *that has no discrete component and is such that the probability that* $X_1, \ldots, X_n$ *be collinear is zero* $(n \geqslant 3)$. *Then, for* $n \geqslant 5$, *the empirical* ρ*-dispersion operator corresponding to* $X_1, \ldots, X_n$ *exists and is almost surely unique. Moreover, if* $\hat{\mu}$ *is consistent for a location parameter* μ, *then the empirical dispersion operator around* $\hat{\mu}$ *is itself consistent for the dispersion operator around* μ.

We remark, for example, that the empirical functional median, i.e., the empirical centre corresponding to $\rho(u) = u$, was proven to be consistent for its theoretical counterpart in Gervini (2008). In fact, in the setting of Corollary 1, this result can be extended to location parameters corresponding to strictly increasing convex ρ-functions.

It is seen from (1) or (3) that the ρ-dispersion operator is self-adjoint. Moreover, from the spectral decomposition found in Proposition 2, it will follow that the ρ-dispersion operator is positive semidefinite. Although many results derived in this paper are valid for a wide class of functions ρ, the choice $\rho(u) = u^q$ for some $q > 0$ is especially attractive as the resulting centre is scale invariant and the dispersion is scale equivariant. For general ρ, it would be more appropriate to use a suitably studentized version of the objective functions; to this end, one can insert a preliminary estimator of the trace into the objective function.

We now provide explicit formulae for two main choices of the ρ-function.

When choosing $\rho(u) = u^2$, the score determining the ρ-dispersion operator equals $\mathscr{G}(\mathsf{P}; \mathscr{R}, \mu) = E_{\mathsf{P}}[2\{\mathscr{R} - \mathscr{P}(X; \mu)\}]$. Thus, $\mathscr{R}(\mathsf{P})$ can be found explicitly as $\mathscr{R}(\mathsf{P}) = E_{\mathsf{P}}\{\mathscr{P}(X; \mu)\}$. As the score for the ρ-centre is $G(\mathsf{P}; \mu) = E_{\mathsf{P}}\{2(\mu - X)\}$, the solution is $\mu(\mathsf{P}) = E_{\mathsf{P}}(X)$. Hence, the dispersion operator is the usual covariance operator.

The choice $\rho(u) = u$ is expected to place less emphasis on influential observations and result in more resistant procedures. The corresponding score operators for the dispersion and centre are

$$\mathscr{G}(\mathsf{P}; \mathscr{R}, \mu) = E_{\mathsf{P}}\left\{\frac{\mathscr{R} - \mathscr{P}(X; \mu)}{\|\mathscr{R} - \mathscr{P}(X; \mu)\|}\right\}, \quad G(\mathsf{P}; \mu) = E_{\mathsf{P}}\left(\frac{\mu - X}{\|\mu - X\|}\right).$$

The parameter $\mu(\mathsf{P})$ has been studied by a number of authors under different names in the multivariate as well as functional settings. In the multivariate context Chaudhuri (1996) calls $\mu(\mathsf{P})$ the geometric median; other authors (Serfling, 2004; Sirkiä et al., 2009) use the name spatial median and some authors (Huber & Ronchetti, 2009; Fritz et al., 2012) use the term L^1-centre or L^1-median. In the functional setting, $\mu(\mathsf{P})$ was studied by Locantore et al. (1999) and by Gervini (2008), who calls it the functional or spatial median. We use the term spatial median for $\mu(\mathsf{P})$ and, similarly, we call $\mathscr{R}(\mathsf{P})$ the spatial dispersion operator. To clarify the terminology, we recall that

$$\mathscr{S}(\mathsf{P}) = E_{\mathsf{P}}\left\{\frac{(X - \mu) \otimes (X - \mu)}{\|X - \mu\|^2}\right\}$$

is called the spherical covariance operator (Locantore et al., 1999). Unlike the parameters under the L^2-type loss function, the spatial median and spatial dispersion are not available explicitly. Their empirical counterparts $\hat{\mu} = \mu(\hat{\mathsf{P}})$ and $\hat{\mathscr{R}} = \mathscr{R}(\hat{\mathsf{P}})$ can, however, be obtained numerically, employing a Newton–Raphson algorithm, as explained in the Appendix.

The score function $\rho'(u) = qu^{q-1}$ corresponding to $\rho(u) = u^q$ is unbounded unless $q = 1$. Therefore, the estimator of the spatial dispersion operator, $q = 1$, is resistant, whereas other choices are nonresistant due to the effect of outliers, $q > 1$, or inliers, $q < 1$.

Although the dispersion operator is in general different from the covariance operator unless $\rho(u) = u^2$, it carries useful information on second-order properties of the distribution. There is an interesting link between the spectra of the dispersion and covariance operator. Let X admit the Karhunen–Loève expansion $X = \mu + \sum_{k=1}^{\infty} \lambda_k^{1/2} \beta_k \varphi_k$, where $\beta_1, \beta_2, \ldots$ are zero-mean unit-variance uncorrelated random variables, $\{\lambda_k : k \geqslant 1\}$ are the nonincreasing nonnegative eigenvalues and $\{\varphi_k : k \geqslant 1\}$ are the complete orthonormal eigenfunctions of the covariance operator $\mathscr{C}(\mathsf{P}) = E_{\mathsf{P}}\{(X - \mu) \otimes (X - \mu)\} = \sum_{k=1}^{\infty} \lambda_k \varphi_k \otimes \varphi_k$. We now investigate the eigendecomposition of the theoretical ρ-dispersion operator $\mathscr{R}(\mathsf{P})$ defined via M-estimation as the solution to (3). The main result is as follows.

PROPOSITION 2. *Assume that the Fourier coefficient sequence $\{\beta_k\}_{k=1}^{\infty}$ has a joint distribution that is invariant under the change of the sign of any component. Then, the dispersion operator $\mathscr{R}(\mathsf{P})$ has the same eigenfunctions as the covariance operator $\mathscr{C}(\mathsf{P})$, i.e., there exists a nonnegative sequence $\{\delta_k\}_{k=1}^{\infty}$ such that $\mathscr{R}(\mathsf{P}) = \sum_{k=1}^{\infty} \delta_k \varphi_k \otimes \varphi_k$. Furthermore, the eigenvalues $\delta_1, \delta_2, \ldots$ satisfy the conditions*

$$\delta_k = \lambda_k \frac{E\left(\dfrac{\rho'[\{\sum_i (\delta_i - \lambda_i \beta_i^2)^2 + \sum_{i \neq l} \lambda_i \lambda_l \beta_i^2 \beta_l^2\}^{1/2}]}{\{\sum_i (\delta_i - \lambda_i \beta_i^2)^2 + \sum_{i \neq l} \lambda_i \lambda_l \beta_i^2 \beta_l^2\}^{1/2}} \beta_k^2\right)}{E\left(\dfrac{\rho'[\{\sum_i (\delta_i - \lambda_i \beta_i^2)^2 + \sum_{i \neq l} \lambda_i \lambda_l \beta_i^2 \beta_l^2\}^{1/2}]}{\{\sum_i (\delta_i - \lambda_i \beta_i^2)^2 + \sum_{i \neq l} \lambda_i \lambda_l \beta_i^2 \beta_l^2\}^{1/2}}\right)} \quad (k = 1, 2, \ldots).$$

A similar result relating the covariance operator and the spherical covariance operator $\mathscr{S}(\mathsf{P})$ was obtained by Gervini (2008, Theorem 3) who showed that, under the assumption of exchangeability of the coefficient sequence, both operators have the same eigenfunctions in the same order; see also Marden (1999) and Boente & Fraiman (1999). Our proposition shows that the ρ-dispersion operator also has the same set of eigenfunctions. We conjecture that, potentially under further assumptions, the order of the eigenfunctions is also the same; computational experiments back this conjecture. Gervini (2008) assumed that the Karhunen–Loève expansion has only finitely many terms, i.e., that the distribution is concentrated on a finite-dimensional subspace, whereas our results hold even for processes with infinite series expansions. On the other hand, Gervini (2008) needed no moment assumptions, whereas we need to assume finite second moments: without moment assumptions the convergence of an infinite Karhunen–Loève series is not guaranteed, while a finite sum is always well defined regardless of the properties of the random summands.

2·3. *The two-sample test*

Having defined the notion of a dispersion operator, we now construct a two-sample second-order test based upon it. Let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ be two independent random samples from distributions $\mathsf{P}_1, \mathsf{P}_2$ on $\mathcal{H}$, whose ρ-centres are $\mu(\mathsf{P}_1)$, $\mu(\mathsf{P}_2)$ and ρ-dispersion operators are $\mathscr{R}(\mathsf{P}_1)$, $\mathscr{R}(\mathsf{P}_2)$. The goal is to test the null hypothesis H_0: $\mathscr{R}(\mathsf{P}_1) = \mathscr{R}(\mathsf{P}_2)$ against the general alternative H_1: $\mathscr{R}(\mathsf{P}_1) \neq \mathscr{R}(\mathsf{P}_2)$. Note that $\mu(\mathsf{P}_1)$, $\mu(\mathsf{P}_2)$ can be equal or different, as neither H_0 nor H_1 specifies their relation. We propose to employ the general idea of score tests, that is, to base the test on the estimating score for the general model, without assuming H_0, evaluated at the null estimate of the parameter.

As the centres $\mu(\mathsf{P}_1)$, $\mu(\mathsf{P}_2)$ are not restricted under the null hypothesis, they can be estimated separately by minimizing $L(\hat{\mathsf{P}}_1; \mu_1)$, $L(\hat{\mathsf{P}}_2; \mu_2)$, i.e., by solving $G(\hat{\mathsf{P}}_1; \mu_1) = 0$, $G(\hat{\mathsf{P}}_2; \mu_2) = 0$, respectively. Denote $\mu(\hat{\mathsf{P}}_j)$ by $\hat{\mu}_j$ $(j = 1, 2)$. On the other hand, the null estimator of the dispersion is based on both samples. As we now have two samples, we need to extend our notation to cover situations with two distributions, empirical or theoretical, mixed at proportions a and $1 - a$ for $a \in (0, 1)$. We denote

$$M(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}_1, \mathscr{R}_2, \mu_1, \mu_2) = aM(\mathsf{P}_1; \mathscr{R}_1, \mu_1) + (1 - a)M(\mathsf{P}_2; \mathscr{R}_2, \mu_2).$$

The common null value $\mathscr{R}$ of the dispersion operator is estimated by $\hat{\mathscr{R}}$, which minimizes $M(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \mathscr{R}, \mathscr{R}, \hat{\mu}_1, \hat{\mu}_2)$ where $a_n = n_1/n$ with $n = n_1 + n_2$. Equivalently, $\hat{\mathscr{R}}$ solves $\mathscr{G}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \mathscr{R}, \hat{\mu}_1, \hat{\mu}_2) = \mathscr{O}$, the null estimating equation, where $\mathscr{G}(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2) = a\mathscr{G}(\mathsf{P}_1; \mathscr{R}, \mu_1) + (1 - a)\mathscr{G}(\mathsf{P}_2; \mathscr{R}, \mu_2)$.

Using the reparameterization $\mathscr{R} = (\mathscr{R}_1 + \mathscr{R}_2)/2$, $\mathscr{T} = (\mathscr{R}_1 - \mathscr{R}_2)/2$, we have $\mathscr{R}_1 = \mathscr{R} + \mathscr{T}$, $\mathscr{R}_2 = \mathscr{R} - \mathscr{T}$ and we need to test H_0: $\mathscr{T} = \mathscr{O}$ against H_1: $\mathscr{T} \neq \mathscr{O}$. For the test, we need the score in the general model

$$\frac{\partial}{\partial(\mathscr{R}, \mathscr{T})^{\mathrm{T}}} M(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \mathscr{R} + \mathscr{T}, \mathscr{R} - \mathscr{T}, \hat{\mu}_1, \hat{\mu}_2) = \begin{pmatrix} \mathscr{G}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \mathscr{R}, \hat{\mu}_1, \hat{\mu}_2) \\ \mathscr{B}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \mathscr{R}, \hat{\mu}_1, \hat{\mu}_2) \end{pmatrix}$$

where $\mathscr{B}(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2) = a\mathscr{G}(\mathsf{P}_1; \mathscr{R}, \mu_1) - (1 - a)\mathscr{G}(\mathsf{P}_2; \mathscr{R}, \mu_2)$. The score test is based on this general score at the null estimator. When evaluated at $(\mathscr{R}, \mathscr{T}) = (\hat{\mathscr{R}}, \mathscr{O})$, the score is zero in the first component. Thus, the test can be based on the second component $\mathscr{B}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2)$.

When the null hypothesis holds, the score operator $\mathscr{B}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2)$ is expected to be close to the zero operator, otherwise it should be far from the zero operator. To perform the test, we need to measure the distance of $\mathscr{B}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2)$ from the zero operator and assess the significance of the resulting test statistic.

One way to measure the distance of the score operator from zero is to use its Hilbert–Schmidt norm. A drawback of this approach is that the resulting statistic does not have a tractable asymptotic distribution. The score operator turns out to be asymptotically Gaussian, but its Hilbert–Schmidt norm is not asymptotically distribution-free. In the context of comparison of covariance operators, Boente et al. (2011) use a simulation procedure to approximate the distribution of the statistic.

Another idea is to mimic the standard procedure from settings where the parameter of interest is Euclidean. In such settings, the difference of the score vector from zero is measured with the help of a quadratic form involving the score vector and the inverse of its covariance matrix. The quadratic statistic is usually asymptotically chi-square distributed and the null hypothesis is then rejected when the value of the statistic is significantly large. In the functional context, the score $\mathscr{B}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2)$ is infinite dimensional. Due to the noninvertibility of its covariance operator, one cannot construct a quadratic statistic. We overcome this problem by regularizing the score operator using spectral truncation.

The test object $\mathscr{B}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2)$ is an element of the space of operators $\mathrm{HS}(\mathcal{H}, \mathcal{H})$. Recall that $\mathrm{HS}(\mathcal{H}, \mathcal{H})$ is a Hilbert space with inner product defined as

$$\langle \mathscr{A}_1, \mathscr{A}_2 \rangle = \sum_{k=1}^{\infty} \langle \mathscr{A}_1 e_k, \mathscr{A}_2 e_k \rangle = \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \langle e_j, \mathscr{A}_1 e_k \rangle \langle e_j, \mathscr{A}_2 e_k \rangle, \quad \mathscr{A}_1, \mathscr{A}_2 \in \mathrm{HS}(\mathcal{H}, \mathcal{H}),$$

where $\{e_k : k = 1, 2, \dots\}$ is an arbitrary complete orthonormal basis of $\mathcal{H}$. For any complete orthonormal basis $\{\mathscr{E}_k : k = 1, 2, \dots\}$ of $\mathrm{HS}(\mathcal{H}, \mathcal{H})$, an operator $\mathscr{A} \in \mathrm{HS}(\mathcal{H}, \mathcal{H})$ and the square of its Hilbert–Schmidt norm can be written as

$$\mathscr{A} = \sum_{k=1}^{\infty} \langle \mathscr{A}, \mathscr{E}_k \rangle \mathscr{E}_k, \quad \|\mathscr{A}\|^2 = \sum_{k=1}^{\infty} \langle \mathscr{A}, \mathscr{E}_k \rangle^2.$$

Instead of this infinite series, one can use a truncated version. If $\mathscr{U} \subset \mathrm{HS}(\mathcal{H}, \mathcal{H})$ is a suitably chosen finite-dimensional linear subspace with an orthonormal basis $\{\mathscr{U}_1, \dots, \mathscr{U}_L\}$, then instead of $\|\mathscr{B}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2)\|^2$ one can use

$$\begin{aligned} \|\pi_{\mathscr{U}} \mathscr{B}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2)\|^2 &= \|\mathscr{B}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2) \pi_{\mathscr{U}}\|^2 \\ &= \sum_{l=1}^{L} \langle \mathscr{B}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2), \mathscr{U}_l \rangle^2, \end{aligned}$$

where $\pi_{\mathscr{U}}$ is the projection onto the subspace $\mathscr{U}$. That is, the test can be based on a score vector with components

$$S_l = \langle \mathscr{B}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2), \mathscr{U}_l \rangle \quad (l = 1, \dots, L). \tag{4}$$

One particular way of choosing the basis elements $\mathscr{U}_l$ is to derive them from a basis of the Hilbert space $\mathcal{H}$. If U is a K-dimensional linear subspace of $\mathcal{H}$ with an orthonormal basis $\{u_1, \dots, u_K\}$,

then one may use the $L = K(K+1)/2$ orthonormal operators of the form

$$\mathscr{U}_{jk} = \begin{cases} u_j \otimes u_j & (j = k), \\ (u_j \otimes u_k + u_k \otimes u_j)/2^{1/2} & (j < k). \end{cases} \tag{5}$$

There is yet another way of motivating the above truncation. Instead of measuring the difference of $\mathscr{B}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2)$ from zero on the entire Hilbert space $\mathcal{H}$, we can measure how it differs from the zero operator when attention is restricted to the linear subspace U. More precisely, instead of $\mathscr{B}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2)$, we use the operator $\pi_U \mathscr{B}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2)\pi_U$, where π_U is the projection operator on U. Its squared Hilbert–Schmidt norm

$$\|\pi_U \mathscr{B}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2)\pi_U\|^2 = \sum_{j=1}^{K}\sum_{k=1}^{K} \langle u_j, \mathscr{B}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2)u_k\rangle^2$$

is a truncated version of

$$\|\mathscr{B}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2)\|^2 = \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \langle e_j, \mathscr{B}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2)e_k\rangle^2,$$

where $\{e_j : j = 1, 2, \dots\}$ is any complete orthonormal basis of $\mathcal{H}$. The resulting scores

$$S_{jk} = \langle u_j, \mathscr{B}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2)u_k\rangle \quad (1 \leqslant j \leqslant k \leqslant K)$$

are equivalent to (4) with $\mathscr{U}_l$ of the form (5).

It is natural to use the basis operators of the form (5) with $u_1, \dots, u_K$ being the first K eigenfunctions of the dispersion operator $\mathscr{R}$ because, in light of Mercer's theorem, they carry the main portion of information about the dispersion operator. In practice, the eigenfunctions of $\mathscr{R}$ are not known, so one uses the eigenfunctions of the pooled sample estimator $\hat{\mathscr{R}}$. The number of components K can be selected as the minimal number for the cumulative proportion of dispersion explained by the subspace to exceed a certain threshold, e.g., 80% of the trace of the corresponding pooled sample dispersion operator. The proportion of dispersion, corresponding to the eigenvalues of the dispersion operator, is in general not equivalent to the proportion of variability, corresponding to the eigenvalues of the covariance operator.

To construct the test statistic, instead of simply summing squares of the terms S_l of the form (4), one combines them in a quadratic form reflecting their covariance structure.

The formal test will be based on the asymptotic distribution of the test statistic. Let n_1, n_2 be such that $n_1 \to \infty$, $n_2 \to \infty$ and $a_n = n_1/n \to a \in (0, 1)$. Assume that $\|G(\mathsf{P}_j; \mu)\|^2$, $\|\mathscr{G}(\mathsf{P}_j; \mathscr{R}, \mu)\|^2$ $(j = 1, 2)$ are finite. Let the function ρ: $\mathbb{R}_0^+ \to \mathbb{R}_0^+$ be twice differentiable, strictly increasing, and convex with $\rho(0) = 0$. Assume that the laws P_1, P_2 satisfy the conditions of Corollary 1 and the expectations $E_{\mathsf{P}_j}\{\rho'(\|X - \mu\|)^2\}$, $E_{\mathsf{P}_j}[\rho'\{\|\mathscr{P}(X; \mu) - \mathscr{R}\|\}^2]$, $E_{\mathsf{P}_j}\{\rho''(\|X - \mu\|)\}$, $E_{\mathsf{P}_j}[\rho''\{\|\mathscr{P}(X; \mu) - \mathscr{R}\|\}]$ and

$$E_{\mathsf{P}_j}\left\{\frac{\rho'(\|X - \mu\|)}{\|X - \mu\|}\right\}, \quad E_{\mathsf{P}_j}\left[\frac{\rho'\{\|\mathscr{P}(X; \mu) - \mathscr{R}\|\}}{\|\mathscr{P}(X; \mu) - \mathscr{R}\|}\right] \quad (j = 1, 2)$$

are finite. Assume that the derivatives $\mathscr{D}(\mathsf{P}_j; \mu)$, $\mathfrak{D}(\mathsf{P}_j; \mathscr{R}, \mu)$, $\mathbb{D}(\mathsf{P}_j; \mathscr{R}, \mu)$ given in (A1)–(A3) in the Appendix exist for $j = 1, 2$.

Let S be a score vector of length L of the form (4) for some linearly independent operators $\mathscr{U}_l = \mathscr{U}_l^{(n)}$. Let the operators $\mathscr{U}_l$ be either nonrandom, independent of n, or convergent in probability to some nonrandom limits, up to a possible sign ambiguity in the sense that there

exist some operators $\mathscr{U}_l^{\infty}$ such that $|\langle \mathscr{U}_l^{(n)}, \mathscr{U}_l^{\infty} \rangle|$ converges to 1. In this set-up, we have the following theorem.

THEOREM 1. *Under the null hypothesis* $H_0 : \mathscr{R}(\mathsf{P}_1) = \mathscr{R}(\mathsf{P}_2)$, *the score* $n^{1/2}\mathscr{B}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2)$ *converges weakly to a mean zero Gaussian random operator with covariance operator, which can be consistently estimated by* $\mathfrak{W}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2)$ *given in* (A5) *in the Appendix. The asymptotic distribution of the score vector* $n^{1/2}S$ *is* L*-variate zero-mean Gaussian with a covariance matrix that is consistently estimated by a matrix* W *with entries* $W_{j,l} = \langle \mathscr{U}_j, \mathfrak{W}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2)\mathscr{U}_l \rangle$ $(j, l = 1, \ldots, L)$. *The test statistic* $T = nS^{\mathrm{T}}W^{-1}S$ *asymptotically follows a* χ^2 *distribution with* L *degrees of freedom.*

We now deal with the two main cases, spatial and L^2-type, explicitly. In the spatial case, $\rho(u) = u$, we test the null hypothesis that the spatial dispersion operators are equal in both samples. The score operator takes the form

$$\mathscr{B}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2) = \frac{1}{n}\sum_{i=1}^{n_1} \frac{\hat{\mathscr{R}} - \mathscr{P}(X_i; \hat{\mu}_1)}{\|\hat{\mathscr{R}} - \mathscr{P}(X_i; \hat{\mu}_1)\|} - \frac{1}{n}\sum_{i=1}^{n_2} \frac{\hat{\mathscr{R}} - \mathscr{P}(Y_i; \hat{\mu}_2)}{\|\hat{\mathscr{R}} - \mathscr{P}(Y_i; \hat{\mu}_2)\|}.$$

The Fréchet derivatives $\mathscr{D}(\mathsf{P}; \mu)$, $\mathfrak{D}(\mathsf{P}; \mathscr{R}, \mu)$ involved in the covariance operator of the score are

$$\mathscr{D}(\mathsf{P}; \mu) = E_{\mathsf{P}}\left[\frac{1}{\|X - \mu\|}\left\{\mathscr{I} - \frac{(X-\mu)\otimes(X-\mu)}{\|X-\mu\|^2}\right\}\right],$$

$$\mathfrak{D}(\mathsf{P}; \mathscr{R}, \mu) = E_{\mathsf{P}}\left(\frac{1}{\|\mathscr{P}(X;\mu) - \mathscr{R}\|}\left[\mathfrak{I} - \frac{\{\mathscr{P}(X;\mu) - \mathscr{R}\}\otimes\{\mathscr{P}(X;\mu) - \mathscr{R}\}}{\|\mathscr{P}(X;\mu) - \mathscr{R}\|^2}\right]\right),$$

and the derivative $\mathbb{D}(\mathsf{P}; \mathscr{R}, \mu)$ evaluated at $f \in \mathcal{H}$ is

$$\mathbb{D}(\mathsf{P}; \mathscr{R}, \mu)f = E_{\mathsf{P}}\left[\frac{-\mathbb{Q}(X;\mu)f}{\|\mathscr{P}(X;\mu) - \mathscr{R}\|} + \frac{\langle \mathscr{P}(X;\mu) - \mathscr{R}, \mathbb{Q}(X;\mu)f\rangle}{\|\mathscr{P}(X;\mu) - \mathscr{R}\|^3}\{\mathscr{P}(X;\mu) - \mathscr{R}\}\right].$$

When the L^2 approach, $\rho(u) = u^2$, is employed, the hypothesis to be tested states that the covariance operators in both samples are equal. The null estimator of $\mathscr{R}$ takes the form $\hat{\mathscr{R}} = a_n\hat{\mathscr{R}}_1 + (1 - a_n)\hat{\mathscr{R}}_2$, that is, the pooled covariance estimator. The test score operator equals

$$\mathscr{B}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2) = a_n 2(\hat{\mathscr{R}} - \hat{\mathscr{R}}_1) - (1 - a_n)2(\hat{\mathscr{R}} - \hat{\mathscr{R}}_2) = 4a_n(1 - a_n)(\hat{\mathscr{R}}_2 - \hat{\mathscr{R}}_1),$$

which is a multiple of the difference of the empirical covariance operators. So, the test is equivalent to a Wald-type test proposed by Panaretos et al. (2010). This is different from the spatial test for which the score does not simplify to the difference of the spatial dispersions, so the score test differs from the Wald test. To compute the covariance operator of the test score, we first notice that $\mathbb{D}(\mathsf{P}; \mathscr{R}, \mu) = -2\,E_{\mathsf{P}}\{\mathbb{Q}(X; \mu)\}$ equals zero at $\mu = \mu(\mathsf{P}) = E_{\mathsf{P}}(X)$; see (A4) in the Appendix. Consequently, the fact that the centres of the two distributions must be estimated does not affect the asymptotic distribution, as could be expected. Also, $\mathfrak{D}(\mathsf{P}; \mathscr{R}, \mu) = 2\mathfrak{I}$. Hence, after straightforward calculations, the estimator of the covariance operator of the test operator is

$$\mathfrak{W}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2) = 4a_n(1 - a_n)\{(1 - a_n)\mathfrak{J}(\hat{\mathsf{P}}_1; \hat{\mathscr{R}}, \hat{\mu}_1) + a_n\mathfrak{J}(\hat{\mathsf{P}}_2; \hat{\mathscr{R}}, \hat{\mu}_2)\}$$

$$= 16a_n(1-a_n)$$
$$\times \left[(1-a_n)\frac{1}{n_1}\sum_{i=1}^{n_1}\{\mathscr{P}(X_i;\hat{\mu}_1)-\hat{\mathscr{R}}_1\}\otimes\{\mathscr{P}(X_i;\hat{\mu}_1)-\hat{\mathscr{R}}_1\}\right.$$
$$\left. + a_n\frac{1}{n_2}\sum_{i=1}^{n_2}\{\mathscr{P}(Y_i;\hat{\mu}_2)-\hat{\mathscr{R}}_2\}\otimes\{\mathscr{P}(Y_i;\hat{\mu}_2)-\hat{\mathscr{R}}_2\}\right].$$

In Panaretos et al. (2010), the limiting covariance of the L^2 score for the Wald-type test was investigated in the special case of Gaussian data and a simpler formula was found.

3. A simulation study

In order to investigate the performance of the testing procedure introduced in § 2·3, we generate random samples of size n_1, n_2 of curves of the form

$$X(t) = \mu_1(t) + \sum_{k=1}^{10}\lambda_{1k}^{1/2}a_{1k}2^{1/2}\sin\{2\pi k(t+\gamma_{1k})\} + \sum_{k=1}^{10}\nu_{1k}^{1/2}b_{1k}2^{1/2}\cos\{2\pi k(t+\delta_{1k})\},$$
$$Y(t) = \mu_2(t) + \sum_{k=1}^{10}\lambda_{2k}^{1/2}a_{2k}2^{1/2}\sin\{2\pi k(t+\gamma_{2k})\} + \sum_{k=1}^{10}\nu_{2k}^{1/2}b_{2k}2^{1/2}\cos\{2\pi k(t+\delta_{2k})\},$$

where the coefficients a_{jk}, b_{jk} are mutually independent random variables with zero-mean and unit variance. Three symmetric coefficient distributions are considered: normal, uniform and t_5, all scaled to have unit variance. As the test procedures are invariant with respect to the location shift of one or both samples, we set $\mu_1(t) = \mu_2(t) = 0$. Unless stated otherwise, we set $\gamma_{jk} = \delta_{jk} = 0$ in all situations. We perform the nonresistant L^2 test and the proposed spatial dispersion test at the nominal level $\alpha = 0{\cdot}05$. The sample sizes are $n_1 = n_2 = 50$. The basis of the subspace for dimension reduction consists of several leading eigenfunctions of the pooled sample estimator of the dispersion operator; that is, the pooled sample empirical covariance for the L^2 test and the pooled sample empirical spatial dispersion for the spatial test. The number of components K included in the basis is selected as the minimal number needed to explain at least 80% of the dispersion.

We first study the behaviour of the test procedures under the null hypothesis. We set $\lambda_{1k} = \lambda_{2k} = k^{-3}$ and $\nu_{1k} = \nu_{2k} = (1/3)^k$.

We begin with uncontaminated samples to verify that the tests maintain the prescribed nominal level. The first row of Table 1 shows that, in general, the asymptotic distribution approximates the distribution of both test statistics reasonably well. The asymptotic approximation for the L^2 method is slightly less accurate and tends to be liberal for distributions with light tails, i.e., normal and uniform.

Next we simulate datasets contaminated by atypical observations. Mean contamination, i.e., observations whose mean is different from the mean of the central distribution, usually impacts the level more seriously than pure covariance contamination, i.e., observations with the same mean but different covariance structure. Thus, we focus on mean contamination, i.e., outliers, in the study of the resistance of the level. In one or both samples, m_j out of n_j observations were replaced by observations that have mean function μ_j^{cont} instead of μ_j and the same covariance structure as the original distribution. We consider various distances of the contamination distribution from the central distribution and various contamination proportions, as indicated in Tables 1

Table 1. *Empirical rejection probabilities* (%) *at the nominal level* $\alpha = 5\%$ *under the null hypothesis. Samples of size* $n_1 = n_2 = 50$ *are contaminated by* m_1, m_2 *observations with mean functions* μ_1^{cont}, μ_2^{cont}, *respectively, and the same covariance structure as the central distribution. Estimates are based on* 2000 *simulation runs*

				Normal		t_5		Uniform	
m_1	$\mu_1^{\text{cont}}(t)$	m_2	$\mu_2^{\text{cont}}(t)$	L^2	Spatial	L^2	Spatial	L^2	Spatial
0		0		7·1	5·0	5·4	5·3	7·8	4·6
5	1	5	$1{\cdot}5 - 3\sin(\pi t)$	9·2	6·6	8·2	6·4	10·0	4·6
5	1·5	5	$1{\cdot}5 - 3\sin(\pi t)$	14·4	6·4	14·6	6·8	14·6	4·6
5	2·5	5	$1{\cdot}5 - 3\sin(\pi t)$	22·9	6·0	23·0	7·2	23·0	5·1
5	1	5	$2 - 4\sin(\pi t)$	11·2	7·2	10·3	7·7	11·7	5·2
5	1·5	5	$2 - 4\sin(\pi t)$	18·8	7·2	19·8	7·8	20·0	5·4
5	2·5	5	$2 - 4\sin(\pi t)$	30·4	7·2	32·4	8·2	30·8	6·4
5	1	5	$2{\cdot}5 - 5\sin(\pi t)$	14·1	8·2	14·0	8·0	15·0	6·4
5	1·5	5	$2{\cdot}5 - 5\sin(\pi t)$	25·9	8·2	25·4	8·4	27·8	6·5
5	2·5	5	$2{\cdot}5 - 5\sin(\pi t)$	41·8	8·3	46·4	9·0	42·4	7·2
5	1	0		7·4	6·0	6·4	5·4	8·6	5·0
5	1·5	0		12·6	5·9	11·2	5·7	13·4	4·6
5	2·5	0		19·0	6·1	17·8	6·0	17·8	4·7
0		5	$1{\cdot}5 - 3\sin(\pi t)$	9·0	6·0	7·2	6·6	9·8	5·6
0		5	$2 - 4\sin(\pi t)$	12·3	6·8	10·8	7·7	13·0	6·6
0		5	$2{\cdot}5 - 5\sin(\pi t)$	16·4	7·6	14·4	8·7	16·8	7·6

Table 2. *Empirical rejection probabilities* (%) *at the nominal level* $\alpha = 5\%$ *under the null hypothesis. Samples of size* $n_1 = n_2 = 50$ *are contaminated by* m_1, m_2 *observations with mean functions* $\mu_1^{\text{cont}}(t) = 1{\cdot}5$, $\mu_2^{\text{cont}}(t) = 2 - 4\sin(\pi t)$, *respectively, and the same covariance structure as the central distribution. Estimates are based on* 2000 *simulation runs*

	$m_1 = m, m_2 = 0$		$m_1 = 0, m_2 = m$		$m_1 = m_2 = m$	
m	L^2	Spatial	L^2	Spatial	L^2	Spatial
0	7·1	5·0	7·1	5·0	7·1	5·0
1	7·0	5·4	6·7	5·1	7·2	5·6
2	6·8	5·0	7·5	5·4	7·8	5·6
3	6·9	5·3	8·7	5·6	8·4	6·2
4	8·4	6·2	10·7	6·2	11·2	6·4
5	12·6	5·9	12·3	6·8	18·8	7·2
6	24·8	6·5	14·8	7·5	39·2	8·1
7	57·8	7·4	17·2	8·6	71·6	10·2
8	89·2	7·9	20·8	9·2	93·0	17·6
9	99·0	11·9	24·7	11·4	99·0	28·2
10	99·8	18·4	28·2	13·6	100·0	42·7

and 2. We consider only atypical observations that are not very far from the central distribution. These are the most insidious because they are often hidden in the main, apparently typical part of the dataset, do not stand out and thus are not easily identified visually, yet they often have a devastating impact on the behaviour of the nonresistant test. To illustrate this, we plot in Fig. 1 typical simulated samples with $m_1 = 5$, $\mu_1^{\text{cont}}(t) = 1{\cdot}5$ and $m_2 = 5$, $\mu_2^{\text{cont}}(t) = 2 - 4\sin(\pi t)$. When looking at the plots, one would be unable to identify atypical observations, if they were not highlighted. Visually, many of them do not seem to be very different from most curves, whereas some curves from the central distribution could be considered unusual.

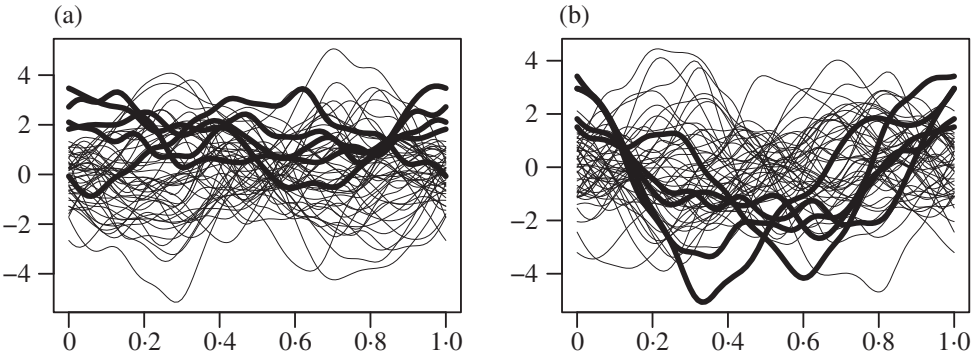

Fig. 1. Simulated contaminated samples. (a) Samples with $m_1 = 5$ atypical observations with $\mu_1^{\text{cont}}(t) = 1{\cdot}5$; (b) Samples with $m_2 = 5$ atypical observations with $\mu_2^{\text{cont}}(t) = 2 - 4\sin(\pi t)$. Atypical obsevations plotted in bold.

Table 1 shows that the proposed spatial test is much more resistant to contamination than the L^2-type test. For instance, notice that for $m_1 = m_2 = 5$, i.e., 10% contamination of both samples, the level of the spatial test in all situations considered is only slightly inflated, while the actual level of the L^2-type test exceeds 40%. Similarly, if one of the samples contains five atypical observations and the other is not contaminated, i.e., 10% contamination of one sample with 5% contamination overall, the spatial test rejects with probability close to the nominal level, while the level of the L^2-type test is as high as 19%. As the magnitude of atypical observations increases, the true level of the L^2 test, unlike that of the spatial one, increases dramatically. Comparing the behaviour of the tests across the various coefficient distributions, we observe no important differences. The higher resistance of the spatial method is also documented in Table 2, where the dependence of the level on the amount of contamination is studied for Gaussian data. The spatial procedure can tolerate much more contamination than can the L^2-type method.

Now we focus on the behaviour of the tests under alternatives. We consider five alternative scenarios. Under all of them, the parameters of the distribution of the first sample are $\lambda_{1k} = k^{-3}$ and $\nu_{1k} = (2/5)^k$. The parameters of the second sample are as follows. Under scenario I, we have $\lambda_{2k} = 1{\cdot}6\lambda_{1k}$ and $\nu_{2k} = 1{\cdot}6\nu_{1k}$ $(k = 1, \ldots, 10)$, so the samples differ only in scale, their covariance structure is otherwise the same. Under scenario II, we use $\lambda_{21} = 1{\cdot}5$, $\nu_{21} = 0{\cdot}8$ and $\lambda_{2k} = \lambda_{1k}$ and $\nu_{2k} = \nu_{1k}$ $(k = 2, \ldots, 10)$, so the covariance operators differ in the two leading eigenvalues, which however correspond to the same eigenfunctions. Scenario III has $\lambda_{2k} = \lambda_{1k}$ $(k = 1, \ldots, 10)$ and $\nu_{21} = 0{\cdot}2$, $\nu_{22} = 0{\cdot}35$ and $\nu_{2k} = \nu_{1k}$ $(k = 3, \ldots, 10)$; here the difference is on the second and third eigenvalues whose corresponding eigenfunctions are the same but in the opposite order. Under scenario IV, we set $\lambda_{22} = \lambda_{13}$, $\lambda_{23} = \lambda_{12}$, $\nu_{22} = \nu_{13}$, $\nu_{23} = \nu_{12}$ and $\lambda_{2k} = \lambda_{1k}$, $\nu_{2k} = \nu_{1k}$ $(k \notin \{2, 3\})$, so the difference occurs further down in the spectrum; eigenfunctions with indices 3, 4, 5, 6 are permuted, the leading two eigen-elements do not differ. Under scenario V, we use $\lambda_{2k} = \lambda_{1k}$, $\nu_{2k} = \nu_{1k}$ and $\gamma_{2k} = \delta_{2k} = 0{\cdot}15$ $(k = 1, \ldots, 10)$; in this case, the whole eigenbases are different but the eigenvalues remain the same in both samples.

First, we compare the power of the proposed spatial method with the L^2-type method for samples without contamination. Table 3 shows that in most cases the power of the spatial test is lower than the power of the L^2-type test for distributions with light tails. The lower efficiency of the spatial method is the price we pay for its increased resistance. Both methods have comparable power in the heavy tailed case under most scenarios. Under scenario IV the spatial method outperforms the L^2-type method. This is due to the automatic selection of K: for instance in the normal case, for the L^2-type test K equals 3 in 91 percent of cases while, for the spatial test, K equals 4 in 96 percent of cases; as the covariance operators differ on the third to sixth eigen-elements, K equal to 4 captures more of the difference.

Table 3. *Empirical rejection probabilities* (%) *at the nominal level* $\alpha = 5\%$ *under various alternative scenarios for samples of size* $n_1 = n_2 = 50$ *without contamination. Estimates are based on* 1000 *simulation runs*

	Normal		t_5		Uniform	
	L^2	Spatial	L^2	Spatial	L^2	Spatial
I	55	40	28	30	93	62
II	53	29	28	22	92	48
III	74	53	36	38	99	85
IV	38	61	24	53	49	73
V	76	58	53	51	96	72

Table 4. *Empirical rejection probabilities* (%) *of the spatial test at the nominal level* $\alpha = 5\%$ *under various alternative scenarios for samples of size* $n_1 = n_2 = 50$ *contaminated by* m_1, m_2 *atypical observations. Estimates are based on* 1000 *simulation runs*

Contamination configuration	m_1	m_2	I	II	III	IV	V
	0	0	40	29	53	61	58
A	5	5	12	16	57	64	59
	5	0	34	25	54	62	58
	0	5	15	16	56	63	61
B	5	5	29	22	36	39	55
	5	0	33	28	46	74	55
	0	5	40	28	49	34	57
C	5	5	24	18	34	39	52
	5	0	32	22	43	50	62
	0	5	31	24	43	49	48

Next, we investigate the impact of contamination on the power of the spatial test; we do not study the L^2-type test as we have seen before that its level is unreliable for contaminated data. The goal is to study if and how contamination can decrease the power. Similarly to the null scenario, here we also observed that mean contamination usually increases the rejection probability. Therefore, it is more interesting to contaminate data with curves with atypical covariance structure. We experimented with many configurations of atypical observations such that it is difficult to identify them visually and found that often even covariance contamination increases the rejection probability. Nevertheless, we were able to find some configurations for which we observed a decrease of the power in some situations. The central distributions follow the same scenarios I–V as before with normally distributed coefficients. Contamination configurations are as follows. Under configuration A, the contamination distribution has $\lambda_{1k}^{\mathrm{cont}} = 1{\cdot}4\lambda_{1k}$, $\nu_{1k}^{\mathrm{cont}} = 1{\cdot}4\nu_{1k}$, $\lambda_{2k}^{\mathrm{cont}} = 0{\cdot}25\lambda_{2k}$ and $\nu_{2k}^{\mathrm{cont}} = 0{\cdot}25\nu_{2k}$ $(k = 1, \ldots, 10)$, other parameters of the contamination distribution are the same as for the central distribution. Under configuration B, we set $\lambda_{1k}^{\mathrm{cont}} = 0{\cdot}3\lambda_{1k}$ and $\lambda_{2k}^{\mathrm{cont}} = 0{\cdot}3\lambda_{2k}$ $(k = 1, \ldots, 10)$, $\nu_{1k}^{\mathrm{cont}} = 0{\cdot}3\nu_{1k}$ and $\nu_{2k}^{\mathrm{cont}} = 0{\cdot}3\nu_{2k}$ $(k = 3, \ldots, 10)$, and $\nu_{11}^{\mathrm{cont}} = \nu_{21}^{\mathrm{cont}} = 1$ and $\nu_{12}^{\mathrm{cont}} = \nu_{22}^{\mathrm{cont}} = 0{\cdot}9$, while other parameters remain unchanged. Under configuration C, atypical observations in the first sample follow the central distribution of the second sample and atypical observations in the second sample follow the central distribution of the first sample.

The simulation results are presented in Table 4. We report only configurations with some detrimental effect on the power, while many configurations not reported here do not have such an effect. Under configuration A, we can see a decrease of the rejection probability for scenarios I and II. Configuration A was specifically designed to decrease the power under scenario I:

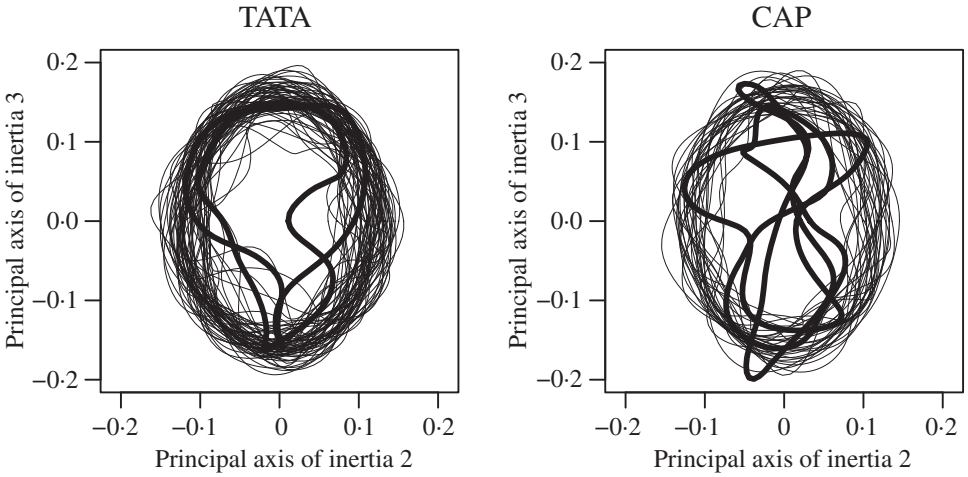

Fig. 2. Projection of DNA minicircle curves on the first principal plane spanned by the second and third principal axis of inertia. Atypical observations plotted in bold.

atypical observations deviate from the central distribution against the direction of the alternative; specifically, both the central and contamination distributions have proportional covariance operators but in the opposite direction. A similar phenomenon is seen for scenario II, where the directions of the alternative and of the contamination distribution are in a similar relationship. On the other hand, we observe no important effect of contamination of type A under scenarios III–V because in these cases atypical observations do not go against the alternative. Under configuration B, the power decreases mainly for scenarios III and IV. Configuration B downweights components other than the first and second cosine component, where it puts higher weight equal for both samples. As these are components carrying an important part of the difference between the covariances, one expects some decrease of the rejection probability, especially under scenarios III and IV. Under configuration C, the two samples are partly mixed, i.e., one sample contaminates the other sample and vice versa. This blurs the difference and somewhat decreases the power under some of the scenarios.

4. An illustration: DNA minicircle data

We illustrate the proposed methods on a dataset consisting of reconstructed three-dimensional electron microscope images of loops called minicircles obtained from short strands of DNA (Amzallag et al., 2006). The dataset contains 99 DNA minicircles of two types, TATA, 65 observations, and CAP, 34 observations, with identical base-pair sequences, except for a short subsequence where they differ. The main question is whether this difference affects the flexibility properties of the DNA minicircles. One way to formalize the flexibility properties is through the fluctuation pattern around the mean minicircle shape. This naturally leads one to consider two-sample second-order functional comparisons. DNA minicircles are closed curves in $\mathbb{R}^3$. In the original dataset, each curve was randomly rotated and shifted in $\mathbb{R}^3$ and had no starting point and no orientation. In Panaretos et al. (2010), an alignment procedure based on the moment of inertia tensor was used as a means of alignment of the curves in a common coordinate system. Figure 2 shows projections of aligned curves on the plane spanned by the two principal axes of inertia.

Using inverse weights induced by Gervini's (2008) spatial median, Panaretos et al. (2010) identified five unusual curves, possible outliers, and removed them from the analysis of the covariance structure. These atypical curves, plotted in thick lines in Fig. 2, are visibly different from the remaining curves. Panaretos et al. (2010) analysed the data without the atypical observations using a test comparing empirical covariance operators under the assumption that

the curves are Gaussian. Under this assumption, they observed significant differences at the 5% level. These differences were highly significant with a numerically zero p-value, when the comparison was restricted to the eigenvalues of the covariance operators; the corresponding empirical eigenfunctions suggested that the eigenfunction structure of the two operators was very similar.

Taking advantage of the results in the present paper, we may run an L^2-type test without assuming normality. When doing so, with the atypical observations still removed, the p-value of the L^2-type score test of the equality of covariance operators equals 0·023 with the dimension of the subspace on which the test operator is projected equal to $K = 6$, suggesting persistence of the effect, independently of a Gaussian assumption. Instead of removing apparently atypical observations manually, one might also wish to run an analysis on the complete dataset. However, the performance of L^2-type procedures was seen to be highly unstable in the presence of atypical observations, such as the ones in the present dataset, see Tables 1 and 2. By contrast, the spatial dispersion test was seen to maintain a level close to nominal in our simulations, especially in outlier scenarios similar to the one in the minicircle data. There may be further influential observations lurking in the sample. For this reason, we applied the score test based on the spatial dispersion operator, using the full minicircle dataset. In contrast to the other procedures, this yielded the p-value 0·353 indicative of a lack of significant differences in the spatial dispersions. The value of K was selected as the minimal number of components needed to explain 80% of the trace of the underlying null dispersion estimator. No further outliers were detected by the resistant test. The discordance between the L^2 and spatial tests is probably due to the reduced efficiency of the resistant procedure when the two samples share common eigenfunctions, as seems to be the case in the minicircle dataset; recall that the dispersion operator shares the same eigenfunctions with the covariance operator, possibly up to order. It was seen in our simulations that, in general, though the level of the spatial test was conserved, in the presence of influential observations its power was appreciably reduced when differences were only in the eigenvalues, i.e., under scenarios I and II in Table 4, as compared to scenarios where differences exist between the eigenfunctions, too, i.e., scenarios III–V in Table 4. Moreover the present framework does not immediately yield a special version of the test that would concentrate only on the eigenvalue structure; the complete structure of the operator is taken into account.

ACKNOWLEDGEMENT

We thank the editor, associate editor, and two anonymous referees for their extensive, constructive, and in-depth comments and suggestions. This research was supported in part by the European Research Council.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes proofs of Proposition 1, Corollary 1, Proposition 2, Theorem 1 and a technical lemma needed in the proof of Theorem 1.

APPENDIX

Computation

Assume that the observations $X_i \in \mathcal{H}$ are represented as linear combinations of some known fixed basis elements ψ_j, that is, $X_i = \sum_{j=1}^{p} \xi_{ij} \psi_j$. This representation is usually obtained by a least squares procedure, possibly with smoothing, from some form of discrete original observations of X_i. The exact form of the original data depends on the particular application. For instance, when $\mathcal{H}$ is a functional, L^2, space indexed by one-dimensional time, the original data usually consist of observations $X_i(t_k)$ $(k = 1, \ldots, m)$ for a grid

of points $t_1 < \cdots < t_m$. Now suppose that the original data are observed discretely but exactly, i.e., without noise; later we explain how to handle noisy discrete observations.

The methods proposed in this paper have the advantage that all required quantities and operations can be expressed in terms of basis coefficients; thus, from the computational point of view the task is multivariate. To estimate the centre, it is enough to find the vector of coefficients m_j in its basis expansion $\mu = \sum_{j=1}^{p} m_j \psi_j$. Similarly, for the dispersion operator, we need to find the matrix of coefficients $R_{jj'}$ in the expansion

$$\mathscr{R} = \sum_{j=1}^{p} \sum_{j'=1}^{p} R_{jj'} \psi_j \otimes \psi_{j'}.$$

For simplicity, we first assume that the basis $\psi_1, \ldots, \psi_p$ is orthonormal. Then, the norm in the objective function for μ is simply the norm of the coefficient vector, i.e., $\|X_i - \mu\|^2 = \|\xi_i - m\|^2 = \sum_{j=1}^{p} (\xi_{ij} - m_j)^2$, and the score operator $G(\hat{\mathsf{P}}; \mu)$ is equivalent to the p-vector

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\rho'(\|\xi_i - m\|)}{\|\xi_i - m\|} (m - \xi_i).$$

The Hilbert–Schmidt norm in the objective function for $\mathscr{R}$ is the Frobenius norm of the coefficient matrix, i.e.,

$$\|\mathscr{P}(X_i; \mu) - \mathscr{R}\|^2 = \|(\xi_i - m)(\xi_i - m)^{\mathrm{T}} - R\|^2 = \sum_{j=1}^{p} \sum_{j'=1}^{p} \{(\xi_{ij} - m_j)(\xi_{ij'} - m_{j'}) - R_{jj'}\}^2,$$

and the score operator $\mathscr{G}(\hat{\mathsf{P}}; \mathscr{R}, \mu)$ is equivalent to the $p \times p$ matrix

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\rho'\{\|(\xi_i - m)(\xi_i - m)^{\mathrm{T}} - R\|\}}{\|(\xi_i - m)(\xi_i - m)^{\mathrm{T}} - R\|} \{R - (\xi_i - m)(\xi_i - m)^{\mathrm{T}}\}.$$

For the two-sample test, the operator $\mathscr{B}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2)$ and the basis elements $\mathscr{U}_l$ for dimension reduction are equivalent to matrices, and the score components S_l are computed as their inner products. Similarly, all quantities involved in the covariance matrix of the score vector are computed in a multivariate setting. When the basis $\psi_1, \ldots, \psi_p$ is not orthonormal, one simply multiplies each coefficient vector ξ_i by the matrix $A^{1/2}$ where A has entries $a_{jj'} = \langle \psi_j, \psi_{j'} \rangle$, and performs all computations, i.e., estimation of the centre and dispersion, eigen-decomposition and the two-sample test, with these transformed multivariate inputs. This corresponds to switching from the original basis to the orthonormal basis $A^{-1/2}(\psi_1, \ldots, \psi_p)^{\mathrm{T}}$. If needed, the centre and the eigenfunctions can then be obtained in the original basis by multiplying their coefficient vectors by $A^{-1/2}$ and in the dispersion by multiplying its coefficient matrix by $A^{-1/2}$ from both sides. We refer to Ramsay & Silverman (2005, § 8.4.2) for a detailed explanation of a similar problem of computing functional principal components from coefficients with respect to a general non-orthonormal basis.

To estimate the centre and dispersion one solves the corresponding multivariate optimization problem. If $\rho(u) = u^2$, the solutions are the sample mean and covariance matrix of the coefficient vectors; otherwise an iterative procedure is used. We use the Broyden–Fletcher–Goldfarb–Shanno quasi-Newton method implemented in the R package (R Development Core Team, 2012) in the function optim, initialized by the componentwise median of ξ_i for the centre and the componentwise median of $(\xi_i - m)(\xi_i - m)^{\mathrm{T}}$ for the dispersion. This numerical procedure was reliable and reasonably fast in our experiments. This is in agreement with a detailed study of the numerical performance of various algorithms for the spatial median presented by Fritz et al. (2012).

In functional settings one can directly use the functional values on a grid of points instead of computing with basis coefficients. The basis approach is slightly more general than the discretization approach because it can be used for any separable Hilbert space, not only a functional space, and in the functional

case it does not require a common grid for all functions. Standard software for functional data analysis, such as the fda package in R, uses basis representations of data.

In many applications, the original functional values on a grid of points are observed with noise. In such situations, some degree of smoothing is necessary for the reconstruction of the underlying functional data. Ramsay & Silverman (2005, Chapter 5) describe how roughness penalties can be used to compute the basis coefficients of the functions. After this preliminary step, our methods can be applied to the reconstructed curve, i.e., their basis coefficients, as described above.

In the case of the spatial median, Gervini (2008, pp. 589–590) proposes an alternative method to deal with noise in discretely observed functions. Rather than on denoising and reconstructing the curves, his procedure is based on removing the bias, which is due to the errors, in the norm in the objective function with the help of a consistent estimate of the variance of the errors. He uses this idea in connection with numerical integration on a grid, but it can be adapted to the basis approach as well. However, this method is less practical for second-order problems, as one would also need to estimate higher order moments of the errors and use convoluted formulae to remove the bias from the norm in the objective functional.

Technical material

We now derive several key expressions pertaining to the assumptions, statement and discussion of Theorem 1. We use the script font, e.g., $\mathscr{D}$, $\mathscr{J}$, $\mathscr{I}$, for linear operators on $\mathcal{H}$, i.e., linear mappings $\mathcal{H} \to \mathcal{H}$, the fraktur font, e.g., $\mathfrak{D}$, $\mathfrak{J}$, $\mathfrak{I}$, $\mathfrak{H}$, $\mathfrak{W}$, for linear operators on Hilbert–Schmidt operators on $\mathcal{H}$, i.e., linear mappings $\mathrm{HS}(\mathcal{H}, \mathcal{H}) \to \mathrm{HS}(\mathcal{H}, \mathcal{H})$, and the blackboard bold font, e.g., $\mathbb{D}$, $\mathbb{J}$, $\mathbb{H}$, $\mathbb{Q}$, for linear operators from $\mathcal{H}$ to Hilbert–Schmidt operators on $\mathcal{H}$, i.e., linear mappings $\mathcal{H} \to \mathrm{HS}(\mathcal{H}, \mathcal{H})$.

First, we introduce certain derivatives in the Fréchet sense as follows. Denote by $\mathscr{I}$ and $\mathfrak{I}$ the identity operators on $\mathcal{H}$ and $\mathrm{HS}(\mathcal{H}, \mathcal{H})$, respectively. The derivative

$$\mathscr{D}(\mathsf{P}; \mu) = \frac{\partial}{\partial \mu} G(\mathsf{P}; \mu) = E_{\mathsf{P}} \left[\frac{\rho'(\|X - \mu\|)}{\|X - \mu\|} \mathscr{I} + \left\{ \frac{\rho''(\|X - \mu\|)}{\|X - \mu\|^2} - \frac{\rho'(\|X - \mu\|)}{\|X - \mu\|^3} \right\} \mathscr{P}(X; \mu) \right] \quad \text{(A1)}$$

is a linear mapping from $\mathcal{H}$ to $\mathcal{H}$. The derivative

$$\begin{aligned} \mathfrak{D}(\mathsf{P}; \mathscr{R}, \mu) = \frac{\partial}{\partial \mathscr{R}} \mathscr{G}(\mathsf{P}; \mathscr{R}, \mu) = E_{\mathsf{P}} \Bigg(& \frac{\rho'\{\|\mathscr{P}(X; \mu) - \mathscr{R}\|\}}{\|\mathscr{P}(X; \mu) - \mathscr{R}\|} \mathfrak{I} \\ & + \left[\frac{\rho''\{\|\mathscr{P}(X; \mu) - \mathscr{R}\|\}}{\|\mathscr{P}(X; \mu) - \mathscr{R}\|^2} - \frac{\rho'\{\|\mathscr{P}(X; \mu) - \mathscr{R}\|\}}{\|\mathscr{P}(X; \mu) - \mathscr{R}\|^3} \right] \mathfrak{P}(X; \mathscr{R}, \mu) \Bigg), \end{aligned} \quad \text{(A2)}$$

where we denote $\mathfrak{P}(x; \mathscr{R}, \mu) = \{\mathscr{P}(x; \mu) - \mathscr{R}\} \otimes \{\mathscr{P}(x; \mu) - \mathscr{R}\}$, is a linear mapping from $\mathrm{HS}(\mathcal{H}, \mathcal{H})$ to $\mathrm{HS}(\mathcal{H}, \mathcal{H})$. We define

$$\mathbb{D}(\mathsf{P}; \mathscr{R}, \mu) = \frac{\partial}{\partial \mu} \mathscr{G}(\mathsf{P}; \mathscr{R}, \mu), \quad \text{(A3)}$$

which is a linear mapping from $\mathcal{H}$ to $\mathrm{HS}(\mathcal{H}, \mathcal{H})$. To compute it, we first compute

$$\mathbb{Q}(x; \mu) = \frac{\partial}{\partial \mu} \mathscr{P}(x; \mu).$$

We consider its value at some $f \in \mathcal{H}$, i.e., we investigate the operator $\mathbb{Q}(x; \mu) f \in \mathrm{HS}(\mathcal{H}, \mathcal{H})$. This is done through its coordinate representation as follows. For any $g_1, g_2 \in \mathcal{H}$, we have

$$\begin{aligned} \langle g_1, \{\mathbb{Q}(x; \mu) f\} g_2 \rangle &= \left\langle g_1, \left[\left\{ \frac{\partial}{\partial \mu} \mathscr{P}(x; \mu) \right\} f \right] g_2 \right\rangle = \left\{ \frac{\partial}{\partial \mu} \langle g_1, \mathscr{P}(x; \mu) g_2 \rangle \right\} f \\ &= \left\{ \frac{\partial}{\partial \mu} (\langle x - \mu, g_1 \rangle \langle x - \mu, g_2 \rangle) \right\} f = -(\langle x - \mu, g_2 \rangle g_1 + \langle x - \mu, g_1 \rangle g_2) f \\ &= -\langle x - \mu, g_2 \rangle \langle g_1, f \rangle - \langle x - \mu, g_1 \rangle \langle g_2, f \rangle. \end{aligned} \quad \text{(A4)}$$

Then, the derivative of $\mathscr{G}(\mathsf{P}; \mathscr{R}, \mu)$ with respect to μ evaluated at $f \in \mathcal{H}$ is

$$\begin{aligned}\mathbb{D}(\mathsf{P}; \mathscr{R}, \mu) f = &- E_{\mathsf{P}} \left[\frac{\rho'\{\|\mathscr{P}(X; \mu) - \mathscr{R}\|\}}{\|\mathscr{P}(X; \mu) - \mathscr{R}\|} \mathbb{Q}(X; \mu) f \right] \\ &- E_{\mathsf{P}} \left(\left[\frac{\rho''\{\|\mathscr{P}(X; \mu) - \mathscr{R}\|\}}{\|\mathscr{P}(X; \mu) - \mathscr{R}\|^2} - \frac{\rho'\{\|\mathscr{P}(X; \mu) - \mathscr{R}\|\}}{\|\mathscr{P}(X; \mu) - \mathscr{R}\|^3} \right] \right. \\ &\left. \times \langle \mathscr{P}(X; \mu) - \mathscr{R}, \mathbb{Q}(X; \mu) f \rangle \{\mathscr{P}(X; \mu) - \mathscr{R}\} \right).\end{aligned}$$

We set

$$\begin{aligned}\mathfrak{D}_0(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2) &= a\mathfrak{D}(\mathsf{P}_1; \mathscr{R}, \mu_1) + (1 - a)\mathfrak{D}(\mathsf{P}_2; \mathscr{R}, \mu_2), \\ \mathfrak{D}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2) &= a\mathfrak{D}(\mathsf{P}_1; \mathscr{R}, \mu_1) - (1 - a)\mathfrak{D}(\mathsf{P}_2; \mathscr{R}, \mu_2).\end{aligned}$$

Next, using the notation $f^{\otimes 2} = f \otimes f$ for $f \in \mathcal{H}$ and $\mathscr{A}^{\otimes 2} = \mathscr{A} \otimes \mathscr{A}$ for $\mathscr{A} \in \mathrm{HS}(\mathcal{H}, \mathcal{H})$, we define

$$\begin{aligned}\mathscr{J}(\mathsf{P}; \mu) &= E_{\mathsf{P}} \left[\left\{ \frac{\rho'(\|X - \mu\|)}{\|X - \mu\|} (\mu - X) - G(\mathsf{P}; \mu) \right\}^{\otimes 2} \right] \\ \mathfrak{J}(\mathsf{P}; \mathscr{R}, \mu) &= E_{\mathsf{P}} \left(\left[\frac{\rho'\{\|\mathscr{P}(X; \mu) - \mathscr{R}\|\}}{\|\mathscr{P}(X; \mu) - \mathscr{R}\|} \{\mathscr{R} - \mathscr{P}(X; \mu)\} - \mathscr{G}(\mathsf{P}; \mathscr{R}, \mu) \right]^{\otimes 2} \right)\end{aligned}$$

and

$$\begin{aligned}\mathbb{J}(\mathsf{P}; \mathscr{R}, \mu) = E_{\mathsf{P}} \left(\left[\frac{\rho'\{\|\mathscr{P}(X; \mu) - \mathscr{R}\|\}}{\|\mathscr{P}(X; \mu) - \mathscr{R}\|} \{\mathscr{R} - \mathscr{P}(X; \mu)\} - \mathscr{G}(\mathsf{P}; \mathscr{R}, \mu) \right] \right. \\ \left. \otimes \left\{ \frac{\rho'(\|X - \mu\|)}{\|X - \mu\|} (\mu - X) - G(\mathsf{P}; \mu) \right\} \right).\end{aligned}$$

Next, we denote

$$\begin{aligned}\mathfrak{H}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2) &= \mathfrak{I} - \mathfrak{D}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)\mathfrak{D}_0(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)^{-1}, \\ \mathbb{H}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2) &= \mathfrak{H}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)\mathbb{D}(\mathsf{P}_1; \mathscr{R}, \mu_1)\mathscr{D}(\mathsf{P}_1; \mu_1)^{-1}, \\ \mathfrak{H}_2(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2) &= \mathfrak{I} + \mathfrak{D}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)\mathfrak{D}_0(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)^{-1}, \\ \mathbb{H}_2(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2) &= \mathfrak{H}_2(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)\mathbb{D}(\mathsf{P}_2; \mathscr{R}, \mu_2)\mathscr{D}(\mathsf{P}_2; \mu_2)^{-1},\end{aligned}$$

where $\mathfrak{I}$ stands for the identity operator on $\mathrm{HS}(\mathcal{H}, \mathcal{H})$. Finally, we set

$$\mathfrak{W}(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2) = a\mathfrak{W}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2) + (1 - a)\mathfrak{W}_2(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2), \quad \text{(A5)}$$

where

$$\begin{aligned}\mathfrak{W}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2) = &\mathfrak{H}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)\mathfrak{J}(\mathsf{P}_1; \mathscr{R}, \mu_1)\mathfrak{H}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)^* \\ &- \mathfrak{H}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)\mathbb{J}(\mathsf{P}_1; \mathscr{R}, \mu_1)\mathbb{H}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)^* \\ &- \mathbb{H}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)\mathbb{J}(\mathsf{P}_1; \mathscr{R}, \mu_1)^*\mathfrak{H}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)^* \\ &+ \mathbb{H}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)\mathscr{J}(\mathsf{P}_1; \mathscr{R}, \mu_1)\mathbb{H}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)^*\end{aligned}$$

with * denoting adjoint operators, and $\mathfrak{W}_2(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)$ is defined analogously with $\mathbb{H}_2, \mathfrak{H}_2$ in place of $\mathbb{H}_1, \mathfrak{H}_1$, respectively, and P_2 instead of P_1 in $\mathfrak{J}, \mathbb{J}, \mathscr{J}$.

REFERENCES

ADLER, R. J. (1990). *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 12. Hayward: Institute of Mathematical Statistics.

AMZALLAG, A., VAILLANT, C., JACOB, M., UNSER, M., BEDNAR, J., KAHN, J. D., DUBOCHET, J., STASIAK, A. & MADDOCKS, J. H. (2006). 3D reconstruction and comparison of shapes of DNA minicircles observed by cryo-electron microscopy. *Nucleic Acids Res.* **34**, e125.

ANDERSON, M. J. (2006). Distance-based tests for homogeneity of multivariate dispersions. *Biometrics* **62**, 245–53.

BALI, L., BOENTE, G., TYLER, D. E. & WANG, J.-L. (2012). Robust functional principal components: A projection-pursuit approach. *Ann. Statist.* **39**, 2852–82.

BENKO, M., HÄRDLE, W. & KNEIP, A. (2009). Common functional principal components. *Ann. Statist.* **37**, 1–34.

BOENTE, G. & FRAIMAN, R. (1999). Comment on a paper by Locantore et al. *Test* **8**, 28–35.

BOENTE, G., RODRIGUEZ, D. & SUED, M. (2011). Testing the equality of covariance operators. In *Recent Advances in Functional Data Analysis and Related Topics*, Ed. F. Ferraty, pp. 49–53. Heidelberg: Physica-Verlag.

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

BOX, G. E. P. (1953). Non-normality and tests on variances. *Biometrika* **40**, 318–35.

CHAUDHURI, P. (1996). On a geometric notion of quantiles for multivariate data. *J. Am. Statist. Assoc.* **91**, 862–72.

DAUXOIS, J., POUSSE, A. & ROMAIN, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *J. Mult. Anal.* **12**, 136–54.

FRITZ, H., FILZMOSER, P. & CROUX, C. (2012). A comparison of algorithms for the multivariate L_1-median. *Comp. Statist.*, to appear. doi: 10.1007/s00180-011-0262-4.

GABRYS, R. & KOKOSZKA, P. (2007). Portmanteau test of independence for functional observations. *J. Am. Statist. Assoc.* **102**, 1338–48.

GERVINI, D. (2006). Free-knot spline smoothing for functional data. *J. R. Statist. Soc.* B **68**, 671–87.

GERVINI, D. (2008). Robust functional estimation using the median and spherical principal components. *Biometrika* **95**, 587–600.

HALL, P. & HOSSEINI-NASAB, M. (2006). On properties of functional principal components analysis. *J. R. Statist. Soc.* B **68**, 109–26.

HALL, P., MÜLLER, H.-G. & WANG, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* **34**, 1493–517.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. & STAHEL, W. A. (1986). *Robust Statistics*. New York: Wiley.

HORVÁTH, L., HUŠKOVÁ, M. & KOKOSZKA, P. (2010). Testing the stability of the functional autoregressive process. *J. Mult. Anal.* **101**, 352–67.

HUBER, P. J. & RONCHETTI, E. M. (2009). *Robust Statistics*. Hoboken: Wiley.

LAYARD, M. W. J. (1974). A Monte Carlo comparison of tests for equality of convariance matrices. *Biometrika* **61**, 461–5.

LI, G. & CHEN, Z. (1985). Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and Monte Carlo. *J. Am. Statist. Assoc.* **80**, 759–66.

LOCANTORE, N., MARRON, J. S., SIMPSON, D. G., TRIPOLI, N., ZHANG, J. T. & COHEN, K. L. (1999). Robust principal component analysis for functional data. *Test* **8**, 1–73.

MARDEN, J. I. (1999). Some robust estimates of principal components. *Statist. Prob. Lett.* **43**, 349–59.

O'BRIEN, P. C. (1992). Robust procedures for testing equality of covariance matrices. *Biometrics* **48**, 819–27.

OLSON, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *J. Am. Statist. Assoc.* **69**, 894–08.

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.

R DEVELOPMENT CORE TEAM (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

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

SERFLING, R. (2004). Nonparametric multivariate descriptive measures based on spatial quantiles. *J. Statist. Plan. Infer.* **123**, 259–78.

SIRKIÄ, S., TASKINEN, S., OJA, H. & TYLER, D. E. (2009). Tests and estimates of shape based on spatial signs and ranks. *J. Nonparam. Statist.* **21**, 155–76.

SUN, Y. & GENTON, M. G. (2011). Functional boxplots. *J. Comp. Graph. Statist.* **20**, 316–334.

TIKU, M. L. & BALAKRISHNAN, N. (1985). Testing the equality of variance-covariance matrices the robust way. *Commun. Statist.* A **14**, 3033–51.

YAO, F. & LEE, T. C. M. (2006). Penalized spline models for functional principal component analysis. *J. R. Statist. Soc.* B **68**, 3–25.

ZHANG, J., PANTULA, S. G. & BOOS, D. D. (1991). Robust methods for testing the pattern of a single covariance matrix. *Biometrika* **78**, 787–95.

[*Received April* 2011. *Revised May* 2012]

Biometrika (2012), **xx**, x, *pp*. 1–8

Printed in Great Britain

Supplementary file: Dispersion operators and resistant second-order functional data analysis

BY DAVID KRAUS AND VICTOR M. PANARETOS

Section de Mathématiques, Ecole Polytechnique Fédérale de Lausanne, EPFL Station 8, 1015 Lausanne, Switzerland

david.kraus@epfl.ch victor.panaretos@epfl.ch

SUMMARY

This supplementary file contains proofs of Proposition 1, Corollary 1, Proposition 2, Theorem 1 and a technical lemma needed in the proof of Theorem 1. Equations in this supplement are numbered (S1), (S2), ...; equation numbers such as (1), (2), ... or (A1), (A2), ... refer to the main body of the paper.

PROOF OF PROPOSITION 1

It suffices to prove that the finitely-valued objective functional $M(\mathsf{P}; \mathscr{R}, \mu)$ given in equation (2) in the paper admits a unique minimizer on the space of Hilbert–Schmidt operators acting on $\mathcal{H}$. By the triangle inequality, monotonicity and convexity of ρ we have that

$$\begin{aligned} &E_{\mathsf{P}}(\rho[\|\mathscr{P}(X;\mu) - \{\lambda\mathscr{R} + (1-\lambda)\mathscr{R}'\}\|] - \rho\{\|\mathscr{P}(X;\mu)\|\}) \\ &\qquad \le E_{\mathsf{P}}[\rho\{\lambda\|\mathscr{P}(X;\mu) - \mathscr{R}\| + (1-\lambda)\|\mathscr{P}(X;\mu) - \mathscr{R}'\|\} - \rho\{\|\mathscr{P}(X;\mu)\|\}] \\ &\qquad \le \lambda\, E_{\mathsf{P}}[\rho\{\|\mathscr{P}(X;\mu) - \mathscr{R}\|\} - \rho\{\|\mathscr{P}(X;\mu)\|\}] \\ &\qquad\quad + (1-\lambda)\, E_{\mathsf{P}}[\rho\{\|\mathscr{P}(X;\mu) - \mathscr{R}'\|\} - \rho\{\|\mathscr{P}(X;\mu)\|\}] \end{aligned}$$

for any $\lambda \in [0,1]$ and arbitrary Hilbert–Schmidt operators $\mathscr{R}, \mathscr{R}'$. Notice that since ρ is strictly increasing, the first inequality is strict unless $\mathscr{P}(X;\mu) - \mathscr{R}$ and $\mathscr{P}(X;\mu) - \mathscr{R}'$ are collinear almost surely. Equivalently, the inequality is strict whenever the distribution of $\mathscr{P}(X;\mu)$ is not concentrated on the line $\{t\mathscr{R} + (1-t)\mathscr{R}' : t \in \mathbb{R}\}$.

We now investigate what this condition means geometrically in the space $\mathcal{H}$. First, notice that as the rank of $\mathscr{P}(X;\mu)$ is 1, the rank of $t\mathscr{R} + (1-t)\mathscr{R}'$ has to be 1 also. Now we distinguish two cases.

First, if $\mathscr{R}, \mathscr{R}'$ are collinear, then the line is of the form $\{\alpha\mathscr{R} : \alpha \in \mathbb{R}\}$, which by the condition on the rank is $\{\alpha u \otimes u : \alpha \in \mathbb{R}\}$ for some $u \in \mathcal{H}$. Since $\mathscr{P}(X;\mu)$ is positive semidefinite, we in fact have $\{\alpha u \otimes u : \alpha \ge 0\}$. Thus, the operator $\mathscr{P}(X;\mu)$ lying on this line is equivalent to X lying on the line $\{\mu + \beta u : \beta \in \mathbb{R}\}$.

Second, if $\mathscr{R}, \mathscr{R}'$ are not collinear, then operators of the form $t\mathscr{R} + (1-t)\mathscr{R}'$ have rank 1 for at most two values of t. To see this, notice that the rank condition implies that for all $i < j$,

$$\det\left\{ t \begin{pmatrix} R_{ii} & R_{ij} \\ R_{ji} & R_{jj} \end{pmatrix} + (1-t) \begin{pmatrix} R'_{ii} & R'_{ij} \\ R'_{ji} & R'_{jj} \end{pmatrix} \right\} = 0,$$

where $R_{ij} = \langle e_i, \mathscr{R} e_j \rangle$, $R'_{ij} = \langle e_i, \mathscr{R}' e_j \rangle$. This system of quadratic equations has at most two solutions. Thus, the set $\{t\mathscr{R} + (1-t)\mathscr{R}' : t \in \mathbb{R}\}$ reduces at most to the set $\{\alpha_1 u_1 \otimes u_1, \alpha_2 u_2 \otimes$

$u_2\}$ for some nonnegative α_1, α_2 and some $u_1, u_2 \in \mathcal{H}$. Hence, the operator $\mathscr{P}(X;\mu)$ belonging to this set is equivalent to X belonging to the set of at most four points $\{\mu \pm \beta_1 u_1, \mu \pm \beta_2 u_2\}$.

Therefore, if the distribution P is not concentrated on a line or on four points, the objective function to be minimized is strictly convex. It follows that the minimum of the functional exists and is unique.

PROOF OF COROLLARY 1

The empirical version of the functional defining the dispersion operator is the expectation with respect to the empirical distribution $\hat{\mathsf{P}}$. Under our assumptions on P, the empirical distribution $\hat{\mathsf{P}}$ is almost surely not concentrated on a line or on four points. Therefore, strict convexity, and thus existence and uniqueness, follows with probability 1 by applying Proposition 1 to the empirical distribution $\hat{\mathsf{P}}$. Consistency then follows from strict convexity and the consistency of $\hat{\mu}$, using standard arguments.

PROOF OF PROPOSITION 2

Consider $\mathscr{R}$ of the form $\sum_{k=1}^{\infty} \delta_k \varphi_k \otimes \varphi_k$ for some sequence $\delta_1, \delta_2, \ldots$ We will prove that such an operator solves the estimating equation (5) showing that $\mathscr{R}$ and $\mathscr{C}$ have the same set of eigenfunctions, and that the sequence $\delta_1, \delta_2, \ldots$ satisfies the condition (6).

We investigate the coordinates of the left-hand side of (5), with the aim of showing that the values

$$\left\langle \varphi_j, E_{\mathsf{P}}\left[\frac{\rho'\{\|\mathscr{R}-\mathscr{P}(X;\mu)\|\}}{\|\mathscr{R}-\mathscr{P}(X;\mu)\|}\{\mathscr{R}-\mathscr{P}(X;\mu)\}\right]\varphi_k\right\rangle \tag{S1}$$

are zero for all j, k. By the orthonormality of $\varphi_1, \varphi_2, \ldots$, we have that

$$\begin{aligned}\|\mathscr{R}-\mathscr{P}(X;\mu)\|^2 &= \left\|\sum_{k=1}^{\infty}\delta_k\varphi_k\otimes\varphi_k - \sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\lambda_j^{1/2}\lambda_k^{1/2}\beta_j\beta_k\varphi_j\otimes\varphi_k\right\|^2 \\ &= \sum_k(\delta_k-\lambda_k\beta_k^2)^2 + \sum_{k\neq j}\lambda_j\lambda_k\beta_j^2\beta_k^2.\end{aligned}$$

First, we compute the off-diagonal coordinates with $j \neq k$. The first summand in (S1) is zero because $\langle\varphi_j, \mathscr{R}\varphi_k\rangle = 0$. To show that the second summand in (S1) is zero, we use the fact that, by assumption, the sequence $\{s_i\beta_i\}_{i=1}^{\infty}$ with $s_i = (-1)^{1\{i=j\}}$ has the same joint distribution as $\{\beta_i\}_{i=1}^{\infty}$. Compute

$$\begin{aligned}A_{jk} &= \left\langle \varphi_j, E_{\mathsf{P}}\left[\frac{\rho'\{\|\mathscr{R}-\mathscr{P}(X;\mu)\|\}}{\|\mathscr{R}-\mathscr{P}(X;\mu)\|}\mathscr{P}(X;\mu)\right]\varphi_k\right\rangle \\ &= E\left(\frac{\rho'[\{\sum_i(\delta_i-\lambda_i\beta_i^2)^2+\sum_{i\neq l}\lambda_i\lambda_l\beta_i^2\beta_l^2\}^{1/2}]}{\{\sum_i(\delta_i-\lambda_i\beta_i^2)^2+\sum_{i\neq l}\lambda_i\lambda_l\beta_i^2\beta_l^2\}^{1/2}}\lambda_j^{1/2}\lambda_k^{1/2}\beta_j\beta_k\right) \\ &= E\left\{\frac{\rho'([\sum_i\{\delta_i-\lambda_i(s_i\beta_i)^2\}^2+\sum_{i\neq l}\lambda_i\lambda_l(s_i\beta_i)^2(s_l\beta_l)^2]^{1/2})}{[\sum_i\{\delta_i-\lambda_i(s_i\beta_i)^2\}^2+\sum_{i\neq l}\lambda_i\lambda_l(s_i\beta_i)^2(s_l\beta_l)^2]^{1/2}}\lambda_j^{1/2}\lambda_k^{1/2}s_j\beta_j s_k\beta_k\right\} \\ &= -A_{jk}.\end{aligned}$$

Thus, $A_{jk} = 0$. Therefore, the operator $\mathscr{R}$ is diagonalized by the same functions $\varphi_1, \varphi_2, \ldots$ as $\mathscr{C}$. By computing the diagonal coordinates with $j = k$ in (5) we obtain (6).

A TECHNICAL LEMMA

LEMMA 1. *Under the assumptions of Theorem 1,*

(a) the linear operator $\mathscr{D}(\mathsf{P};\mu)$ defined in equation (A1) is a bijection of $\mathcal{H}$ onto itself, it is bounded and has bounded inverse,

(b) the linear operator $\mathfrak{D}(\mathsf{P};\mathscr{R},\mu)$ defined in equation (A2) is a bijection of $\mathrm{HS}(\mathcal{H},\mathcal{H})$ onto itself, it is bounded and has bounded inverse.

Proof. We prove part (a); the proof of part (b) is similar. The proof uses and extends the steps of the proof of Lemma 1 (iii) of Gervini (2008) modified for the present context of general ρ and generalized to the case of infinitely many components in the Karhunen–Loève expansion.

Recall that

$$\mathscr{D}(\mathsf{P};\mu) = E_{\mathsf{P}}\left[\frac{\rho'(\|X-\mu\|)}{\|X-\mu\|}\mathscr{I} + \left\{\frac{\rho''(\|X-\mu\|)}{\|X-\mu\|^2} - \frac{\rho'(\|X-\mu\|)}{\|X-\mu\|^3}\right\}\mathscr{P}(X;\mu)\right];$$

see the appendix of the main body of the paper. To show that $\mathscr{D}(\mathsf{P};\mu)$ is a bijection, we need to find for any $h \in \mathcal{H}$ a unique element $f \in \mathcal{H}$ such that $\mathscr{D}(\mathsf{P};\mu)f = h$. The set of orthonormal eigenfunctions $\{\varphi_k\}_{k=1}^{\infty}$ of $\mathscr{C}$ can be extended to an orthonormal basis of $\mathcal{H}$ by possibly adding some functions $\{\psi_k\}_{k=1}^{q}$ with q finite or infinite or zero. It is then enough to verify the relation $\mathscr{D}(\mathsf{P};\mu)f = h$ in terms of the Fourier coefficients of both sides with respect to the basis $\{\varphi_k\}_{k=1}^{\infty} \cup \{\psi_k\}_{k=1}^{q}$, i.e., to show that $\langle \mathscr{D}(\mathsf{P};\mu)f, \varphi_k\rangle = \langle h, \varphi_k\rangle$ for all $k = 1, 2, \ldots$ and $\langle \mathscr{D}(\mathsf{P};\mu)f, \psi_k\rangle = \langle h, \psi_k\rangle$ for all $k = 1, \ldots, q$. As $\langle \mathscr{D}(\mathsf{P};\mu)f, \varphi_k\rangle = \langle f, \mathscr{D}(\mathsf{P};\mu)\varphi_k\rangle$ and $\langle \mathscr{D}(\mathsf{P};\mu)f, \psi_k\rangle = \langle f, \mathscr{D}(\mathsf{P};\mu)\psi_k\rangle$, we first investigate $\mathscr{D}(\mathsf{P};\mu)\varphi_k$ and $\mathscr{D}(\mathsf{P};\mu)\psi_k$.

We begin by exploring the structure of the operator $\mathscr{D}(\mathsf{P};\mu)$. We can rewrite

$$E_{\mathsf{P}}\left\{\frac{\rho'(\|X-\mu\|)}{\|X-\mu\|^3}\mathscr{P}(X;\mu)\right\} = E_{\mathsf{P}}(\tilde{\varepsilon}\otimes\tilde{\varepsilon}),$$

where

$$\tilde{\varepsilon} = \frac{\rho'(\|X-\mu\|)^{1/2}}{\|X-\mu\|^{3/2}}(X-\mu) = \sum_{k=1}^{\infty}\lambda_k^{1/2}\frac{\rho'(\|X-\mu\|)^{1/2}}{\|X-\mu\|^{3/2}}\beta_k\varphi_k = \sum_{k=1}^{\infty}\tilde{\lambda}_k^{1/2}\tilde{\beta}_k\varphi_k \qquad \text{(S2)}$$

with

$$\tilde{\lambda}_k = \lambda_k\, E_{\mathsf{P}}\left\{\frac{\rho'(\|X-\mu\|)}{\|X-\mu\|^3}\beta_k^2\right\},$$

$$\tilde{\beta}_k = \frac{\rho'(\|X-\mu\|)^{1/2}}{\|X-\mu\|^{3/2}}\beta_k \Big/ \left[E_{\mathsf{P}}\left\{\frac{\rho'(\|X-\mu\|)}{\|X-\mu\|^3}\beta_k^2\right\}\right]^{1/2}.$$

Thus, we need to find the covariance operator of $\tilde{\varepsilon}$. The series expansion (S2) of $\tilde{\varepsilon}$ is a Karhunen–Loève expansion because the coefficients $\tilde{\beta}_k$ have zero mean and unit variance and are uncorrelated (which follows from the fact that the distribution of $\{\beta_k\}$ is invariant under the change of the sign of any component). Therefore, since $E_{\mathsf{P}}(\|\tilde{\varepsilon}\|^2) < \infty$, which follows immediately from the assumption that

$$E_{\mathsf{P}}\left\{\frac{\rho'(\|X-\mu\|)}{\|X-\mu\|}\right\} < \infty,$$

the operator of interest, as the covariance operator of $\tilde{\varepsilon}$, takes the form

$$E_{\mathsf{P}}\left\{\frac{\rho'(\|X-\mu\|)}{\|X-\mu\|^3}\mathscr{P}(X;\mu)\right\} = \sum_{k=1}^{\infty}\tilde{\lambda}_k\varphi_k\otimes\varphi_k = \sum_{k=1}^{\infty}E_{\mathsf{P}}\left\{\frac{\rho'(\|X-\mu\|)}{\|X-\mu\|^3}\lambda_k\beta_k^2\right\}\varphi_k\otimes\varphi_k.$$

Using analogous arguments for

$$\dot{\varepsilon} = \frac{\rho''(\|X-\mu\|)^{1/2}}{\|X-\mu\|}(X-\mu),$$

we can show that

$$E_{\mathsf{P}}\left\{\frac{\rho''(\|X-\mu\|)}{\|X-\mu\|^2}\mathscr{P}(X;\mu)\right\} = \sum_{k=1}^{\infty}E_{\mathsf{P}}\left\{\frac{\rho''(\|X-\mu\|)}{\|X-\mu\|^2}\lambda_k\beta_k^2\right\}\varphi_k\otimes\varphi_k.$$

Hence, we finally obtain $\mathscr{D}(\mathsf{P};\mu)$ in the form

$$\begin{aligned}\mathscr{D}(\mathsf{P};\mu) = {} & E_{\mathsf{P}}\left\{\frac{\rho'(\|X-\mu\|)}{\|X-\mu\|}\right\}\mathscr{I} \\ & + \sum_{k=1}^{\infty}E_{\mathsf{P}}\left[\left\{\frac{\rho''(\|X-\mu\|)}{\|X-\mu\|^2} - \frac{\rho'(\|X-\mu\|)}{\|X-\mu\|^3}\right\}\lambda_k\beta_k^2\right]\varphi_k\otimes\varphi_k.\end{aligned}$$

Therefore, for $k = 1, 2, \dots$ we have

$$\mathscr{D}(\mathsf{P};\mu)\varphi_k = E_{\mathsf{P}}\left\{\frac{\rho'(\|X-\mu\|)}{\|X-\mu\|}\right\}\varphi_k + E_{\mathsf{P}}\left[\left\{\frac{\rho''(\|X-\mu\|)}{\|X-\mu\|^2} - \frac{\rho'(\|X-\mu\|)}{\|X-\mu\|^3}\right\}\lambda_k\beta_k^2\right]\varphi_k$$

and, for $k = 1, \dots, q$, we have

$$\mathscr{D}(\mathsf{P};\mu)\psi_k = E_{\mathsf{P}}\left\{\frac{\rho'(\|X-\mu\|)}{\|X-\mu\|}\right\}\psi_k.$$

Thus, we obtain

$$\begin{aligned}\langle\mathscr{D}(\mathsf{P};\mu)f,\varphi_k\rangle &= \nu_k\langle f,\varphi_k\rangle \quad (k=1,2,\dots), \\ \langle\mathscr{D}(\mathsf{P};\mu)f,\psi_k\rangle &= \eta\langle f,\psi_k\rangle \quad (k=1,\dots,q),\end{aligned}$$

where

$$\nu_k = E_{\mathsf{P}}\left\{\frac{\rho'(\|X-\mu\|)}{\|X-\mu\|}\right\} + \lambda_k\, E_{\mathsf{P}}\left[\left\{\frac{\rho''(\|X-\mu\|)}{\|X-\mu\|^2} - \frac{\rho'(\|X-\mu\|)}{\|X-\mu\|^3}\right\}\beta_k^2\right] \quad (k=1,2,\dots)$$

and

$$\eta = E_{\mathsf{P}}\left\{\frac{\rho'(\|X-\mu\|)}{\|X-\mu\|}\right\}.$$

So f, the candidate for $\mathscr{D}(\mathsf{P};\mu)^{-1}h$, should have Fourier coefficients $\langle f,\varphi_k\rangle$, $\langle f,\psi_k\rangle$ satisfying the system of equations

$$\nu_k\langle f,\varphi_k\rangle = \langle h,\varphi_k\rangle \quad (k=1,2,\dots), \quad \eta\langle f,\psi_k\rangle = \langle h,\psi_k\rangle \quad (k=1,\dots,q).$$

To be able to write $\langle f,\varphi_k\rangle = \langle h,\varphi_k\rangle/\nu_k$, we need to show that ν_k $(k=1,2,\dots)$ and η are nonzero and finite. Then, f will be uniquely determined by the formula

$$f = \sum_{k=1}^{\infty}\frac{\langle h,\varphi_k\rangle}{\nu_k}\varphi_k + \sum_{k=1}^{q}\frac{\langle h,\psi_k\rangle}{\eta}\psi_k$$

provided that f is a well-defined element of $\mathcal{H}$, that is,

$$\|f\|^2 = \sum_{k=1}^{\infty} \frac{\langle h, \varphi_k \rangle^2}{\nu_k^2} + \sum_{k=1}^{q} \frac{\langle h, \psi_k \rangle^2}{\eta^2} < \infty. \tag{S3}$$

We assumed that $\eta < \infty$ and we immediately see that $\eta > 0$ because ρ is strictly increasing. We now deal with ν_k ($k = 1, 2, \dots$). We will show that there exist $0 < a \leq b < \infty$ such that $\nu_k \in [a, b]$ for all $k = 1, 2, \dots$

First we establish the lower bound a. Using the Karhunen–Loève expansion (S2) we can rewrite

$$E_{\mathsf{P}}\left\{ \frac{\rho'(\|X - \mu\|)}{\|X - \mu\|} \right\} = E_{\mathsf{P}}(\|\tilde{\varepsilon}\|^2) = \sum_{k=1}^{\infty} \tilde{\lambda}_k = \sum_{k=1}^{\infty} \lambda_k \, E_{\mathsf{P}}\left\{ \frac{\rho'(\|X - \mu\|)}{\|X - \mu\|^3} \beta_k^2 \right\}. \tag{S4}$$

Each term in the series on the right hand side of (S4) is obviously positive and by finiteness of the left hand side it is finite, and thus the differences

$$E_{\mathsf{P}}\left\{ \frac{\rho'(\|X - \mu\|)}{\|X - \mu\|} \right\} - \lambda_k \, E_{\mathsf{P}}\left\{ \frac{\rho'(\|X - \mu\|)}{\|X - \mu\|^3} \beta_k^2 \right\}, \tag{S5}$$

which appear in the expression for ν_k, are positive and bounded away from zero by a constant a. The remaining term

$$\lambda_k \, E_{\mathsf{P}}\left\{ \frac{\rho''(\|X - \mu\|)}{\|X - \mu\|^2} \beta_k^2 \right\} \tag{S6}$$

appearing in ν_k is nonnegative as $\rho'' \geq 0$ because ρ is convex. It follows that $\nu_k \geq a$ for all $k = 1, 2, \dots$

Now we find the upper bound b. By applying the same idea as in (S4) to $\dot{\varepsilon}$, we obtain

$$E_{\mathsf{P}}\{\rho''(\|X - \mu\|)\} = \sum_{k=1}^{\infty} \lambda_k \, E_{\mathsf{P}}\left\{ \frac{\rho''(\|X - \mu\|)}{\|X - \mu\|^2} \beta_k^2 \right\}. \tag{S7}$$

In view of (S7), the terms (S6) are smaller than or equal to $E_{\mathsf{P}}\{\rho''(\|X - \mu\|)\}$. The differences (S5) are smaller than

$$E_{\mathsf{P}}\left\{ \frac{\rho'(\|X - \mu\|)}{\|X - \mu\|} \right\}.$$

Therefore, we have that $\nu_k \leq b$ for all $k = 1, 2, \dots$ with

$$b = E_{\mathsf{P}}\left\{ \frac{\rho'(\|X - \mu\|)}{\|X - \mu\|} \right\} + E_{\mathsf{P}}\{\rho''(\|X - \mu\|)\}.$$

Finally, it remains to show (S3), which is now straightforward because

$$\|f\|^2 = \sum_{k=1}^{\infty} \frac{\langle h, \varphi_k \rangle^2}{\nu_k^2} + \sum_{k=1}^{q} \frac{\langle h, \psi_k \rangle^2}{\eta^2} \leq \frac{\sum_{k=1}^{\infty} \langle h, \varphi_k \rangle^2 + \sum_{k=1}^{q} \langle h, \psi_k \rangle^2}{\min(a, \eta)} = \frac{\|h\|^2}{\min(a, \eta)} < \infty.$$

This shows that f is a well defined element of $\mathcal{H}$ and thus the linear operator $\mathscr{D}(\mathsf{P}; \mu)$ is a bijection of $\mathcal{H}$ onto itself. It also shows that the inverse $\mathscr{D}(\mathsf{P}; \mu)^{-1}$ is a bounded operator. Hence also the operator $\mathscr{D}(\mathsf{P}; \mu)$ is bounded by the bounded inverse theorem or by direct verification. □

Remark: As ν_k are bounded away from zero and bounded from above, the operator $\mathscr{D}(\mathsf{P};\mu)$ is only a small perturbation of a multiple of the identity. This gives an intuitive explanation why it inherits its bijectivity and boundedness.

PROOF OF THEOREM 1

It is enough to prove the weak convergence of $n^{1/2}\mathscr{B}(\hat{\mathsf{P}}_1,\hat{\mathsf{P}}_2,a_n;\hat{\mathscr{R}},\hat{\mu}_1,\hat{\mu}_2)$. The weak convergence of the vector with components S_l will then follow directly from Slutsky's theorem. The continuous mapping theorem and Slutsky's theorem will then establish the weak convergence of the statistic T. Applying a Taylor expansion (Nelson, 1969, Theorem 6, p. 12) of $\mathscr{B}(\hat{\mathsf{P}}_1,\hat{\mathsf{P}}_2,a_n;\hat{\mathscr{R}},\hat{\mu}_1,\hat{\mu}_2)$ around the true values of the parameters yields

$$\begin{aligned} n^{1/2}\mathscr{B}(\hat{\mathsf{P}}_1,\hat{\mathsf{P}}_2,a_n;\hat{\mathscr{R}},\hat{\mu}_1,\hat{\mu}_2) &= n^{1/2}\mathscr{B}(\hat{\mathsf{P}}_1,\hat{\mathsf{P}}_2,a_n;\mathscr{R},\mu_1,\mu_2) \\ &\quad + \mathfrak{D}_1(\hat{\mathsf{P}}_1,\hat{\mathsf{P}}_2,a_n;\mathscr{R}^\diamond,\mu_1^\diamond,\mu_2^\diamond)n^{1/2}(\hat{\mathscr{R}}-\mathscr{R}) \\ &\quad + a_n^{1/2}\mathbb{D}(\hat{\mathsf{P}}_1;\mathscr{R}^\diamond,\mu_1^\diamond)n_1^{1/2}(\hat{\mu}_1-\mu_1) \\ &\quad - (1-a_n)^{1/2}\mathbb{D}(\hat{\mathsf{P}}_2;\mathscr{R}^\diamond,\mu_2^\diamond)n_2^{1/2}(\hat{\mu}_2-\mu_2), \end{aligned} \tag{S8}$$

where

$$\begin{aligned} \mathfrak{D}_1(\mathsf{P}_1,\mathsf{P}_2,a;\mathscr{R},\mu_1,\mu_2) &= \frac{\partial}{\partial\mathscr{R}}\mathscr{B}(\mathsf{P}_1,\mathsf{P}_2,a;\mathscr{R},\mu_1,\mu_2) \\ &= a\mathfrak{D}(\mathsf{P}_1;\mathscr{R},\mu_1)-(1-a)\mathfrak{D}(\mathsf{P}_2;\mathscr{R},\mu_2) \end{aligned}$$

and

$$\mathfrak{D}(\mathsf{P};\mathscr{R},\mu)=\frac{\partial}{\partial\mathscr{R}}\mathscr{G}(\mathsf{P};\mathscr{R},\mu),\quad \mathbb{D}(\mathsf{P};\mathscr{R},\mu)=\frac{\partial}{\partial\mu}\mathscr{G}(\mathsf{P};\mathscr{R},\mu).$$

See the Appendix in the main body of the paper for explicit formulae.

We now turn to develop certain asymptotic representations for $\hat{\mu}_1$, $\hat{\mu}_2$ and $\hat{\mathscr{R}}$. Using the Taylor expansion, law of large numbers and consistency of $\hat{\mu}_1$ we get

$$\begin{aligned} 0=n_1^{1/2}G(\hat{\mathsf{P}}_1;\hat{\mu}_1) &= n_1^{1/2}G(\hat{\mathsf{P}}_1;\mu_1)+\mathscr{D}(\hat{\mathsf{P}}_1;\mu_1^\dagger)n_1^{1/2}(\hat{\mu}_1-\mu_1) \\ &= n_1^{1/2}G(\hat{\mathsf{P}}_1;\mu_1)+\mathscr{D}(\mathsf{P}_1;\mu_1)n_1^{1/2}(\hat{\mu}_1-\mu_1)+o_P(1), \end{aligned}$$

where the term $o_P(1)$ is due to the fact that we replace $\mathscr{D}(\hat{\mathsf{P}}_1;\mu_1)$ by its limit $\mathscr{D}(\mathsf{P}_1;\mu_1)$. From this and an analogous expansion for μ_2 we obtain

$$\begin{aligned} n_1^{1/2}(\hat{\mu}_1-\mu_1) &= -\mathscr{D}(\mathsf{P}_1;\mu_1)^{-1}n_1^{1/2}G(\hat{\mathsf{P}}_1;\mu_1)+o_P(1), \\ n_2^{1/2}(\hat{\mu}_2-\mu_2) &= -\mathscr{D}(\mathsf{P}_2;\mu_2)^{-1}n_2^{1/2}G(\hat{\mathsf{P}}_2;\mu_2)+o_P(1). \end{aligned} \tag{S9}$$

The existence of the bounded inverse operators in the above equations, as well as of other inverse operators appearing later in the proof, is shown in Lemma 1. The Taylor expansion of the estimating score for $\mathscr{R}$ around the true values is

$$\begin{aligned} \mathscr{O}=n^{1/2}\mathscr{G}(\hat{\mathsf{P}}_1,\hat{\mathsf{P}}_2,a_n;\hat{\mathscr{R}},\hat{\mu}_1,\hat{\mu}_2) &= n^{1/2}\mathscr{G}(\hat{\mathsf{P}}_1,\hat{\mathsf{P}}_2,a_n;\mathscr{R},\mu_1,\mu_2) \\ &\quad + \mathfrak{D}_0(\hat{\mathsf{P}}_1,\hat{\mathsf{P}}_2,a_n;\mathscr{R}^\ddagger,\mu_1^\ddagger,\mu_2^\ddagger)n^{1/2}(\hat{\mathscr{R}}-\mathscr{R}) \\ &\quad + a_n^{1/2}\mathbb{D}(\hat{\mathsf{P}}_1;\mathscr{R}^\ddagger,\mu_1^\ddagger)n_1^{1/2}(\hat{\mu}_1-\mu_1) \\ &\quad + (1-a_n)^{1/2}\mathbb{D}(\hat{\mathsf{P}}_2;\mathscr{R}^\ddagger,\mu_2^\ddagger)n_2^{1/2}(\hat{\mu}_2-\mu_2), \end{aligned}$$

where $\mathfrak{D}_0(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2) = a\mathfrak{D}(\mathsf{P}_1; \mathscr{R}, \mu_1) + (1-a)\mathfrak{D}(\mathsf{P}_2; \mathscr{R}, \mu_2)$. This yields

$$\begin{aligned} n^{1/2}(\hat{\mathscr{R}} - \mathscr{R}) = &-\mathfrak{D}_0(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)^{-1} \\ &\{n^{1/2}\mathscr{G}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \mathscr{R}, \mu_1, \mu_2) + a_n^{1/2}\mathbb{D}(\mathsf{P}_1; \mathscr{R}^\ddagger, \mu_1^\ddagger) n_1^{1/2}(\hat{\mu}_1 - \mu_1) \\ &+ (1-a_n)^{1/2}\mathbb{D}(\mathsf{P}_2; \mathscr{R}^\ddagger, \mu_2^\ddagger) n_2^{1/2}(\hat{\mu}_2 - \mu_2)\} \\ &+ o_P(1); \end{aligned} \tag{S10}$$

here again the term $o_P(1)$ is present because we replace the empirical distributions by their theoretical counterparts in $\mathfrak{D}_0$ and $\mathbb{D}$.

The different Taylor expansions we have used contain various elements denoted by $^\diamond$, †, ‡ which lie on the line segments between the true and estimated corresponding parameters. We will replace all of these elements by the true values of the parameters. Due to the consistency of the estimators, the difference between a quantity at the true value of the parameters and at a value on the line segment between the true value and the estimator converges in probability to zero. Moreover, the quantities involving elements marked with $^\diamond$, † or ‡ are always multiplied by a term that is bounded in probability (by its convergence in distribution which will be seen later). Hence, the change we make by replacing the elements marked with $^\diamond$, † or ‡ by their true values is asymptotically negligible. The reason for doing this is that we obtain simpler formulas.

Denote

$$\begin{aligned} \mathfrak{H}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2) &= \mathfrak{I} - \mathfrak{D}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)\mathfrak{D}_0(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)^{-1}, \\ \mathbb{H}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2) &= \mathfrak{H}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)\mathbb{D}(\mathsf{P}_1; \mathscr{R}, \mu_1)\mathscr{D}(\mathsf{P}_1; \mu_1)^{-1}, \\ \mathfrak{H}_2(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2) &= \mathfrak{I} + \mathfrak{D}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)\mathfrak{D}_0(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)^{-1}, \\ \mathbb{H}_2(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2) &= \mathfrak{H}_2(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)\mathbb{D}(\mathsf{P}_2; \mathscr{R}, \mu_2)\mathscr{D}(\mathsf{P}_2; \mu_2)^{-1}, \end{aligned}$$

where $\mathfrak{I}$ stands for the identity operator on $\mathrm{HS}(\mathcal{H}, \mathcal{H})$. Inserting (S9) and (S10) into (S8), we obtain

$$\begin{aligned} n^{1/2}\mathscr{B}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2) = &\, a_n^{1/2}\mathfrak{H}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2) n_1^{1/2}\mathscr{G}(\hat{\mathsf{P}}_1; \mathscr{R}, \mu_1) \\ &- a_n^{1/2}\mathbb{H}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2) n_1^{1/2} G(\hat{\mathsf{P}}_1; \mu_1) \\ &- (1-a_n)^{1/2}\mathfrak{H}_2(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2) n_2^{1/2}\mathscr{G}(\hat{\mathsf{P}}_2; \mathscr{R}, \mu_2) \\ &+ (1-a_n)^{1/2}\mathbb{H}_2(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2) n_2^{1/2} G(\hat{\mathsf{P}}_2; \mu_2) \\ &+ o_P(1). \end{aligned}$$

The term $o_P(1)$ is due to the fact that we have replaced the quantities marked with $^\diamond$, †, ‡ by their true counterparts.

By the central limit theorem for Hilbert spaces (Bosq, 2000, Theorem 2.7), the operators $n_1^{1/2}\mathscr{G}(\hat{\mathsf{P}}_1; \mathscr{R}, \mu_1)$, $n_1^{1/2}G(\hat{\mathsf{P}}_1; \mu_1)$ jointly converge in distribution to a zero-mean Gaussian random variable in $\mathrm{HS}(\mathcal{H}, \mathcal{H}) \times \mathcal{H}$. The asymptotic covariance operator of $n_1^{1/2}\mathscr{G}(\hat{\mathsf{P}}_1; \mathscr{R}, \mu_1)$, i.e., an operator on operators on $\mathcal{H}$, can be estimated by the empirical covariance $\mathfrak{J}(\hat{\mathsf{P}}_1; \hat{\mathscr{R}}, \hat{\mu}_1)$, where

$$\mathfrak{J}(\mathsf{P}; \mathscr{R}, \mu) = E_{\mathsf{P}}\left(\left[\frac{\rho'\{\|\mathscr{P}(X;\mu) - \mathscr{R}\|\}}{\|\mathscr{P}(X;\mu) - \mathscr{R}\|}\{\mathscr{R} - \mathscr{P}(X;\mu)\} - \mathscr{G}(\mathsf{P}; \mathscr{R}, \mu)\right]^{\otimes 2}\right)$$

with the notation $\mathscr{A}^{\otimes 2} = \mathscr{A} \otimes \mathscr{A}$ for $\mathscr{A} \in \mathrm{HS}(\mathcal{H}, \mathcal{H})$, the asymptotic covariance operator of $n_1^{1/2} G(\hat{\mathsf{P}}_1; \mu_1)$, i.e., an operator on $\mathcal{H}$, can be estimated by $\hat{\mathscr{J}}(\hat{\mathsf{P}}_1; \hat{\mu}_1)$, where

$$\mathscr{J}(\mathsf{P}; \mu) = E_{\mathsf{P}}\left[\left\{\frac{\rho'(\|X-\mu\|)}{\|X-\mu\|}(\mu - X) - G(\mathsf{P}; \mu)\right\}^{\otimes 2}\right]$$

with $f^{\otimes 2} = f \otimes f$ for $f \in \mathcal{H}$, and the asymptotic cross-covariance operator of $n_1^{1/2}\mathscr{G}(\hat{\mathsf{P}}_1; \mathscr{R}, \mu_1)$ and $n_1^{1/2} G(\hat{\mathsf{P}}_1; \mu_1)$, i.e., an operator from $\mathcal{H}$ to operators on $\mathcal{H}$, can be estimated by $\mathbb{J}(\hat{\mathsf{P}}_1; \hat{\mathscr{R}}, \hat{\mu}_1)$, where

$$\begin{aligned}\mathbb{J}(\mathsf{P}; \mathscr{R}, \mu) = E_{\mathsf{P}}\Bigg(&\left[\frac{\rho'\{\|\mathscr{P}(X;\mu) - \mathscr{R}\|\}}{\|\mathscr{P}(X;\mu) - \mathscr{R}\|}\{\mathscr{R} - \mathscr{P}(X;\mu)\} - \mathscr{G}(\mathsf{P}; \mathscr{R}, \mu)\right] \\ &\otimes \left\{\frac{\rho'(\|X-\mu\|)}{\|X-\mu\|}(\mu - X) - G(\mathsf{P};\mu)\right\}\Bigg).\end{aligned}$$

Similarly, $n_2^{1/2}\mathscr{G}(\hat{\mathsf{P}}_2; \mathscr{R}, \mu_2)$, $n_2^{1/2} G(\hat{\mathsf{P}}_2; \mu_2)$ jointly converge in distribution to a zero-mean Gaussian random element with covariance estimators analogous to those mentioned above for the sample from P_1. As the samples are independent, all four random variables jointly converge in distribution.

Finally, it follows by Slutsky's theorem that the test operator $n^{1/2}\mathscr{B}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2)$ is asymptotically distributed as a zero-mean Gaussian operator whose covariance operator can be consistently estimated by

$$\begin{aligned}\mathfrak{W}(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2) = a_n \mathfrak{W}_1(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2) \\ + (1 - a_n)\mathfrak{W}_2(\hat{\mathsf{P}}_1, \hat{\mathsf{P}}_2, a_n; \hat{\mathscr{R}}, \hat{\mu}_1, \hat{\mu}_2),\end{aligned}$$

where

$$\begin{aligned}&\mathfrak{W}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2) \\ &\quad= \mathfrak{H}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)\mathfrak{J}(\mathsf{P}_1; \mathscr{R}, \mu_1)\mathfrak{H}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)^* \\ &\quad\quad - \mathfrak{H}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)\mathbb{J}(\mathsf{P}_1; \mathscr{R}, \mu_1)\mathbb{H}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)^* \\ &\quad\quad - \mathbb{H}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)\mathbb{J}(\mathsf{P}_1; \mathscr{R}, \mu_1)^*\mathfrak{H}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)^* \\ &\quad\quad + \mathbb{H}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)\mathscr{J}(\mathsf{P}_1; \mathscr{R}, \mu_1)\mathbb{H}_1(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)^*\end{aligned}$$

with * denoting adjoint operators, and $\mathfrak{W}_2(\mathsf{P}_1, \mathsf{P}_2, a; \mathscr{R}, \mu_1, \mu_2)$ is defined analogously with $\mathbb{H}_2, \mathfrak{H}_2$ in place of $\mathbb{H}_1, \mathfrak{H}_1$, respectively, and P_2 instead of P_1 in $\mathfrak{J}, \mathbb{J}, \mathscr{J}$.

REFERENCES

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

GERVINI, D. (2008). Robust functional estimation using the median and spherical principal components. *Biometrika* **95**, 587–600.

NELSON, E. (1969). *Topics in Dynamics. I: Flows*. Princeton: Princeton University Press.

C. Components and completion of partially observed functional data

By David Kraus

Journal of the Royal Statistical Society. Series B. Statistical Methodology, 77(4):777–801, 2015

DOI: 10.1111/rssb.12087

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2015)
77, *Part* 4, *pp.* 777–801

Components and completion of partially observed functional data

David Kraus

University Hospital Lausanne, Switzerland

[Received June 2013. Final revision July 2014]

Summary. Functional data are traditionally assumed to be observed on the same domain. Motivated by a data set of heart rate temporal profiles, we develop methodology for the analysis of incomplete functional samples where each curve may be observed on a subset of the domain and unobserved elsewhere. We formalize this observation regime and develop the fundamental procedures of functional data analysis for this framework: estimation of parameters (mean and covariance operator) and principal component analysis. Principal scores of a partially observed function cannot be computed directly and we solve this challenging issue by estimating their best predictions as linear functionals of the observed part of the trajectory. Next, we propose a functional completion procedure that recovers the missing part by using the observed part of the curve. We construct prediction intervals for principal scores and bands for missing parts of trajectories. The prediction problems are seen to be ill-posed inverse problems; regularization techniques are used to obtain a stable solution. A simulation study shows the good performance of our methods. We illustrate the methods on the heart rate data and provide practical computational algorithms and theoretical arguments and proofs of all results.

Keywords: Functional data analysis; Incomplete observation; Inverse problem; Prediction; Principal component analysis; Regularization

1. Introduction

Contemporary data sets often consist of data units that are complex objects, such as functions, curves or images; see, for example, Ramsay and Silverman (2005), Ferraty and Vieu (2006), Ferraty and Romain (2011) and Horváth and Kokoszka (2012). It is standard in the field of functional data analysis to assume that all functions are observed on the same domain. In this paper, we develop methods of analysis for functional data that are observed incompletely in the sense that each function might be observed only on a subset of the domain, whereas no information about the curve is available on the complement of this subset.

Our work is motivated by an ambulatory blood pressure monitoring data set that is part of the 'Swiss kidney project on genes in hypertension' (Pruijm *et al.*, 2013) which is a multicentre cross-sectional study focusing on the role of kidney function and genes in blood pressure regulation and hypertension. In ambulatory blood pressure monitoring, participants wear a calibrated automatic device that is programmed to record systolic and diastolic blood pressure and heart rate at frequent intervals during 24 h (every 15 min during the day and every 30 min during the night). Ideally, this design should provide enough information for each continuous temporal profile to be reconstructed by standard smoothing techniques; the resulting sample of curves would then be analysed by traditional methods of functional data analysis. In reality,

Address for correspondence: David Kraus, Institute of Social and Preventive Medicine, University Hospital Lausanne, Route de la Corniche 10, Lausanne 1010, Switzerland.
E-mail: kraus.stat@gmail.com

 1369–7412/15/77777

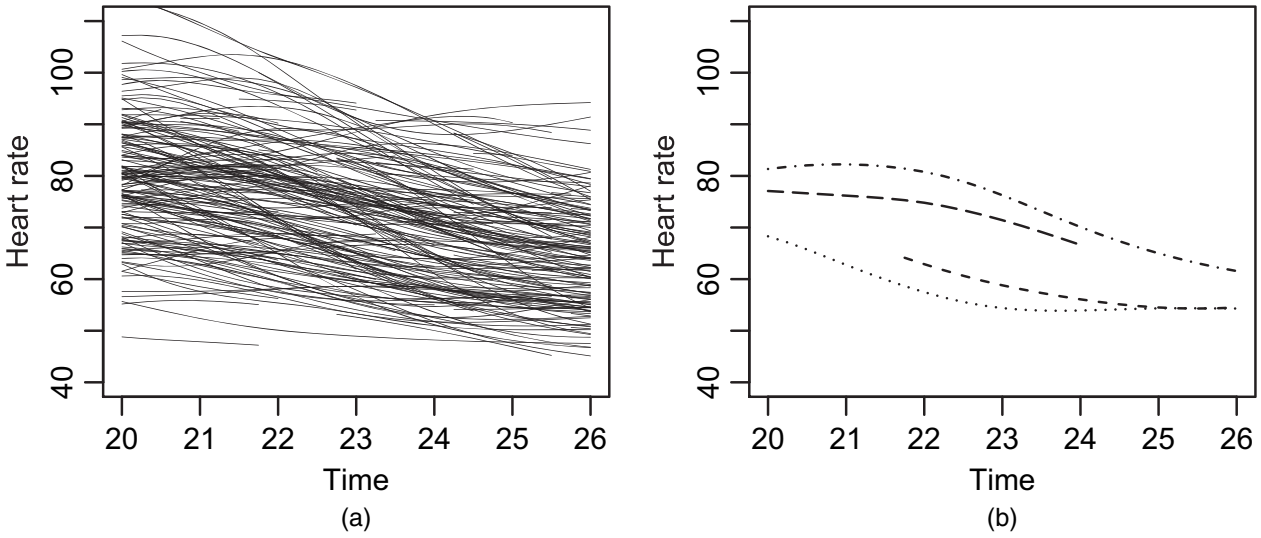

Fig. 1. (a) Subset of the sample of heart rate profiles and (b) several curves in detail

however, some values have not been measured and the time points corresponding to unobserved values form series (intervals) of non-negligible length. There are two main reasons why no measurements are available for certain periods: first is the participant's discomfort (the participants can remove the device when they feel uncomfortable) and second is the failure of the device to take measurements. However, there are series of frequent, properly recorded measurements. It is therefore possible to reconstruct the underlying profiles in continuous time on these periods. Fig. 1(a) displays a subset of 685 heart rate profiles (values in beats per minute); we focus on the time interval [20, 26] (i.e. from 8 p.m. of one day to 2 a.m. next day) that is of particular medical interest because it is the transition period between the day and night regime. In Fig. 1(b), we plot separately four profiles to illustrate the type of available data: whereas some curves (dotted and chain curves) are observed completely (on the entire domain [20, 26]), other curves (the two broken curves) have unobserved periods. The percentage of incomplete functions is 31% for blood pressure profiles and 44% for heart rate profiles. This is a considerable fraction of the data, and we therefore wish to avoid removing the incomplete curves from the analysis.

The partial observation regime that we encounter in this data set is of general interest in applications as often, despite the failure to observe the curves in some regions, there is enough observed information in the rest of the domain. The mechanism that causes the absence of data can be random, like in our data, but the curves may also be partially observed by design. Moreover, data need not necessarily be curves indexed by time; methods that we develop can be extended to more general object data subject to incomplete observation, such as partially observed images, spatial curves or surfaces. Hence this kind of functional data is worth systematic investigation. Interestingly and surprisingly, this observation pattern, however natural and likely to occur in many applications it is, has received relatively little attention in the literature. James *et al*. (2000) and James and Hastie (2001) used parametric mixed effects models for principal components analysis and classification of partially observed curves. Bugni (2012) developed a goodness-of-fit test under circumstances that were similar to those of our paper. Delaigle and Hall (2013) dealt with classification of functional data when only fragments of curves are available. Liebl (2013) studied low rank extensions of curves observed on subdomains. Goldberg *et al*. (2014) propose a prediction procedure for the continuation of a low rank functional observation.

In this paper we introduce a formal framework for analysing incompletely observed functional data and develop basic non-parametric, fully functional (infinite dimensional) inferential

procedures. When exploring functional data, one often finds interesting information in their covariance structure; see Ramsay and Silverman (2005) for some examples and, for example, Benko *et al.* (2009), Sangalli *et al.* (2009) or Panaretos *et al.* (2010) for other illustrations. Therefore, we first focus on the main building blocks of the analysis of the second-order properties: estimation of the covariance operator and principal component analysis. We propose an estimator of the covariance operator and its eigenvalues and eigenfunctions for partially observed functions and derive their properties. We deal with the estimation of projections (principal scores) of individual incomplete functions which is especially challenging. We develop a procedure that enables us to predict the value of a principal score of a function when only a fragment of the function is available and direct computation is thus impossible. Next, we propose a method that can recover the unobserved part of the function from the observed part, using the information about the distribution of the data that it learns from the sample. We develop automatic procedures for the selection of the tuning parameter of the method that is based on generalized cross-validation for incompletely observed functions. We quantify the uncertainty of the predictions of unobserved quantities and provide approximate prediction regions (intervals and bands) covering the unobserved random quantity with high probability. Simulations confirm the usefulness and good performance of the methodology proposed.

Both the prediction of principal scores and the reconstruction of an incomplete function or its derivatives are important problems. Principal scores are key elements in the exploration of complex data and can be used as input quantities in many inferential procedures. Their usefulness in the multivariate setting is well described, for example, in Krzanowski (2000) and Jolliffe (2002). In the functional context Ramsay and Silverman (2005) provided some real data examples illustrating how principal scores help to understand the properties of the data. Further applications can be found in Ramsay and Silverman (2002) and Ramsay *et al.* (2009). Horváth and Kokoszka (2012) have given a comprehensive account of the utility of principal scores in procedures like two-sample tests, linear and non-linear regression, clustering and classification, time series analysis or change point analysis. In this paper, we shall see in Section 6 that the first three principal components of the heart rate profiles and their derivatives explain a large proportion of the total variability and are sufficiently flexible to describe interesting features of the curves. Hence the corresponding scores provide an effectively reduced representation of the complex individual heart rate profiles. To perform graphical or formal analyses of the scores, we need to be able to compute them, which is not straightforward in the partial observation regime. Also, when an individual curve, surface or image is observed incompletely, one is interested in visualizing and studying the shape of the missing part, for instance to forecast the continuation of the natural or social process that is described by the functional variable. Our paper provides solutions to these problems by developing methods that predict unobserved quantities via their conditional expectation given the observed data. In addition to their direct application to data, these methods will be an important tool in future research: for instance, advanced techniques of missing data analysis in the multivariate setting involve conditional expectations in some form, and our results will be helpful in extending them to the functional case.

To our knowledge, no results of the kind that we provide here exist for functional data that are fully (densely in practice) observed on subsets of the domain. A related but different (in terms of applicability, used methods and achievable results) type of imperfectly observed functional data was studied by Yao *et al.* (2005a) who considered sparsely observed functions, i.e. situations where only a few observed values are available for each function, making it impossible to reconstruct each curve from these values. Our approach is novel in that it enables us, under the assumed observation regime, to investigate some genuinely functional aspects of the data. From the theoretical point of view, exploiting the continuous time nature of the observed data, we can

obtain stronger results than in the sparse regime. For example, the rates of convergence of estimators of parameters (the covariance operator and eigenelements) are parametric, unlike with sparsely observed data (see also Hall *et al*. (2006)). Also, the consistency result for our functional completion procedure is fully functional, whereas the restrictions of the sparse regime enabled Yao *et al*. (2005a) to achieve pointwise or finite dimensional convergence of the reconstructed trajectory. From the practical perspective, an important advantage of our method is that derivatives can be readily analysed in our setting whereas with methods for sparsely observed functions it is complicated. The method of Liu and Müller (2009) is a variant of that of Yao *et al*. (2005a) that can deal with derivatives in the sparse regime to some extent. Although the method of Liu and Müller (2009) can reconstruct derivatives, it does not provide insight into their covariance structure because it neither estimates the covariance operator of the derivatives nor performs principal component analysis of the derivatives (it is based on derivatives of eigenfunctions rather than on eigenfunctions of derivatives). Since derivatives describe the dynamics of the underlying real world process, the analysis of derivatives, and especially of the principal sources of their variability, is often revealing in many applications, including the one we consider in this paper.

Mathematically, the problem that we need to solve for the computation of unobserved quantities (prediction of principal scores or reconstruction of missing parts of trajectories) is seen to be an ill-posed inverse problem (e.g. Groetsch (1993)), and regularization techniques need to be applied. Such problems previously appeared in the literature on complete functional data mainly in the area of functional regression modelling; see, for example, Cardot *et al*. (1999, 2007), Müller and Stadtmüller (2005), Cai and Hall (2006), Hall and Horowitz (2007) or He *et al*. (2010). Inverse problems similar to those which we encounter here also arise in connection with functional canonical correlations (e.g. He *et al.* (2003)) or with tests of hypotheses on parameters of functional data (e.g. Mas (2007), Horváth *et al.* (2010, 2013), Aston and Kirch (2012), Kraus and Panaretos (2012) and Jarušková (2013)). Our problem is related to the task of prediction that was previously studied in the literature on functional time series; see, for example, Bosq (2000), Antoniadis and Sapatinas (2003) or Kargin and Onatski (2008). None of these references, however, assumes the partial observation pattern that we consider in this paper.

The paper is organized as follows. In Section 2 we formalize the mechanism of partial observation of functional data and deal with the estimation of the mean function and covariance operator. Section 3 develops principal component analysis for incompletely observed functions. In Section 4, a method is proposed to reconstruct the missing part of a partially observed curve. Sections 5 and 6 present a simulation study and a data example. Appendix A contains proofs of the main theoretical results (theorems 1 and 2). A supplementary document available on line contains proofs of propositions 1–4 and a detailed description of computational procedures.

The programs that were used to analyse the data and some example data can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Partially observed functional data

Functional data $X_1, \ldots, X_n$ are seen as independent identically distributed random variables in the separable Hilbert space of square integrable functions on a bounded domain. Without loss of generality, we consider the space $L^2([0,1])$ with inner product $\langle f, g\rangle = \int_0^1 f(t)\, g(t)\, \mathrm{d}t$, $f, g \in L^2([0,1])$ and norm $\|f\| = \langle f, f\rangle^{1/2}$. It is possible to extend our results to vector-valued functions or more general domains for applications with spatial curves, surfaces, images etc.

In traditional functional data analysis, it is assumed that the functions $X_1, \ldots, X_n$ are observed on the whole interval $[0, 1]$. We consider situations where each curve X_i is observed only on a subset of $[0, 1]$. Specifically, let the observation periods be $O_i \subset [0, 1]$, $i = 1, \ldots, n$. Then the observed data for the ith curve are $X_i(t)$, $t \in O_i$. (In practice, the raw data are most often in the form of possibly noisy observations on a dense grid of points in O_i, which enables us to assume that the curves are observed fully in O_i, as is explained by Hall *et al.* (2006).) We collectively denote the observed part of the curve as X_{iO_i}, which can be seen as a random element of the space $L^2(O_i)$. The values of X_i on the complement of O_i, $M_i = [0, 1] \setminus O_i$, are not observed; the missing part of the trajectory is denoted as X_{iM_i}. The observation periods O_i, $i = 1, \ldots, n$, are modelled as random subsets of $[0, 1]$. We assume that each realization of O_i is the union of a finite number of intervals. This assumption is not restrictive for practical applications, although some generalizations are probably possible. We assume that the observation periods are independent of the functions $X_1, \ldots, X_n$, i.e. the data are missing completely at random. (Under this assumption, the observation periods can also be seen as fixed when inference is made about the curves.)

The main characteristics of the distribution that generates the data are the mean function and the covariance operator. Let the mean function be $\mu = E(X_1)$. The covariance operator $\mathscr{R} : L^2([0, 1]) \to L^2([0, 1])$ is defined as

$$\mathscr{R} f = E\{\langle f, X_1 - \mu \rangle (X_1 - \mu)\} = \int_0^1 \rho(\cdot, t)\, f(t)\, \mathrm{d}t,$$

where $\rho(s, t) = \mathrm{cov}\{X_1(s), X_1(t)\}$ is the covariance kernel of the stochastic process X_1.

Like in the multivariate case, the mean function μ at point $t \in [0, 1]$ can be estimated by the sample mean of observed values at this point. Formally, the estimator can be written as

$$\hat{\mu}(t) = \frac{J(t)}{\sum_{i=1}^n O_i(t)} \sum_{i=1}^n O_i(t)\, X_i(t),$$

where the notation $O_i(t)$ is used for the indicator $\mathbf{1}_{O_i}(t)$ and $J(t) = \mathbf{1}_{[\sum_{i=1}^n O_i(t) > 0]}$. The values of $X_i(t)$ are available only if $O_i(t) = 1$; otherwise, the contribution $O_i(t) X_i(t)$ in the sum above is zero. The term $J(t)$ is included to avoid division by 0: if $J(t) = 0$, the estimate of the mean is 0 (or arbitrary, as such situations vanish asymptotically).

The estimator $\hat{\mathscr{R}}$ of the covariance operator $\mathscr{R}$ is defined through an estimator of its covariance kernel ρ. We estimate $\rho(s, t)$ by the sample covariance computed from all complete pairs of functional values at s and t. The estimator equals

$$\hat{\rho}(s, t) = \frac{I(s, t)}{\sum_{i=1}^n U_i(s, t)} \sum_{i=1}^n U_i(s, t)\{X_i(s) - \hat{\mu}_{st}(s)\}\{X_i(t) - \hat{\mu}_{st}(t)\}, \tag{1}$$

where $U_i(s, t) = O_i(s)\, O_i(t)$ and $I(s, t) = \mathbf{1}_{[\sum_{i=1}^n U_i(s,t) > 0]}$. The estimator of the mean function used here is

$$\hat{\mu}_{st}(s) = \frac{I(s, t)}{\sum_{i=1}^n U_i(s, t)} \sum_{i=1}^n U_i(s, t)\, X_i(s),$$

i.e., for the computation of the covariance at s, t, functional values are centred at the sample mean computed from complete pairs. (It is also possible to centre by the estimator $\hat{\mu}$ that was introduced before; all results remain valid when $\hat{\mu}$ is used in place of $\hat{\mu}_{st}$.)

The sample covariance operator computed from incomplete functions may be indefinite. This is similar to the multivariate setting. However, unlike with multivariate data, our experience in the functional context is that this problem is unimportant in practice because negative eigenvalues occur far in the tail of the spectrum and are small in comparison with the leading eigenvalues. The corresponding high frequency features of the data are practically never of interest. If needed, the estimate $\hat{\mathscr{R}}$ can be modified by setting negative eigenvalues equal to 0.

It is seen that $\hat{\mu}(t)$ is an unbiased estimator of $\mu(t)$. Similarly, if we subtract 1 in the denominator of $\hat{\rho}(s,t)$, the estimator becomes unbiased for $\rho(s,t)$. For the estimators $\hat{\mu}$ and $\hat{\mathscr{R}}$ to be consistent, we need to assume that the observation pattern asymptotically provides enough information. For the mean function, the right assumption is that

$$\text{there exists } \delta_1 > 0 \text{ such that } \sup_{t\in[0,1]} P\left\{ n^{-1} \sum_{i=1}^{n} O_i(t) \leqslant \delta_1 \right\} = O(n^{-2}) \text{ as } n \to \infty. \tag{2}$$

Similarly, for the covariance operator, we need the stronger assumption that

$$\text{there exists } \delta_2 > 0 \text{ such that } \sup_{(s,t)\in[0,1]^2} P\left\{ n^{-1} \sum_{i=1}^{n} U_i(s,t) \leqslant \delta_2 \right\} = O(n^{-2}) \text{ as } n \to \infty. \tag{3}$$

Assumption (2) is satisfied, for example, when the observation sets $O_1, \ldots, O_n$ are independent and identically distributed and $\pi_0 = \inf_{t\in[0,1]} P\{O_1(t) = 1\} > 0$. To see this, set $\delta_1 = \pi_0/2$ and use Hoeffding's inequality to show that

$$\sup_{t\in[0,1]} P\left\{ n^{-1} \sum_{i=1}^{n} O_i(t) \leqslant \delta_1 \right\} \leqslant \exp(-\pi_0^2 n/2).$$

Analogously, assumption (3) is satisfied when we further assume that $\inf_{(s,t)\in[0,1]^2} P\{U_1(s,t) = 1\} > 0$. Under these weak assumptions, we obtain a consistency result as follows.

Proposition 1.

(a) Let $E(\|X_1\|^2) < \infty$ and assumption (2) be satisfied. Then $E(\|\hat{\mu} - \mu\|^2) = O(n^{-1})$ for $n \to \infty$.
(b) Let $E(\|X_1\|^4) < \infty$ and assumption (3) be satisfied. Then $E(\|\hat{\mathscr{R}} - \mathscr{R}\|_2^2) = O(n^{-1})$ for $n \to \infty$ (here $\|\cdot\|_2$ denotes the Hilbert–Schmidt norm).

Note that the properties of the estimators are unaffected by the fact that the functions are observed only partially. The full (dense) observation regime, albeit only on subsets of the domain, preserves the convergence rates that are known for complete functional data (see Bosq (2000) or Horváth and Kokoszka (2012) for results in the traditional setting).

3. Principal component analysis

3.1. Estimation of eigenfunctions and eigenvalues

Probably the most fundamental method for functional data is functional principal component analysis. It provides insight into the complex covariance structure of functional data and is used to identify main sources of variability and to quantify their importance and to reduce the dimension of the data.

The theoretical foundation of functional principal component analysis is the Karhunen–Loève theorem (e.g. Bosq (2000), theorem 1.5) stating that there are random variables β_{ij} and non-random functions φ_j such that the stochastic process X_i admits the decomposition

$$X_i(t) = \mu(t) + \sum_{j=1}^{\infty} \beta_{ij}\varphi_j(t), \qquad t \in [0,1],$$

where the series converges in mean square, uniformly in t. Here φ_j, $j = 1, 2, \ldots$, are the orthonormal eigenfunctions of the operator $\mathscr{R}$ and β_{ij}, $j = 1, 2, \ldots$, are uncorrelated mean 0 variables with variances λ_j, where $\lambda_1 \geqslant \lambda_2 \geqslant \ldots > 0$ are the eigenvalues of $\mathscr{R}$. Functional principal component analysis is the empirical version of the Karhunen–Loève expansion that aims to estimate the elements involved in the expansion. For background information on this classical topic, we refer to Ramsay and Silverman (2005), chapter 8, for an introduction from an applied perspective, and to Dauxois *et al.* (1982), Bosq (2000) or Hall and Hosseini-Nasab (2006) for theoretical studies.

In the case of completely observed functional data, to estimate the eigenvalues λ_j and eigenfunctions φ_j, one performs eigendecomposition of the usual sample covariance operator. When the functions are observed partially, we can proceed similarly and define the estimators $\hat{\lambda}_j$ and $\hat{\varphi}_j$ as the eigenvalues and eigenfunctions of the operator $\hat{\mathscr{R}}$ given by the kernel $\hat{\rho}$ in equation (1).

It turns out that the asymptotic properties of the empirical eigenvalues and eigenfunctions remain unchanged by the incompleteness of the observed functions. The following proposition shows that, first, the empirical eigenvalues are consistent estimators of the true eigenvalues and this consistency is uniform over all indices and, second, the empirical eigenfunctions are consistent estimators of the true eigenfunctions, up to the usual sign ambiguity.

Proposition 2. Let $E(\|X_1\|^4) < \infty$ and assumption (3) be satisfied. Then $E[\sup_{j\in\mathbb{N}}\{|\hat{\lambda}_j - \lambda_j|^2\}] = O(n^{-1})$. If moreover all eigenvalues of $\mathscr{R}$ have multiplicity 1, then $E(\|\hat{\varphi}_j - \hat{s}_j\varphi_j\|^2) = O(n^{-1})$ for all $j \in \mathbb{N}$, where $\hat{s}_j = \operatorname{sgn}\langle\hat{\varphi}_j, \varphi_j\rangle$.

The rates of convergence are parametric because of the full observation regime on subsets; the situation is different from that of sparsely observed functions, where the estimators of eigenelements (constructed differently) converge at non-parametric rates (Yao *et al.*, 2005a; Hall *et al.*, 2006).

3.2. Estimation of principal component scores

In principal component analysis, one is usually interested not only in estimating the eigenfunctions and eigenvalues but also in the estimation of the principal component scores

$$\beta_{ij} = \langle X_i - \mu, \varphi_j\rangle, \qquad i = 1, \ldots, n, \quad j = 1, 2, \ldots,$$

representing the individual co-ordinates of each curve with respect to the eigenbasis (the expression of the feature φ_j for the ith observation). The leading principal scores provide the optimal finite dimensional representation of each curve and can be further analysed by traditional techniques.

In the standard situation of complete functional data, the scores are easily estimated by $\hat{\beta}_{ij} = \langle X_i - \hat{\mu}, \hat{\varphi}_j\rangle$. When the functional observations are incomplete, the direct computation of $\langle X_i - \hat{\mu}, \hat{\varphi}_j\rangle$ is impossible because the last term in the expression

$$\langle X_i - \hat{\mu}, \hat{\varphi}_j\rangle = \langle X_i O_i - \hat{\mu}_{O_i}, \hat{\varphi}_{jO_i}\rangle + \langle X_{iM_i} - \hat{\mu}_{M_i}, \hat{\varphi}_{jM_i}\rangle$$

is not available. In this equation the subscript O_i or M_i denotes the restriction of the corresponding function to the ith observed or missing period respectively. We develop a procedure to estimate the missing quantity $\langle X_{iM_i} - \hat{\mu}_{M_i}, \hat{\varphi}_{jM_i}\rangle$ from the observed data.

First, we consider the population version of the problem. Let the function X with mean 0 and covariance operator $\mathscr{R}$ be observed on the set O and missing on M. For the following considerations, the sets O and M, which are independent of X, can be regarded as non-random; equivalently, derivations can be made conditionally on them. The goal is to predict $\beta_{jM} = \langle X_M, \varphi_{jM}\rangle$ from the observed part X_O. It is a standard fact that, in terms of the mean-squared prediction error, the best approximation of β_{jM} by a functional of X_O is the conditional expectation $E(\beta_{jM}|X_O)$. The conditional expectation may be a non-linear functional of the condition and thus difficult to estimate. Therefore, we propose to look for the best linear prediction corresponding to a continuous linear functional of the observed curve. This is equivalent to the best linear approximation of the conditional expectation. By the Riesz representation theorem, a continuous linear functional takes the form $\langle a_j, X_O\rangle$, where a_j is an element of $L^2(O)$. The best continuous linear prediction of β_{jM} equals $\tilde{\beta}_{jM} = \langle \tilde{a}_j, X_O\rangle$, where $\tilde{a}_j$ solves the infinite dimensional optimization problem

$$\min_{a_j \in L^2(O)} E\{(\beta_{jM} - \langle a_j, X_O\rangle)^2\}. \tag{4}$$

The objective functional can be rewritten as

$$\begin{aligned} E\{(\beta_{jM} - \langle a_j, X_O\rangle)^2\} &= E\{\langle \varphi_{jM}, X_M\rangle^2 - 2\langle \varphi_{jM}, X_M\rangle\langle a_j, X_O\rangle + \langle a_j, X_O\rangle^2\} \\ &= \langle \varphi_{jM}, \mathscr{R}_{MM}\varphi_{jM}\rangle - 2\langle \varphi_{jM}, \mathscr{R}_{MO}a_j\rangle + \langle a_j, \mathscr{R}_{OO}a_j\rangle, \end{aligned}$$

where $\mathscr{R}_{OO}$ is the covariance operator of X_O and $\mathscr{R}_{MO}$ is the cross-covariance operator of X_M and X_O. It is obvious that the objective functional is convex. If a minimizer exists, it can be found by setting the derivative equal to 0. The derivatives in this context are in the Fréchet sense. In particular, we see that

$$\frac{\partial}{\partial a_j} E\{(\beta_{jM} - \langle a_j, X_O\rangle)^2\} = -2r_j + 2\mathscr{R}_{OO}a_j,$$

where $r_j = \mathscr{R}_{OM}\varphi_{jM}$ with $\mathscr{R}_{OM} = \mathscr{R}_{MO}^*$ (the asterisk denotes the adjoint operator). Thus we need to solve the equation

$$\mathscr{R}_{OO}a_j = r_j. \tag{5}$$

We recognize that this is a linear inverse problem where we need to recover the function $a_j \in L^2(O)$ from its image through the linear operator $\mathscr{R}_{OO}$.

Let λ_{OOk}, $k = 1, 2, \ldots,$ be the decreasing positive eigenvalues and φ_{OOk} the corresponding orthonormal eigenfunctions of the operator $\mathscr{R}_{OO}$. By comparing the coefficients of the left- and right-hand side of equation (5) with respect to the basis φ_{OOk}, we arrive at the system of equations $\lambda_{OOk}\langle a_j, \varphi_{OOk}\rangle = \langle r_j, \varphi_{OOk}\rangle, k = 1, 2, \ldots$. This suggests that a candidate for the solution is

$$\tilde{a}_j = \sum_{k=1}^{\infty} \frac{\langle r_j, \varphi_{OOk}\rangle}{\lambda_{OOk}} \varphi_{OOk}, \tag{6}$$

i.e. $\tilde{a}_j = \mathscr{R}_{OO}^{-1} r_j$. This is a valid solution, if it is an element of $L^2(O)$, i.e. if

$$\sum_{k=1}^{\infty} \frac{\langle r_j, \varphi_{OOk}\rangle^2}{\lambda_{OOk}^2} < \infty. \tag{7}$$

This condition is known in the theory of inverse problems as Picard's condition. A solution to the inverse problem (5) exists if and only if condition (7) is satisfied.

Condition (7) is equivalent to the condition

$$\sum_{k=1}^{\infty} \frac{\operatorname{corr}(\beta_{jM}, \langle X_O, \varphi_{OOk}\rangle)^2}{\operatorname{var}(\langle X_O, \varphi_{OOk}\rangle)} < \infty, \tag{8}$$

which has a clear interpretation. It states that the missing variable β_{jM} must not be strongly correlated with complicated, high frequency components of the observed function. The variability of these components must be sufficiently large to provide enough information for the prediction of β_{jM}. The precise balance between the complexity of the correlation of the unobserved score with the predictor components and the variability of the predictor components is quantified by the requirement on the series above to converge.

In the Gaussian case, the conditional expectation of β_{jM} given the principal scores $\langle X_O, \varphi_{OOk}\rangle, k = 1, 2, \ldots,$ is an infinite linear combination of these scores (an almost surely convergent infinite series). One can show this by conditioning on finitely many components (this multivariate conditional expectation is linear) and applying Lévy's 0–1 law (Kallenberg (2002), theorem 7.23) to obtain the limit. The infinite sum of variances of terms in this series converges, which is equivalent to the convergence of $\Sigma_{k=1}^{\infty} \langle r_j, \varphi_{OOk}\rangle^2 / \lambda_{OOk}$ or $\Sigma_{k=1}^{\infty} \operatorname{corr}(\beta_{jM}, \langle X_O, \varphi_{OOk}\rangle)^2$. If, moreover, condition (7) or (8) is satisfied, then the coefficients in the infinite linear combination for the conditional expectation form an l^2-sequence; hence the conditional expectation is continuous in the condition.

From now on, to guarantee the existence of a continuous solution to condition (5), we assume that condition (7) holds. If it is *a priori* known that the conditional expectation $E(\beta_{jM}|X_O)$ is a continuous linear functional of X_O, then condition (7) is automatically satisfied.

The operator $\mathscr{R}_{OO}$ is a compact operator with infinite dimensional range; therefore, its inverse $\mathscr{R}_{OO}^{-1}$ is not bounded (i.e. not continuous). Consequently, small perturbations of r_j may lead to large perturbations of $\tilde{a}_j = \mathscr{R}_{OO}^{-1} r_j$. It is seen from equation (6) that an overall small change of r_j may result in an arbitrarily large change of $\tilde{a}_j$, if the change of r_j occurs on a coefficient with a sufficiently high index k; the division by a sufficiently small eigenvalue may enormously magnify the perturbation. In other words, the solution $\tilde{a}_j = \mathscr{R}_{OO}^{-1} r_j$ is extremely unstable and the inverse problem (5) is ill posed. It is important for a solution to be stable with respect to perturbations of the right-hand side r_j because r_j is unknown and needs to be estimated. With estimated right-hand side, the solution to the inverse problem may be arbitrarily far from the true solution no matter how accurate the estimate is. This is true even when $\mathscr{R}_{OO}$ is known. Moreover, the operator $\mathscr{R}_{OO}$ is not known either; its estimate has finite rank and therefore is not invertible in $L^2(O)$.

To obtain a stable solution, one needs to use regularization, i.e. to modify the ill-posed inverse problem in such a way that it becomes well posed with a stable solution. We use ridge regularization. Instead of problem (5), we solve the problem $\mathscr{R}_{OO}^{(\alpha)} a_j = r_j$ with $\mathscr{R}_{OO}^{(\alpha)} = \mathscr{R}_{OO} + \alpha \mathscr{I}_O$, where $\alpha > 0$ and $\mathscr{I}_O$ is the identity operator on $L^2(O)$. The inverse $\mathscr{R}_{OO}^{(\alpha)-1}$ of the bounded operator $\mathscr{R}_{OO}^{(\alpha)}$ is bounded and therefore the solution $\tilde{a}_j^{(\alpha)} = \mathscr{R}_{OO}^{(\alpha)-1} r_j$ is stable. Denote the regularized best linear prediction of β_{jM} by $\tilde{\beta}_{jM}^{(\alpha)} = \langle \tilde{a}_j^{(\alpha)}, X_O\rangle$. The stability of the solution increases with α but the bias of the solution increases also because the problem becomes more different from the original problem; conversely, with α decreasing, the solution becomes closer to the exact but unstable solution of the original problem.

We now turn to the practical, empirical version of the problem of computation of principal scores from partially observed functional data. We have a sample of n functions $X_{1O_1}, \ldots, X_{nO_n}$ observed on the sets $O_1, \ldots, O_n$. The mean function μ and the covariance operator $\mathscr{R}$ are

estimated by $\hat{\mu}$ and $\hat{\mathscr{R}}$ introduced in Section 2. The principal score of the ith curve with respect to the jth eigenfunction is estimated by $\hat{\beta}_{ij}^{(\alpha)} = \hat{\beta}_{ijO_i} + \hat{\beta}_{ijM_i}^{(\alpha)}$, where $\hat{\beta}_{ijO_i} = \langle X_{iO_i} - \hat{\mu}_{O_i}, \hat{\varphi}_{jO_i} \rangle$ and $\hat{\beta}_{ijM_i}^{(\alpha)} = \langle \hat{a}_{ij}^{(\alpha)}, X_{iO_i} - \hat{\mu}_{O_i} \rangle$. Here the function $\hat{a}_{ij}^{(\alpha)} = \hat{\mathscr{R}}_{O_iO_i}^{(\alpha)-1} \hat{r}_{ij}$ solves the empirical regularized inverse problem $\hat{\mathscr{R}}_{O_iO_i}^{(\alpha)} a_{ij} = \hat{r}_{ij}$, where $\hat{\mathscr{R}}_{O_iO_i}^{(\alpha)} = \hat{\mathscr{R}}_{O_iO_i} + \alpha \mathscr{I}_{O_i}$ with $\hat{\mathscr{R}}_{O_iO_i}$ being an integral operator on $L^2(O_i)$ with kernel equal to the restriction of the kernel $\hat{\rho}$ of $\hat{\mathscr{R}}$ (see equation (1)) to $O_i \times O_i$, and $\hat{r}_{ij} = \hat{\mathscr{R}}_{O_iM_i} \hat{\varphi}_{jM_i}$ with $\hat{\mathscr{R}}_{O_iM_i}$ defined analogously by restriction of $\hat{\rho}$ to $O_i \times M_i$.

We are ready to state the main convergence result that justifies this method. The difference between the regularized estimator $\hat{\beta}_{ijM_i}^{(\alpha)}$ and the best linear prediction $\tilde{\beta}_{ijM_i}$ can be decomposed into the sum of the estimation error for the regularized prediction and the approximation error due to regularization, i.e. $\hat{\beta}_{ijM_i}^{(\alpha)} - \tilde{\beta}_{ijM_i} = \hat{\beta}_{ijM_i}^{(\alpha)} - \tilde{\beta}_{ijM_i}^{(\alpha)} + \tilde{\beta}_{ijM_i}^{(\alpha)} - \tilde{\beta}_{ijM_i}$. We show that, when the amount of regularization decreases at a suitable rate as the sample size increases, both terms converge to 0 in $L^2(P)$ and thus the regularized estimator of the prediction is consistent.

Theorem 1. Let $E(\|X_1\|^4) < \infty$, assumption (3) be satisfied, all eigenvalues of $\mathscr{R}$ have multiplicity 1 and condition (7) be satisfied for O_i and M_i in place of O and M respectively. Then

$$E\{(\hat{\beta}_{ijM_i}^{(\alpha)} - \tilde{\beta}_{ijM_i})^2\} \leqslant O(\alpha^{-3})\, O(n^{-1}) + O(\alpha)$$

as $\alpha \to 0$ and $n \to \infty$. Hence, if $\alpha = \alpha_n$ such that $\alpha_n \to 0$ and $\alpha_n n^{1/3} \to \infty$ as $n \to \infty$, then $\hat{\beta}_{ijM_i}^{(\alpha_n)}$ is a consistent estimator of the best linear prediction $\tilde{\beta}_{ijM_i}$ of β_{ijM_i}.

Sometimes one is interested in estimating other linear functionals than the principal score $\langle X_i - \mu, \varphi_j \rangle$. Our consistency results remain valid when $\hat{\varphi}_{jO_i}$ is replaced by an arbitrary random or fixed function $\hat{f}_{O_i} \in L^2(O_i)$ such that $E(\|\hat{f}_{O_i} - f_{O_i}\|^2) = O(n^{-1})$ for some deterministic $f_{O_i} \in L^2(O_i)$.

Note that theorem 1 has no strong assumptions. Picard's condition (7) is a basic assumption that is required in all inverse problems to guarantee the existence of a solution. Except this standard requirement, no other condition on the rate of decrease of the eigenvalues $\lambda_{O_iO_ik}$ is needed. This is because we estimate the prediction $\langle \tilde{a}_{ij}, X_{iO_i} \rangle$ rather than the prediction functional $\tilde{a}_{ij}$ itself. Intuitively, the integration in $\langle \tilde{a}_{ij}, X_{iO_i} \rangle$ brings additional smoothness; the exact way that this happens is seen in the proof of theorem 1. In a related context of prediction in functional linear regression, it was observed by Cai and Hall (2006) and Cardot *et al.* (2007) that weaker assumptions are needed and stronger results can be obtained when the focus is on prediction rather than on the estimation of the regression functional. The inverse problem is similar to that solved in the functional linear model (Cardot *et al.*, 1999, 2007; Hall and Horowitz, 2007). However, the way that we arrive at it differs from the functional linear model because, for instance, of the incompleteness of observations there is no collection of response–covariate pairs in the present situation.

As an alternative to ridge regularization, one may consider the spectral truncation approach. Both methods have their advantages and disadvantages. For instance, it is known that the behaviour of spectral cut-off methods depends on the spacings between the eigenvalues of the operator to be inverted which makes them less robust with respect to situations with similar or even identical eigenvalues (see Hall and Horowitz (2007)). Indeed, in a preliminary analysis of our motivating data set we observed some very similar estimated eigenvalues. There is also an important computational advantage of the ridge method. For this method, one needs to solve only a linear equation with $\hat{\mathscr{R}}_{O_iO_i}^{(\alpha)}$ which is very easy and fast. In contrast, the spectral truncation approach requires computing the eigendecomposition of $\hat{\mathscr{R}}_{O_iO_i}$ and projecting on the corresponding subspace. This is computationally more demanding, especially since it must be done repeatedly for each function because different suboperators $\hat{\mathscr{R}}_{O_iO_i}$ of $\hat{\mathscr{R}}$ corresponding to different functions

have different spectral decompositions. Yet another approach may be based on smoothing, for instance, by penalizing the roughness of the solution of the inverse problem.

3.3. Regularization parameter selection

Theorem 1 shows that, for an appropriate choice of α_n, the estimator $\hat{\beta}_{ijM_i}^{(\alpha_n)}$ is consistent for the best prediction $\tilde{\beta}_{ijM_i}$. Theorem 1, however, does not give a practical recommendation on how to select the regularization parameter. It is desirable to have an automatic, data-driven selection procedure.

Since the parameter α is difficult to understand, we first translate it into more comprehensible values. By analogy with ridge regression or various standard smoothing techniques, we define the number of effective degrees of freedom as the trace of the covariance of the predictors composed of its regularized inverse, i.e.

$$\mathrm{df}_i(\alpha) = \mathrm{tr}(\hat{\mathscr{R}}_{O_iO_i}^{(\alpha)-1}\hat{\mathscr{R}}_{O_iO_i}) = \sum_{k=1}^{\infty} \frac{\hat{\lambda}_{O_iO_ik}}{\hat{\lambda}_{O_iO_ik}+\alpha}, \tag{9}$$

which is a decreasing function of α. Unlike in standard situations the covariance operator here is computed from partially observed data. Another way to measure the amount of regularization is the proportion of retained variability like in classical principal component analysis using, for example,

$$\frac{\mathrm{tr}(\hat{\mathscr{R}}_{O_iO_i}\hat{\mathscr{R}}_{O_iO_i}^{(\alpha)-1}\hat{\mathscr{R}}_{O_iO_i}\hat{\mathscr{R}}_{O_iO_i}^{(\alpha)-1}\hat{\mathscr{R}}_{O_iO_i})}{\mathrm{tr}(\hat{\mathscr{R}}_{O_iO_i})} = \frac{\sum_{k=1}^{\infty}\hat{\lambda}_{O_iO_ik}^3/(\hat{\lambda}_{O_iO_ik}+\alpha)^2}{\sum_{k=1}^{\infty}\hat{\lambda}_{O_iO_ik}} \tag{10}$$

or a similar quantity. One can determine α such that the effective degrees of freedom equal some value or the proportion of retained variability exceeds some threshold. These quantities, however, do not measure the predictive performance of the regularized solution.

A universal recipe for situations of this type is to use generalized cross-validation. In traditional settings, the generalized cross-validation score is the residual sum of squares (a measure of goodness of fit) divided by a decreasing function of the effective degrees of freedom (a penalty included to avoid underregularization). The residual sum of squares is the sum of squared differences of the response variables and their predictions, which in our case are $\hat{\beta}_{kjM_i} = \langle X_{kM_i} - \hat{\mu}_{M_i}, \hat{\varphi}_{jM_i}\rangle$ and $\hat{\beta}_{kjM_i}^{(\alpha)} = \langle \hat{a}_{ij}^{(\alpha)}, X_{kO_i} - \hat{\mu}_{O_i}\rangle$, $k = 1, \ldots, n$, respectively. In the situation of partially observed functions, the pair of the response variable $\hat{\beta}_{kjM_i}$ and the explanatory variable X_{kO_i} is not available for all individuals $k = 1, \ldots, n$. The idea is, therefore, to consider the set of completely observed functions with indices $C = \{k : 1 \leqslant k \leqslant n, \int_0^1 O_k(t)\,\mathrm{d}t = 1\}$. If this set is reasonably large, we can compute the residual sum of squares over the complete functions

$$\mathrm{rss}_{ij}(\alpha) = \sum_{k\in C}(\hat{\beta}_{kjM_i} - \hat{\beta}_{kjM_i}^{(\alpha)})^2.$$

The cross-validation score for the regularized estimation of the jth score of the ith function is

$$\mathrm{gcv}_{ij}(\alpha) = \frac{\mathrm{rss}_{ij}(\alpha)}{\{1 - (1/|C|)\mathrm{df}_i(\alpha)\}^2},$$

where $|C|$ is the number of complete functions. One selects the value of α that minimizes this quantity. Separate values of the regularization parameter are used for each function and each score.

3.4. Prediction uncertainty

For a statistical procedure to be useful, it is important to quantify its uncertainty, i.e. to assess how far $\hat{\beta}_{ijM_i}^{(\alpha_n)}$ can be from β_{ijM_i}. The following proposition answers these questions.

Proposition 3. Let the assumptions of theorem 1 be satisfied and let $\alpha_n \to 0$ and $\alpha_n n^{1/4} \to \infty$ as $n \to \infty$. Then $\hat{\beta}_{ijM_i}^{(\alpha_n)} - \beta_{ijM_i}$ is asymptotically distributed as $\tilde{\beta}_{ijM_i} - \beta_{ijM_i}$, which is a zero-mean random variable with variance that can be consistently estimated by

$$\hat{v}_{ij}^2 = \langle \hat{\varphi}_{jM_i}, (\hat{\mathscr{R}}_{M_iM_i} - \hat{\mathscr{R}}_{M_iO_i} \hat{\mathscr{R}}_{O_iO_i}^{(\alpha_n)-1} \hat{\mathscr{R}}_{O_iO_i} \hat{\mathscr{R}}_{O_iO_i}^{(\alpha_n)-1} \hat{\mathscr{R}}_{O_iM_i}) \hat{\varphi}_{jM_i} \rangle.$$

If the distribution of the data is Gaussian, then the limiting variable is Gaussian.

The assumptions of this proposition are similar to those of the consistency result of theorem 1, except that a slower rate of convergence of the regularization parameter to 0 is needed to estimate the limiting variance consistently.

The prediction uncertainty, as expressed by the variance $\hat{v}_{ij}^2$, does not converge to 0 as the sample size converges to ∞. This is because the situation is a prediction problem rather than an estimation problem in the sense that we try to recover a random variable rather than a non-random parameter. Thus, although increasing the sample size eventually removes the uncertainty due to unknown estimated quantities (the mean function and covariance operator) and regularization, there is a fundamental uncertainty that cannot be removed asymptotically. In other words, the knowledge of the principal score will never be precise, if the functional observation is incomplete, and the limits of accuracy of the prediction are given by the asymptotic variance v_{ij}^2. We refer to Didericksen *et al.* (2012) for an interesting discussion of similar questions in somewhat related prediction problems in the context of functional time series.

Proposition 3 immediately enables us to construct a prediction interval for the score. Assume that a Gaussian distribution is a good approximation for the distribution of the data. Then

$$I_{ij;\eta} = (\hat{\beta}_{ij}^{(\alpha_n)} - z_{1-\eta/2} \hat{v}_{ij}, \hat{\beta}_{ij}^{(\alpha_n)} + z_{1-\eta/2} \hat{v}_{ij}), \tag{11}$$

where $z_{1-\eta/2}$ is the $(1-\eta/2)$-quantile of the standard normal distribution, is a prediction interval for β_{ij} with asymptotic coverage probability $1-\eta$, i.e. $P(\beta_{ij} \in I_{ij;\eta}) \to 1-\eta$ as $n \to \infty$.

Since principal component analysis is often used as a dimension reduction procedure and the resulting principal scores are subsequently analysed by traditional techniques, it is useful to have a measure of reliability of the computed scores. The true score β_{ij} is a random variable with variance estimated by $\hat{\lambda}_j$. The predicted score $\hat{\beta}_{ij}^{(\alpha_n)}$ can be seen as the true score contaminated by error with variance estimated by $\hat{v}_{ij}^2$. One can define the relative error

$$\hat{v}_{ij} / \hat{\lambda}_j^{1/2}, \tag{12}$$

which is the ratio of the error variability and the natural intrinsic variability of the score. This value, lying between 0 and 1, can be used as an indicator of observations that are too uncertain, and the scores whose relative error exceeds a certain threshold (e.g. 0.2) can be excluded from the subsequent analysis. The uncertainty will be high when the association between the missing part of the score and the observed fragment is weak.

The high uncertainty of predictions due to a small amount of observed information is one example of situations where we must be cautious. Another such case could be when missingness is very frequent in certain regions or the overlap of observation periods is not sufficiently frequent because then the precision of the estimation of the covariance function will be locally reduced, and consequently the prediction procedure may be less accurate. The performance

of generalized cross-validation may also be negatively influenced. Yet another problem could arise when the data are not missing at random (e.g. when missingness is more likely to occur when functional values are high). In such cases, missing functional chunks may be indeed very insidious because important features of the data distribution may be lost. Furthermore, the presence of functional outliers can be a complication as they may be more difficult to detect when only fragments are available.

4. Functional completion

4.1. Reconstruction of incomplete functions

It is natural to ask whether it is possible to recover not only the missing part of a principal score (and thus to compute the score of an incomplete function) like in Section 3 but also the whole missing part of the trajectory (and thus to reconstruct the whole functional variable). The answer is positive.

In the population version of the problem, the best prediction of X_M by a function of X_O in the sense of the mean integrated prediction squared error is the conditional expectation $E(X_M|X_O)$. It is in general a non-linear operator from $L^2(O)$ to $L^2(M)$ and, similarly to the case of principal scores, we consider its best continuous linear approximation. Assuming for simplicity that the functional variable has mean 0, the minimization problem to be solved is

$$\min_{\mathscr{A}:\|\mathscr{A}\|_\infty<\infty} E(\|X_M - \mathscr{A}X_O\|^2),$$

where the solution is looked for in the class of continuous (bounded) linear operators from $L^2(O)$ to $L^2(M)$ (by $\|\cdot\|_\infty$ we denote the operator norm). We see (by Fréchet differentiation or direct computation) that solving this minimization is equivalent to solving the (normal) equation $\mathscr{A}\mathscr{R}_{OO} = \mathscr{R}_{MO}$. This suggests the solution $\tilde{\mathscr{A}} = \mathscr{R}_{MO}\mathscr{R}_{OO}^{-1}$ and the best linear prediction of X_M in the form $\tilde{X}_M = \tilde{\mathscr{A}}X_O$. From now on, we assume the existence of a bounded solution, i.e. we assume that $\|\mathscr{R}_{MO}\mathscr{R}_{OO}^{-1}\|_\infty < \infty$. Similarly to the case of principal scores, the inverse problem to be solved is ill posed. Using ridge regularization we obtain the solution $\tilde{\mathscr{A}}^{(\alpha)} = \mathscr{R}_{MO}\mathscr{R}_{OO}^{(\alpha)-1}$. The regularized best linear prediction equals $\tilde{X}_M^{(\alpha)} = \tilde{\mathscr{A}}^{(\alpha)}X_O$.

Practically, when the sample $X_{1O_1}, \ldots, X_{nO_n}$ is observed on the subsets $O_1, \ldots, O_n$, we replace the covariance operator by its estimate and set $\hat{\mathscr{A}}_i^{(\alpha)} = \hat{\mathscr{R}}_{M_iO_i}\hat{\mathscr{R}}_{O_iO_i}^{(\alpha)-1}$. The mean function needs to be estimated as well. For the ith curve, the best linear prediction of X_{iM_i} is estimated by

$$\hat{X}_{iM_i}^{(\alpha)} = \hat{\mu}_{M_i} + \hat{\mathscr{A}}_i^{(\alpha)}(X_{iO_i} - \hat{\mu}_{O_i}).$$

To prove the consistency, we assume not only that the solution to the inverse problem (the prediction operator) is bounded but that it is Hilbert–Schmidt. We have a result as follows.

Theorem 2. Let $E(\|X_1\|^4) < \infty$, assumption (3) be satisfied and $\|\mathscr{R}_{M_iO_i}\mathscr{R}_{O_iO_i}^{-1}\|_2 < \infty$. Then

$$E(\|\hat{X}_{iM_i}^{(\alpha)} - \tilde{X}_{iM_i}\|^2) \leqslant O(\alpha^{-3})\,O(n^{-1}) + O(\alpha)$$

as $\alpha \to 0$ and $n \to \infty$. Hence, if $\alpha = \alpha_n$ such that $\alpha_n \to 0$ and $\alpha_n n^{1/3} \to \infty$ as $n \to \infty$, then $\hat{X}_{iM_i}^{(\alpha_n)}$ is a consistent estimator of the best linear prediction $\tilde{X}_{iM_i}$ of X_{iM_i}.

Note that our consistency result is genuinely functional. It is different from theorem 3 of Yao *et al.* (2005a) where it was possible to obtain only a pointwise consistent estimator of the functional variable. The reason is that we assume that the functions are observed fully (or densely in practice) on subsets of the domain whereas Yao *et al.* (2005a) worked in a sparse

observation regime. In other words, we can achieve stronger results because our data contain more information.

The assumption that the prediction operator $\tilde{\mathscr{A}}_i = \mathscr{R}_{M_i O_i} \mathscr{R}_{O_i O_i}^{-1}$ is Hilbert–Schmidt ($\|\tilde{\mathscr{A}}_i\|_2 < \infty$) which is needed for the proof is a strengthening of the basic assumption on the continuity of $\tilde{\mathscr{A}}_i$ ($\|\tilde{\mathscr{A}}_i\|_\infty < \infty$). Assumptions of this type were used in related contexts of, for example, prediction in functional time series (Bosq (2000), chapter 8, and Kargin and Onatski (2008)) and the functional linear model (Yao *et al.*, 2005b; He *et al.*, 2010). It seems possible to replace this assumption by a combination of the condition $\|\tilde{\mathscr{A}}_i\|_\infty < \infty$ and a condition on the eigenvalue sequence $\lambda_{O_i O_i k}$ such that the regularization error can be controlled.

The condition $\|\tilde{\mathscr{A}}_i\|_2 < \infty$ can be written explicitly in terms of the covariance structure of the principal scores of the observed and unobserved part of the function. If the eigendecompositions of $\mathscr{R}_{O_i O_i}$ and $\mathscr{R}_{M_i M_i}$ are

$$\mathscr{R}_{O_i O_i} = \sum_{k=1}^{\infty} \lambda_{O_i O_i k} \varphi_{O_i O_i k} \otimes \varphi_{O_i O_i k},$$
$$\mathscr{R}_{M_i M_i} = \sum_{k=1}^{\infty} \lambda_{M_i M_i k} \varphi_{M_i M_i k} \otimes \varphi_{M_i M_i k}$$

(where '$\otimes$' stands for the tensor product: $(f \otimes g)u = \langle g, u \rangle f$), then we can write

$$\mathscr{R}_{M_i O_i} = \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \gamma_{M_i O_i jk} \varphi_{M_i M_i j} \otimes \varphi_{O_i O_i k},$$

where $\gamma_{M_i O_i jk} = \langle \varphi_{M_i M_i j}, \mathscr{R}_{M_i O_i} \varphi_{O_i O_i k} \rangle = \operatorname{cov}(\langle X_{M_i} - \mu_{M_i}, \varphi_{M_i M_i j} \rangle, \langle X_{O_i} - \mu_{O_i}, \varphi_{O_i O_i k} \rangle)$. Then the operator $\tilde{\mathscr{A}}_i$ is Hilbert–Schmidt whenever

$$\sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \frac{\gamma_{M_i O_i jk}^2}{\lambda_{O_i O_i k}^2} < \infty,$$

which is equivalent to

$$\sum_{j=1}^{\infty} \lambda_{M_i M_i j} \sum_{k=1}^{\infty} \frac{\operatorname{corr}(\langle X_{M_i} - \mu_{M_i}, \varphi_{M_i M_i j} \rangle, \langle X_{O_i} - \mu_{O_i}, \varphi_{O_i O_i k} \rangle)^2}{\lambda_{O_i O_i k}} < \infty.$$

It is seen that this condition combines conditions for the prediction of $\langle X_{M_i} - \mu_{M_i}, \varphi_{M_i M_i j} \rangle$, $j = 1, 2, \ldots$ (compare the inner series above with condition (8)).

4.2. Selection of the regularization parameter

To understand the amount of regularization corresponding to α, we can use the effective degrees of freedom or the proportion of retained variability as defined in equations (9) and (10) respectively. For the selection of α automatically balancing the stability and accuracy of the prediction of X_{iM_i}, we propose a similar cross-validation procedure to that in Section 3.3 for principal scores. The residual sum of squares for the prediction of trajectories on M_i computed for the completely observed curves in the sample is

$$\operatorname{rss}_i(\alpha) = \sum_{k \in C} \|X_{kM_i} - \hat{X}_{kM_i}^{(\alpha)}\|^2.$$

The value of α that is used for the prediction of a function on M_i from its observation on O_i minimizes

$$\mathrm{gcv}_i(\alpha) = \frac{\mathrm{rss}_i(\alpha)}{\{1 - (1/|C|)\mathrm{df}_i(\alpha)\}^2}.$$

4.3. Uncertainty and prediction bands

Theorem 2 shows that $\hat{X}_{iM_i}^{(\alpha_n)}$ consistently estimates the best linear prediction $\tilde{X}_{iM_i}$. We are now interested in the variation of $\hat{X}_{iM_i}^{(\alpha_n)}$ around the target quantity: the unobserved function X_{iM_i}.

Proposition 4. Let the assumptions of theorem 2 be satisfied and let $\alpha_n \to 0$ and $\alpha_n n^{1/4} \to \infty$ as $n \to \infty$. Then $\hat{X}_{iM_i}^{(\alpha_n)} - X_{iM_i}$ is asymptotically distributed (in the sense of weak convergence of probability measures on $L^2([0,1])$) as the mean 0 stochastic process $\tilde{X}_{iM_i} - X_{iM_i}$. The limiting covariance operator is consistently estimated (with respect to the Hilbert–Schmidt norm) by

$$\hat{\mathscr{V}}_i = \hat{\mathscr{R}}_{M_iM_i} - \hat{\mathscr{R}}_{M_iO_i}\hat{\mathscr{R}}_{O_iO_i}^{(\alpha_n)-1}\hat{\mathscr{R}}_{O_iO_i}\hat{\mathscr{R}}_{O_iO_i}^{(\alpha_n)-1}\hat{\mathscr{R}}_{O_iM_i}.$$

If the data are Gaussian, then the limiting stochastic process is Gaussian.

The trace of $\hat{\mathscr{V}}_i$ quantifies the total amount of uncertainty of the linear prediction of X_{iM_i}. It approaches 0 as the Lebesgue measure of the missing region M_i approaches 0, i.e. as we approach a completely observed function. When the measure of the observation period O_i converges to 0, the total prediction uncertainty converges to the trace of $\hat{\mathscr{R}}$, which corresponds to the situation of no information about the ith curve. The scale invariant ratio

$$\mathrm{tr}(\hat{\mathscr{V}}_i)^{1/2}/\mathrm{tr}(\hat{\mathscr{R}})^{1/2} \tag{13}$$

measures the relative prediction error, i.e. the amount of uncertainty about the ith curve as a proportion of the total spread of the distribution of the functional random variable. 1 minus this value corresponds to the reduction of uncertainty that is achieved by the best linear prediction and can be seen as a measure of performance of the completion procedure. Alternatively, we can use $\hat{\mathscr{R}}_{M_iM_i}$ instead of $\hat{\mathscr{R}}$ in the denominator in the relative prediction error, leading to the ratio of the uncertainty about the missing trajectory when the prediction method is used *versus* the uncertainty that there would be about X_{iM_i} if we ignored the observed part.

We use the asymptotic distribution of $\hat{X}_{iM_i}^{(\alpha_n)} - X_{iM_i}$ for the construction of prediction bands for the unobserved part of the trajectory, i.e. regions containing the curve X_{iM_i} with high probability. We consider bands of the form

$$\{(t,x) : \hat{X}_{iM_i}^{(\alpha_n)}(t) - c_{1-\eta}\hat{h}(t) \leqslant x \leqslant \hat{X}_{iM_i}^{(\alpha_n)}(t) + c_{1-\eta}\hat{h}(t),\ t \in M_i\}, \tag{14}$$

where $\hat{h}$ is a function that consistently estimates some limiting function h that is bounded away from zero, and $c_{1-\eta}$ is the $(1-\eta)$-quantile of the random variable $\sup_{t\in M_i} |\tilde{X}_{iM_i}(t) - X_{iM_i}(t)|/h(t)$. This band has asymptotic coverage $1-\eta$. One can choose $\hat{h} = 1$, leading to a band with constant width, but typically one prefers a band whose width at time t reflects the uncertainty of the prediction of the missing function at t. We use $\hat{h}(t) = \max\{\hat{h}_0, \hat{v}_i(t)\}$ where $\hat{v}_i(t)$ is the estimated standard deviation of the limiting predictive distribution at time t, i.e. the square root of the diagonal of the kernel of $\hat{\mathscr{V}}_i$, and $\hat{h}_0$ is a threshold guaranteeing that the limiting function h is bounded away from 0. For example, the choice $\hat{h}_0 = 0.2 \sup_{t\in M_i} \hat{v}_i(t)$ works well in practice. If the distribution of the data can be considered as Gaussian, the quantile $c_{1-\eta}$ can be computed by simulation as follows. Generate a large number of independent realizations of the Gaussian process with mean 0 and covariance operator $\hat{\mathscr{V}}_i$, divide them by $\hat{h}(t)$, compute the maxima of their absolute values and determine the $(1-\eta)$-quantile of this sample. The

simulation of the trajectories and the computation of the maxima are performed on a fine grid of points. Note that the width of the band does not converge to 0 because it is a prediction band, i.e. it must contain, with high probability, a random function.

We conlude this section with a theoretical remark. Although the prediction bands proposed work well in practice, as is documented in the simulation study in Section 5, for a strictly rigorous justification arguments based on proposition 4 (which is a consequence of theorem 2) need to be extended. Proposition 4 guarantees the convergence in distribution in the sense of the topology of the L^2-norm of the Hilbert space $L^2([0,1])$. This justifies the construction of prediction regions in the form of balls in $L^2([0,1])$ which, however, are not practical because they cannot be plotted. For prediction bands, the convergence is needed in the sense of the uniform topology. For this, we need to leave the geometric world of $L^2([0,1])$ and to switch to the space of continuous functions $C([0,1])$. Under modified assumptions (which would include conditions on sample paths, such as Hölder continuity), it seems possible to prove the convergence in the uniform topology. We do not pursue this theoretical study but give arguments indirectly justifying the use of the bands. Suppose that the asymptotic approximation that is suggested by theorem 2 and proposition 4 is considered applicable if the L^2-distance from the limiting variable is sufficiently small. The probability that this L^2-distance exceeds some $\varepsilon > 0$ is, in light of Chebyshev's inequality, bounded as $P(\|\hat{X}_{iM_i}^{(\alpha_n)} - \tilde{X}_{iM_i}\|_2^2 > \varepsilon) \leqslant \varepsilon^{-2} E(\|\hat{X}_{iM_i}^{(\alpha_n)} - \tilde{X}_{iM_i}\|_2^2)$. However, convergence in the L^2-norm does not imply uniform convergence because large deviations may occur on a small set of arguments. Let us compute the Lebesgue measure γ of the set where $|\hat{X}_{iM_i}^{(\alpha_n)} - \tilde{X}_{iM_i}|$ deviates more than ε from 0. We compute $\gamma(\{t : |\hat{X}_{iM_i}^{(\alpha_n)}(t) - \tilde{X}_{iM_i}(t)| > \varepsilon\}) \leqslant \varepsilon^{-2} \|\hat{X}_{iM_i}^{(\alpha_n)} - \tilde{X}_{iM_i}\|_2^2$ by using Chebyshev's inequality. Taking expectations on both sides, we obtain on the right-hand side the same bound as before. Hence, if the bound is considered to be sufficiently small for the asymptotic approximation in the L^2-norm to be applicable, then also the expected Lebesgue measure of the set of large pointwise deviations will be negligible.

5. Simulations

A simulation study was designed to address the following goals: to investigate the performance of generalized cross-validation as a selector of the regularization parameter, to verify the validity and accuracy of the prediction intervals and bands and to explore the effect of the observation pattern.

We generate random samples of curves of the form

$$X(t) = \sum_{k=1}^{100} 2^{1/2} \nu_k^{1/2} \xi_k \cos(2\pi k t) + \sum_{k=1}^{100} 2^{1/2} \omega_k^{1/2} \eta_k \sin(2\pi k t), \qquad t \in [0,1],$$

where ξ_k and η_k are independent standard normal variables and the eigenvalues are of the form $\nu_k = 3^{-(2k-1)}$ and $\omega_k = 3^{-2k}$. The three most important components represent 67%, 22% and 7% of the total variability. For each curve we generate independently a random period on which this curve is not observed. The functional values on this period are removed. For the ith function, the missing period M_i is simulated in the form $M_i = [C_i - E_i, C_i + E_i] \cap [0,1]$ with $C_i = dU_{i,1}^{1/2}$ and $E_i = fU_{i,2}$, where d and f are parameters and $U_{i,1}$ and $U_{i,2}$ are independent variables uniformly distributed on $[0,1]$. The performance of our procedures is measured on one curve in the sample, say X_1. For this curve, we use a fixed (non-random) missing period to guarantee that values computed from different simulation runs have the same meaning. In all simulations, we use $L = 1000$ repetitions.

Table 1. Performance of the generalized cross-validation selection procedure†

Target quantity (and its variability)	n	*MSPE for $\alpha = c\alpha_{\mathrm{gcv}}$ and the following values of c:*					*Median degrees of freedom for $\alpha = \alpha_{\mathrm{gcv}}$*
		0.04	*0.2*	*1*	*5*	*25*	
Score 1 (333)	100	1.91	1.55	1.32	1.61	3.78	7.68
	500	0.60	0.44	0.36	0.42	1.07	12.73
Score 2 (111)	100	0.46	0.37	0.35	0.44	0.80	8.61
	500	0.16	0.13	0.12	0.15	0.27	13.71
Score 3 (37)	100	1.45	1.13	0.95	1.08	2.00	8.62
	500	0.48	0.34	0.28	0.29	0.53	13.71
Missing trajectory (500)	100	10.07	7.90	6.95	8.24	15.16	7.98
	500	4.04	2.79	2.24	2.30	3.48	15.02

†MSPE and the variability of the target quantity are multiplied by 1000.

For the first two sets of simulations, we set $d = 1.4$ and $f = 0.2$. This leads to an observation pattern with similar characteristics to those in our motivating data set. The cross-sectional probability of observation ranges from 99% at time 0 to 79% at time 1. The percentage of complete curves is 39%. The median length of the missing period (given the curve has a missing period) is 0.15. For the curve X_1, on which the performance is measured, we set $M_1 = (0.4, 0.7)$.

First, we investigate the performance of generalized cross-validation based on complete curves. As a measure of quality of the prediction of a missing quantity, we use the mean-squared prediction error MSPE which is the average over all simulation runs of the squared distances of the predicted value and the true value, i.e. $L^{-1}\Sigma_{l=1}^{L}(\hat{\beta}_{1jM_1}^{(\alpha)[l]} - \hat{\beta}_{1jM_1}^{[l]})^2$ for the jth score and $L^{-1}\Sigma_{l=1}^{L}\|\hat{X}_{1M_1}^{(\alpha)[l]} - X_{1M_1}^{[l]}\|^2$ for the missing part of the trajectory, where the superscript $[l]$ indicates that the value pertains to the lth generated sample. Table 1 shows values of the mean-squared prediction error for the first three principal scores and for the missing part of the trajectory. Table 1 also includes the variability of the target quantities (i.e. the true eigenvalues for the scores and the trace of the true covariance operator $\mathscr{R}$ for the trajectory) to put the values into context. The mean-squared prediction error is reported for α set to the value selected by generalized cross-validation and to values slightly smaller or bigger in the form of multiples of the selected value. We see that the method successfully approximates the best value of α and can be recommended as the tuning parameter selector. The accuracy increases with increasing sample size n; however, it should be noted that the mean-squared prediction error cannot converge to 0 because there is always some uncertainty due to the randomness of the target quantity, as discussed in Sections 3.4 and 4.3. The last column of Table 1 reports the median of the effective degrees of freedom corresponding to the selected value of α. It is seen that in all cases the typical number of degrees of freedom is in a reasonable relation to the sample size.

The second set of simulations explores the properties of the approximate distribution of the deviation of the prediction from the predicted quantity that is established in propositions 3 and 4. We simulate from the same distribution and observation pattern as before. The regularization parameter is selected by generalized cross-validation. We consider prediction intervals and bands of the form (11) and (14) respectively, with nominal coverage 95%. We compute bands with both constant and variable width, as discussed in Section 4.3. Empirical coverage probabilities (i.e. the percentage of cases when the unobserved quantity was covered by the constructed region) are reported in Table 2. We see that the intervals and bands proposed have coverage that is close

Table 2. Empirical coverage of prediction regions (intervals for scores; bands with constant and variable width for curves) and the median relative error measure

n	*Results for score 1*		*Results for score 2*		*Results for score 3*		*Results for missing trajectory*		
	Coverage (%)	*Median relative error*	*Coverage (%)*	*Median relative error*	*Coverage (%)*	*Median relative error*	*Coverage (constant width) (%)*	*Coverage (variable width) (%)*	*Median relative error*
100	97.2	0.073	95.2	0.056	94.5	0.143	94.3	96.7	0.123
500	97.4	0.042	95.0	0.036	96.3	0.092	94.2	98.4	0.07

Table 3. Standardized mean-squared prediction error for different observation patterns

n	*Observation pattern (X_1)*	*Results for score 1 and the following observation patterns (sample):*		*Results for score 2 and the following observation patterns (sample):*		*Results for score 3 and the following observation patterns (sample):*		*Results for missing trajectory and the following observation patterns (sample):*	
		A	*B*	*A*	*B*	*A*	*B*	*A*	*B*
100	I	0.022	0.045	0.052	0.093	0.035	0.067	0.040	0.075
	II	0.039	0.073	0.078	0.128	0.107	0.155	0.076	0.136
500	I	0.006	0.013	0.018	0.031	0.010	0.023	0.013	0.024
	II	0.019	0.027	0.037	0.051	0.060	0.076	0.035	0.051

to the nominal coverage and, therefore, provide useful information on the probable values of the scores or the missing trajectory. Table 2 also reports the median of relative error measures (12) and (13). For instance, we can see that the approximate distribution is relatively more spread for less variable (higher index) scores. This is in line with conclusions from Table 1 where we observed a similar relationship between MSPE and the variability of the target quantity. Hence the relative error measures (12) and (13), which can be computed from the data, seem to be valuable indicators of the accuracy of the reconstruction procedure.

In the last set of simulations, we study the effect of the observation pattern on the accuracy of our methods. We vary the amount of observed information both for X_1 (whose characteristics are to be reconstructed) and for the whole sample (which is used to learn the reconstruction procedure). Two settings are used for the missing period of X_1: I, $M_1 = (0.4, 0.7)$; II, $M_1 = (0.4, 0.9)$. For the simulation of the missing periods of other curves in the sample, we simulate M_i of the form given earlier in this section, with parameter pairs A, $d = 1.4$ and $f = 0.2$, and B, $d = 1.4$ and $f = 0.5$. Basic characteristics of the observation pattern for A were discussed before; for B, the cross-sectional observation probability varies from 95% at $t = 0$ to 50% at $t = 1$, 21% of curves are complete and the average length of missing periods (among incomplete curves) is 0.29. Configuration IA was used in the first two sets of simulations; other combinations contain less observed information. Results are reported in Table 3 where mean-squared prediction errors are presented after standardization by the true variance of the predicted quantity, i.e. by the variance of the missing part of the score, $\mathrm{var}(\beta_{1jM_1})$, or by the trace of the covariance operator of the missing part of the trajec-

tory, $\mathrm{tr}(\mathscr{R}_{M_1 M_1})$; after this standardization it is possible to compare values under pattern I with their counterparts computed under II. We see that the precision of estimation decreases as the amount of observed information (either on the curve of interest or on the sample) decreases.

6. An illustration: ambulatory blood pressure monitoring data

Heart rate profiles displayed in Fig. 1 and their first derivative plotted in Fig. 2 were obtained from raw observations by penalized spline smoothing described in the supplementary file that is available on line. The curves were registered by shifting the individual timescales so that every person's bed time is 23 (i.e. 11 p.m.); individual bed times were available from a questionnaire. The methodology that is developed in this paper requires that the observation periods be independent of the curves. The expert opinion is that this is a realistic assumption; in addition, we performed exploratory graphical checks that did not indicate any problem with regard to this assumption.

From the shape of the mean functions of the profiles and their first derivatives it is obvious that on average heart rate profiles have a decreasing shape in this part of the day and they decrease fastest around the bed time. We wish to understand the main sources of variability between individual heart rate profiles. In Fig. 3 we plot the first three eigenfunctions of the profiles and of their derivatives as perturbations of the mean shape (see Ramsay and Silverman (2005), section 8.3.1) i.e. we plot the mean profile plus and minus a suitable multiple of each eigenfunction (the eigenfunctions are multiplied by $0.9\hat{\lambda}_j^{1/2}$). For the profiles, we see that the most important component is the global level of heart rate, followed by a component describing the difference between the day and night values and a component that can be interpreted as a time shift. In terms of the first derivative, the first component quantifies the global level of the speed of decrease, the second component captures a shift in time and the third characterizes whether the individual's heart rate decreases rather suddenly or more gradually. The first three components explain a large proportion of the total variability and provide enough flexibility to capture individual shape features, e.g. the increasing trend of some curves in regions where the mean and most curves decrease.

Let us now focus on the individual level. To illustrate our prediction method for principal scores, we first consider the curve that is plotted as short dashes in Figs 1(b) and 2(b). The functional values are missing on a subset of the time interval and hence the principal scores cannot be computed directly. They can, however, be predicted. We give the results for the profile only

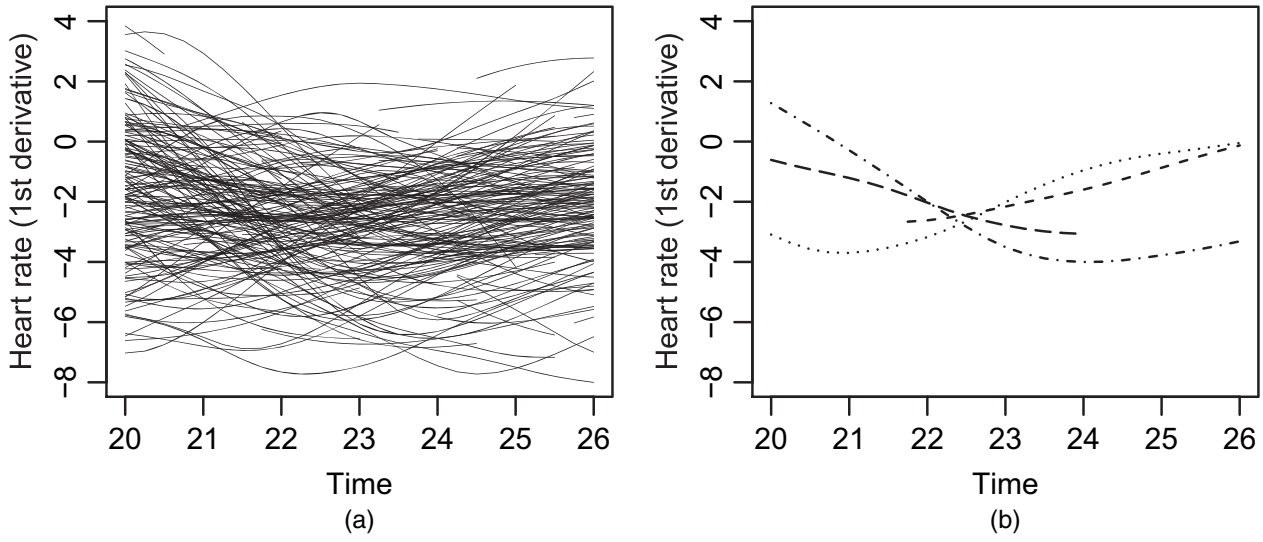

Fig. 2. (a) Subset of the sample of the first derivatives of heart rate profiles and (b) several curves in detail

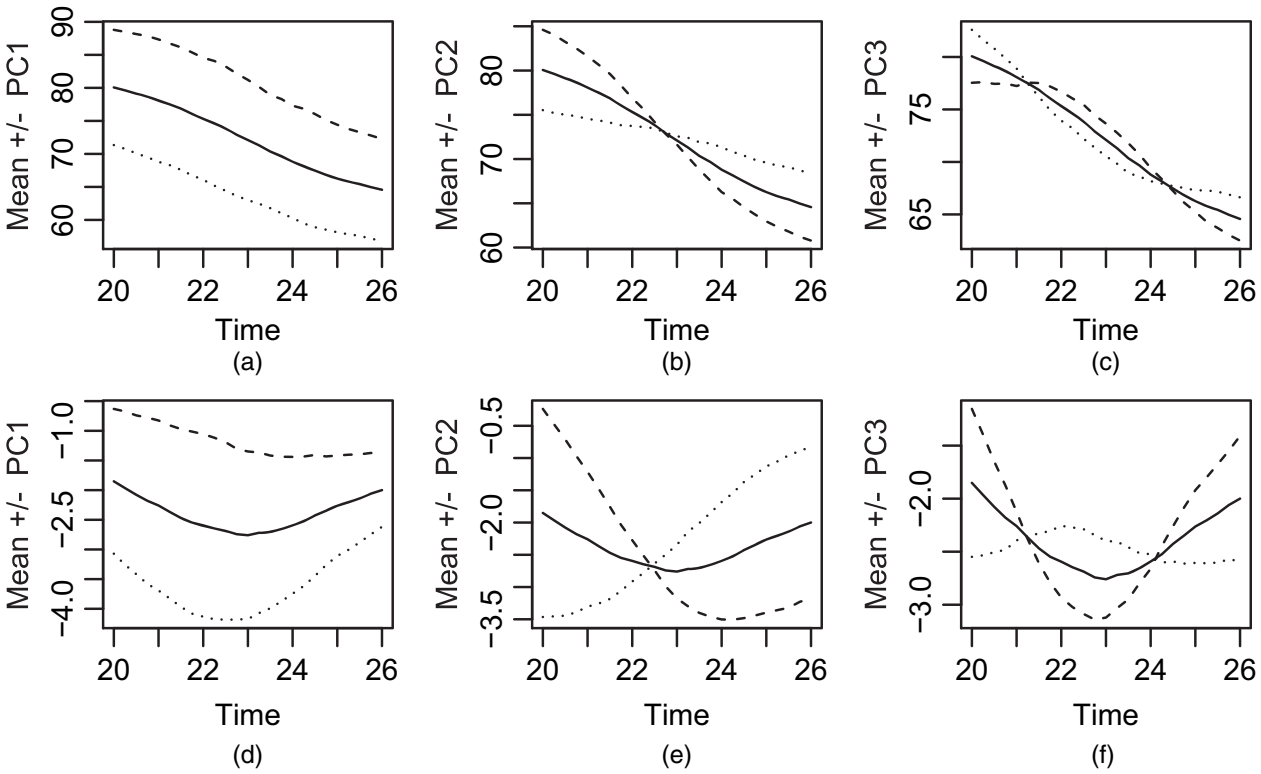

Fig. 3. (a)–(c) First three eigenfunctions of heart rate profiles and (d)–(f) of their first derivative plotted as perturbations (- - - - -, · · · · · ·) of the mean (———): (a) principal component 1, 87.2%; (b) principal component 2, 9.3%; (c) principal component 3, 2.1%; (d) principal component 1, 59.5%; (e) principal component 2, 33.8%; (f) principal component 3, 4.5%

(one can proceed analogously for the first derivative). The predicted values for the first three components are $(-28.7, 2.9, -1.9)$. Their prediction standard deviations quantifying the uncertainty are $(1.7, 2.3, 1.8)$. Mainly for the first two components they are relatively small compared with the standard deviations of the intrinsic variability $(24.0, 7.8, 3.7)$ (the square root of the eigenvalues); the corresponding relative errors are $(0.07, 0.29, 0.48)$. It is not surprising that the best precision is achieved for the first component: this component dominates the spectrum and is quite simple (constant), so even a fraction of the curve provides relatively much information about the score. Next, we illustrate the method on the completely observed function plotted as the chain curves in Figs 1(b) and 2(b) from which we artificially remove observations in the time interval [23.75,26]. Using the remaining part for the prediction, we estimate the scores by $(5.84, 4.43, 4.18)$ (with prediction standard deviations $(2.12, 2.68, 2.01)$), which is quite close to the true values $(5.76, 4.55, 4.32)$ computed from the complete curve (recall, however, that there will always be some random non-vanishing discrepancy between the predicted and true values because we predict random variables by their conditional expectations).

Finally, we illustrate the functional reconstruction procedure. In Fig. 4 we plot the two curves (and their derivatives) that we considered before and the reconstructed missing parts along with 95% prediction bands. For the originally complete function (Figs 4(b) and 4(d)), we chose a difficult scenario: the missing period is relatively large (2.25 h) and it contains a non-trivial change of shape of the curve mainly in terms of the first derivative which is decreasing in the observed region and increasing in the missing period. However, it is seen that the completion procedure can recover the missing part of information as the predicted curve (thick) approximates very well the true function (thin). It is interesting that our method captures to some extent the presence

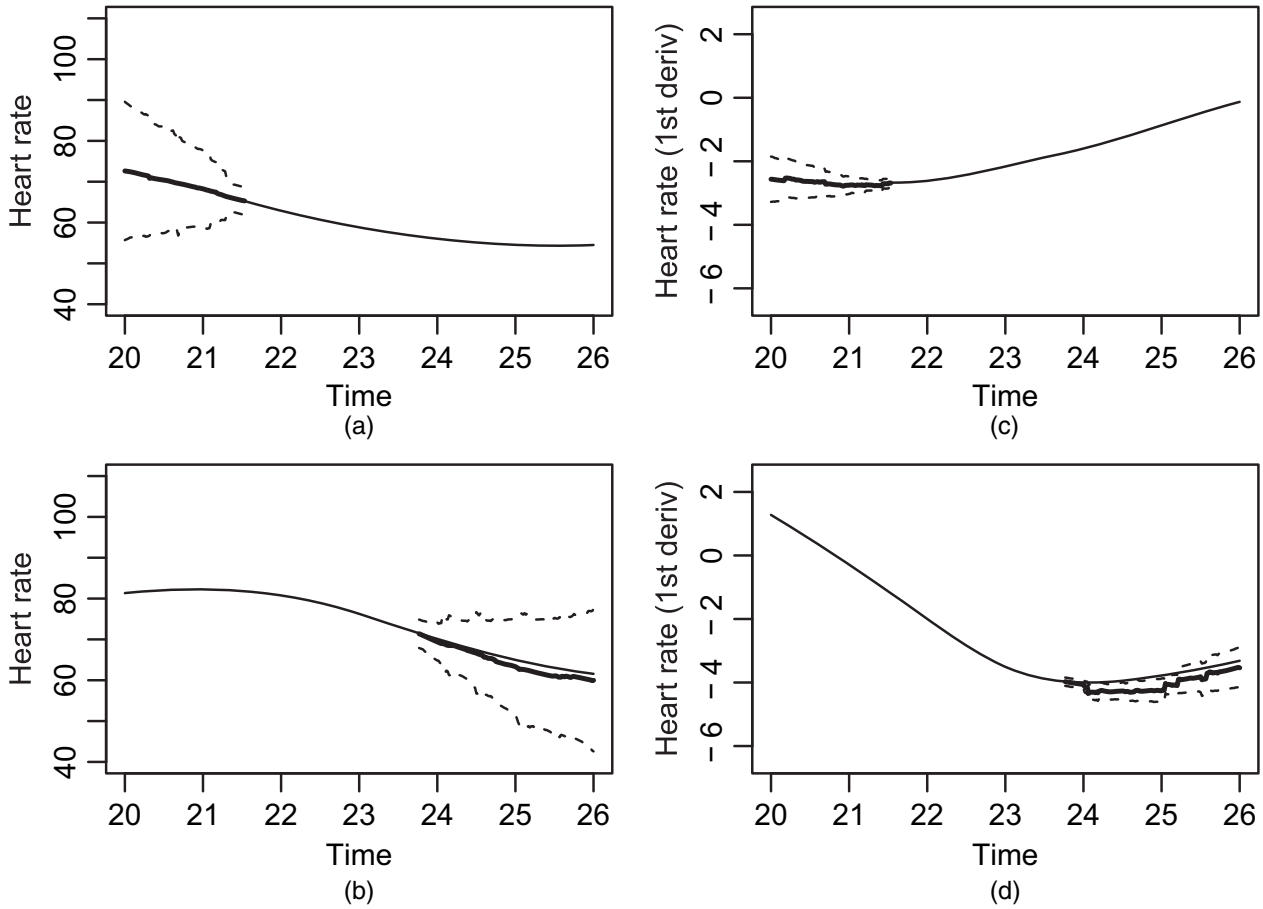

Fig. 4. (a), (b) Observed (———) and reconstructed (———) heart rate profiles and (c), (d) derivatives along with 95% prediction bands for (a), (c) an incompletely observed curve and (b), (d) a complete curve with an artificially introduced missing period

of a local minimum in the first derivative. This illustrates the usefulness of the reconstruction procedure: without it important shape features like this would be concealed from the analyst. At first glance, some of the bands may seem to be wide but one needs to keep in mind that they are prediction (not confidence) bands and, therefore, must cover the random trajectory (rather than a non-random function) with a high probability. The uncertainty of the completion is in fact not big in proportion to the intrinsic variability of the stochastic process: the relative error is 0.10 and 0.11 for the curves in Figs 4(a) and 4(c), and 4(b) and 4(d) respectively. A referee pointed out that the prediction bands for the derivatives are narrower than those for the curves. This is not a general phenomenon: it is possible to construct simple examples with prediction bands for derivatives that are wider than those for curves or examples with no such inequality. Differentiation is an operation that changes the covariance structure of functional data in a complex manner.

We compared our method with that of Yao *et al.* (2005a) applied to the raw heart rate values (not preprocessed by smoothing). Although their method was primarily developed for sparsely observed curves, it can be also used in our situation. Main results regarding the covariance structure of the profiles were similar for both methods. The proportion of variance explained by the first three principal components was 82.9%, 10.8% and 3.4%. The first three eigenfunctions had a similar shape and interpretation with both methods. There was a high degree of agreement between principal scores that were obtained by the two methods. The method of Liu and Müller (2009) can reconstruct derivatives. However, our method seems to be the only currently available

method that can perform principal component analysis of derivatives under incompleteness. This is an important asset of our method over the other approach provided that the data are sufficiently dense on subsets of the domain.

Acknowledgements

This work was done within the 'Swiss kidney project on genes in hypertension', which is a collaboration between Murielle Bochud (Principal Investigator), M. Burnier, O. Devuyst, P.-Y. Martin, M. Mohaupt, F. Paccaud, A. Péchère-Bertschi, B. Vogt, D. Ackermann, H. Alwan, Y. Bouatou, N. Dhayat, G. Ehret, I. Guessous, P. Monney, M.-E. Mueller, B. Ponte, M. Pruijm, S. Reverdin, P. Vuistiner, Z. Kutalik and S. Estoppey. The project was funded by the Swiss National Science Foundation. Special thanks are given to Murielle Bochud for her support and interest, and for her understanding of the importance of methodological developments in statistics. The hospitality of the Institute of Social and Preventive Medicine Lausanne is gratefully acknowledged. I am also grateful to the Joint Editor, the Associate Editor and two referees for their interesting comments and encouragement.

Appendix A: Main proofs

Here we prove theorems 1 and 2. Propositions 1–4 are proven in the supplementary document that is available on line. Recall that we denote by $\|\cdot\|$ the L^2-norm of square integrable functions on a domain S that is obvious from the context (S will be $[0,1]$ or O_i or M_i). For linear operators, the symbols $\|\cdot\|_\infty$ and $\|\cdot\|_2$ are used for the operator norm and the Hilbert–Schmidt norm respectively, where the operator will be a mapping between $L^2(S_1)$ and $L^2(S_2)$ with S_1 and S_2 that is obvious from the context. For definitions of basic notions from operator theory, we refer to Bosq (2000).

A.1. Proof of theorem 1

We neglect the fact that the data are centred by the estimated mean function and assume that the mean is known and equal to 0. The result remains valid when the curves are centred empirically, as the additional terms are negligible. It is enough to prove the inequality in the statement of the theorem; the remaining assertions follow easily. We write $|\hat{\beta}^{(\alpha)}_{ijM_i} - \tilde{\beta}_{ijM_i}| \leqslant |\hat{\beta}^{(\alpha)}_{ijM_i} - \tilde{\beta}^{(\alpha)}_{ijM_i}| + |\tilde{\beta}^{(\alpha)}_{ijM_i} - \tilde{\beta}_{ijM_i}|$, which is a decomposition into the estimation error and approximation error. If we show that both errors converge in $L^2(P)$ to 0, the result will follow.

We denote the approximation error $A_1 = |\tilde{\beta}^{(\alpha)}_{ijM_i} - \tilde{\beta}_{ijM_i}|$ and compute

$$\begin{aligned}
E(A_1^2) &= E\{\langle X_{iO_i}, \tilde{a}^{(\alpha)}_{ij} - \tilde{a}^2_{iji}\rangle\} \\
&= \|\mathscr{R}^{1/2}_{O_iO_i}(\tilde{a}^{(\alpha)}_{ij} - \tilde{a}_{ij})\|^2 \\
&= \|\mathscr{R}^{1/2}_{O_iO_i}(\mathscr{R}^{(\alpha)-1}_{O_iO_i} - \mathscr{R}^{-1}_{O_iO_i})r_{ij}\|^2 \\
&= \sum_{k=1}^{\infty} \lambda_{O_iO_ik}\left(\frac{1}{\lambda_{O_iO_ik}+\alpha} - \frac{1}{\lambda_{O_iO_ik}}\right)^2 \langle r_{ij}, \varphi_{O_iO_ik}\rangle^2 \\
&= \alpha \sum_{k=1}^{\infty} \frac{\alpha\lambda_{O_iO_ik}}{(\lambda_{O_iO_ik}+\alpha)^2} \frac{\langle r_{ij}, \varphi_{O_iO_ik}\rangle^2}{\lambda^2_{O_iO_ik}} \\
&= O(\alpha),
\end{aligned}$$

where $\lambda_{O_iO_ik}$ and $\varphi_{O_iO_ik}$ are the eigenvalues and eigenfunctions of $\mathscr{R}_{O_iO_i}$ and the result follows from the fact that $\alpha\lambda_{O_iO_ik}/(\lambda_{O_iO_ik}+\alpha)^2 \leqslant 1$ and Picard's condition (7).

Let us turn to the estimation error $|\hat{\beta}^{(\alpha)}_{ijM_i} - \tilde{\beta}^{(\alpha)}_{ijM_i}|$. The computation of expectations is complicated by the fact that the quantities $\hat{\mathscr{R}}_{O_iO_i}$ and $\hat{r}_{ij}$ are obtained from the whole sample including the ith function and thus are dependent on the ith function. We overcome this complication by first considering a modified problem with estimates of $\mathscr{R}_{O_iO_i}$ and r_{ij} independent of the ith function and then showing

that this modification is asymptotically negligible. Specifically, we introduce $\hat{\beta}^{(\alpha)}_{ijM_i(-i)} = \hat{\mathscr{R}}^{(\alpha)-1}_{O_iO_i(-i)}\hat{r}_{ij(-i)}$ with $\hat{\mathscr{R}}^{(\alpha)}_{O_iO_i(-i)} = \hat{\mathscr{R}}_{O_iO_i(-i)} + \alpha\mathscr{I}_{O_i}$ and $\hat{r}_{ij(-i)} = \hat{\mathscr{R}}_{O_iM_i(-i)}\hat{\varphi}_{jM_i(-i)}$. Here $\hat{\mathscr{R}}_{O_iO_i(-i)}$ and $\hat{\mathscr{R}}_{O_iM_i(-i)}$ are suboperators of the estimated covariance operator $\hat{\mathscr{R}}_{(-i)}$ that is computed from all functions except the ith, and $\hat{\varphi}_{jM_i(-i)}$ is a subfunction of the jth eigenfunction $\hat{\varphi}_{j(-i)}$ of $\hat{\mathscr{R}}_{(-i)}$. We decompose $|\hat{\beta}^{(\alpha)}_{ijM_i} - \tilde{\beta}^{(\alpha)}_{ijM_i}|$ as follows:

$$|\hat{\beta}^{(\alpha)}_{ijM_i} - \tilde{\beta}^{(\alpha)}_{ijM_i}| \leqslant |\hat{\beta}^{(\alpha)}_{ijM_i} - \hat{\beta}^{(\alpha)}_{ijM_i(-i)}| + |\hat{\beta}^{(\alpha)}_{ijM_i(-i)} - \tilde{\beta}^{(\alpha)}_{ijM_i}|, \tag{15}$$

and we show that both terms converge in $L^2(P)$ to 0.

For the second term on the right-hand side in inequality (15), $A_2 = |\hat{\beta}^{(\alpha)}_{ijM_i(-i)} - \tilde{\beta}^{(\alpha)}_{ijM_i}|$, we have

$$\begin{aligned} E(A_2^2) &= E\{E(|\hat{\beta}^{(\alpha)}_{ijM_i(-i)} - \tilde{\beta}^{(\alpha)}_{ijM_i}|^2 | \{X_{kO_k} : k \neq i\})\} \\ &= E\{E(|\langle X_{iO_i}, \hat{a}^{(\alpha)}_{ij(-i)} - \tilde{a}^{(\alpha)}_{ij}\rangle|^2 | \{X_{kO_k} : k \neq i\})\} \\ &= E\{\|\mathscr{R}^{1/2}_{O_iO_i}(\hat{a}^{(\alpha)}_{ij(-i)} - \tilde{a}^{(\alpha)}_{ij})\|^2\}. \end{aligned}$$

Using the definitions of $\hat{a}^{(\alpha)}_{ij(-i)}$ and $\tilde{a}^{(\alpha)}_{ij}$ and the triangle inequality, we obtain

$$\begin{aligned} \|\mathscr{R}^{1/2}_{O_iO_i}(\hat{a}^{(\alpha)}_{ij(-i)} - \tilde{a}^{(\alpha)}_{ij})\| &\leqslant \|\mathscr{R}^{1/2}_{O_iO_i}\hat{\mathscr{R}}^{(\alpha)-1}_{O_iO_i(-i)}(\hat{\mathscr{R}}_{O_iM_i(-i)} - \mathscr{R}_{O_iM_i})\hat{\varphi}_{jM_i(-i)}\| \\ &\quad + \|\mathscr{R}^{1/2}_{O_iO_i}\hat{\mathscr{R}}^{(\alpha)-1}_{O_iO_i(-i)}\mathscr{R}_{O_iM_i}(\hat{\varphi}_{jM_i(-i)} - \hat{s}_j\varphi_{jM_i})\| \\ &\quad + \|\mathscr{R}^{1/2}_{O_iO_i}(\hat{\mathscr{R}}^{(\alpha)-1}_{O_iO_i(-i)} - \mathscr{R}^{(\alpha)-1}_{O_iO_i})\mathscr{R}_{O_iM_i(-i)}\varphi_{jM_i}\| \end{aligned}$$

with $\hat{s}_j = \operatorname{sgn}\langle\hat{\varphi}_{j(-i)}, \varphi_j\rangle$. Denote these three terms A_{21}, A_{22} and A_{23} respectively. We see that

$$A_{21} \leqslant \|\mathscr{R}^{1/2}_{O_iO_i}\|_\infty \|\hat{\mathscr{R}}^{(\alpha)-1}_{O_iO_i(-i)}\|_\infty \|\hat{\mathscr{R}}_{O_iM_i(-i)} - \mathscr{R}_{O_iM_i}\|_\infty \|\hat{\varphi}_{jM_i(-i)}\|.$$

Here, $\|\mathscr{R}^{1/2}_{O_iO_i}\|_\infty$ is a finite constant, $\|\hat{\mathscr{R}}^{(\alpha)-1}_{O_iO_i(-i)}\|_\infty \leqslant \alpha^{-1}$ and $\|\hat{\varphi}_{jM_i(-i)}\| \leqslant \|\hat{\varphi}_{j(-i)}\| = 1$. Using proposition 1 we obtain $E(A_{21}^2) \leqslant \alpha^{-2}\, O(n^{-1})$. For the term A_{22} we have the bound

$$A_{22} \leqslant \|\mathscr{R}^{1/2}_{O_iO_i}\|_\infty \|\hat{\mathscr{R}}^{(\alpha)-1}_{O_iO_i(-i)}\|_\infty \|\mathscr{R}_{O_iM_i}\|_\infty \|\hat{\varphi}_{jM_i(-i)} - \hat{s}_j\varphi_{jM_i}\|.$$

In light of proposition 2, we see that $E(\|\hat{\varphi}_{jM_i(-i)} - \hat{s}_j\varphi_{jM_i}\|^2) \leqslant E(\|\hat{\varphi}_{j(-i)} - \hat{s}_j\varphi_j\|^2) = O(n^{-1})$. This implies that $E(A_{22}^2) \leqslant \alpha^{-2}\, O(n^{-1})$. For the term A_{23}, first note that

$$\begin{aligned} \hat{\mathscr{R}}^{(\alpha)-1}_{O_iO_i(-i)} - \mathscr{R}^{(\alpha)-1}_{O_iO_i} &= \mathscr{R}^{(\alpha)-1}_{O_iO_i}(\mathscr{R}^{(\alpha)}_{O_iO_i} - \hat{\mathscr{R}}^{(\alpha)}_{O_iO_i(-i)})\hat{\mathscr{R}}^{(\alpha)-1}_{O_iO_i(-i)} \\ &= \mathscr{R}^{(\alpha)-1}_{O_iO_i}(\mathscr{R}_{O_iO_i} - \hat{\mathscr{R}}_{O_iO_i(-i)})\hat{\mathscr{R}}^{(\alpha)-1}_{O_iO_i(-i)}. \end{aligned}$$

Therefore, we see that

$$A_{23} \leqslant \|\mathscr{R}^{1/2}_{O_iO_i}\mathscr{R}^{(\alpha)-1}_{O_iO_i}\|_\infty \|\hat{\mathscr{R}}_{O_iO_i(-i)} - \mathscr{R}_{O_iO_i}\|_\infty \|\hat{\mathscr{R}}^{(\alpha)-1}_{O_iO_i(-i)}\|_\infty \|\mathscr{R}_{O_iM_i}\|_\infty \|\hat{\varphi}_{jM_i(-i)}\|.$$

The first, third and fifth term are dominated by $\alpha^{-1/2}$, α^{-1} and 1 respectively. The fourth term is a finite constant. Using these bounds and proposition 1 we obtain $E(A_{23}^2) \leqslant \alpha^{-3}\, O(n^{-1})$. Hence with the help of the Cauchy–Schwarz inequality we finally obtain that $E(A_2^2) \leqslant \alpha^{-3}\, O(n^{-1})$.

It remains to analyse the first term on the right-hand side of inequality (15). It reflects the effect of omitting the ith observation in the estimation. As this effect is of order $O(n^{-2})$ in terms of mean-squared difference, this term is negligible compared with the second term. In particular, it can be shown that $E\{(\hat{\beta}^{(\alpha)}_{ijM_i} - \hat{\beta}^{(\alpha)}_{ijM_i(-i)})^2\} \leqslant \alpha^{-3}\, O(n^{-2})$. We omit the technical details.

A.2. Proof of theorem 2

To simplify the proof of theorem 2 we assume that the mean is known to be 0 and no centring is performed. The difference due to the estimation of the mean is of negligible order in comparison with other terms. Similarly to the proof of theorem 1, we split the prediction error into the estimation error and regularization error as follows:

$$\|\hat{X}^{(\alpha)}_{iM_i} - \tilde{X}_{iM_i}\| \leqslant \|\hat{X}^{(\alpha)}_{iM_i} - \tilde{X}^{(\alpha)}_{iM_i}\| + \|\tilde{X}^{(\alpha)}_{iM_i} - \tilde{X}_{iM_i}\|.$$

For the regularization error we compute

$$
\begin{aligned}
E(\|\tilde{X}_{iM_i}^{(\alpha)} - \tilde{X}_{iM_i}\|^2) &= \|(\tilde{\mathscr{A}}_i^{(\alpha)} - \tilde{\mathscr{A}}_i)\mathscr{R}_{O_iO_i}^{1/2}\|_2^2 \\
&= \|\alpha \mathscr{R}_{M_iO_i}\mathscr{R}_{O_iO_i}^{-1}\mathscr{R}_{O_iO_i}^{(\alpha)-1}\mathscr{R}_{O_iO_i}^{1/2}\|_2^2 \\
&\leqslant \alpha \|\mathscr{R}_{M_iO_i}\mathscr{R}_{O_iO_i}^{-1}\|_2^2 \, \|\alpha^{1/2}\mathscr{R}_{O_iO_i}^{(\alpha)-1}\mathscr{R}_{O_iO_i}^{1/2}\|_\infty^2 \\
&= \alpha \|\tilde{\mathscr{A}}_i\|_2^2 \left(\sup_{k\in\mathbb{N}} \frac{\alpha^{1/2}\lambda_{O_iO_ik}^{1/2}}{\lambda_{O_iO_ik}+\alpha}\right)^2 \\
&\leqslant O(\alpha).
\end{aligned}
$$

We turn to the estimation error. Similarly to the proof of theorem 1 we avoid the dependence between $\hat{\mathscr{A}}_i^{(\alpha)}$ and X_{iO_i} in $\hat{X}_{iM_i}^{(\alpha)} = \hat{\mathscr{A}}_i^{(\alpha)} X_{iO_i}$ by considering $\hat{X}_{iM_i(-i)}^{(\alpha)} = \hat{\mathscr{A}}_{i(-i)}^{(\alpha)} X_{iO_i}$, where the estimator of the covariance operator in the prediction operator is replaced by its analogue based on all curves except the ith. The difference is negligible in comparison with the remaining terms; for an analogous discussion see the proof of theorem 1. The modified estimation error equals

$$
\begin{aligned}
E(\|\hat{X}_{iM_i(-i)}^{(\alpha)} - \tilde{X}_{iM_i}^{(\alpha)}\|^2) &= E\{\|(\hat{\mathscr{R}}_{M_iO_i(-i)}\hat{\mathscr{R}}_{O_iO_i(-i)}^{(\alpha)-1} - \mathscr{R}_{M_iO_i}\mathscr{R}_{O_iO_i}^{(\alpha)-1})\mathscr{R}_{O_iO_i}^{1/2}\|_2^2\} \\
&\leqslant E\{\|(\hat{\mathscr{R}}_{M_iO_i(-i)} - \mathscr{R}_{M_iO_i})\hat{\mathscr{R}}_{O_iO_i(-i)}^{(\alpha)-1}\mathscr{R}_{O_iO_i}^{1/2}\|_2 \\
&\quad + \|\mathscr{R}_{M_iO_i}(\hat{\mathscr{R}}_{O_iO_i(-i)}^{(\alpha)-1} - \mathscr{R}_{O_iO_i}^{(\alpha)-1})\mathscr{R}_{O_iO_i}^{1/2}\|_2\}^2.
\end{aligned}
$$

The proof is complete on computing

$$
\begin{aligned}
E\{\|(\hat{\mathscr{R}}_{M_iO_i(-i)} - \mathscr{R}_{M_iO_i})\hat{\mathscr{R}}_{O_iO_i(-i)}^{(\alpha)-1}\mathscr{R}_{O_iO_i}^{1/2}\|_2^2\} &\leqslant E(\|\hat{\mathscr{R}}_{M_iO_i(-i)} - \mathscr{R}_{M_iO_i}\|_2^2 \|\hat{\mathscr{R}}_{O_iO_i(-i)}^{(\alpha)-1}\|_\infty^2 \|\mathscr{R}_{O_iO_i}^{1/2}\|_\infty^2) \\
&\leqslant E(\|\hat{\mathscr{R}}_{M_iO_i(-i)} - \mathscr{R}_{M_iO_i}\|_2^2 \alpha^{-2}\lambda_{O_iO_i1}) \\
&= \alpha^{-2} O(n^{-1}),
\end{aligned}
$$

$$
\begin{aligned}
E\{\|\mathscr{R}_{M_iO_i}(\hat{\mathscr{R}}_{O_iO_i(-i)}^{(\alpha)-1} - \mathscr{R}_{O_iO_i}^{(\alpha)-1})\mathscr{R}_{O_iO_i}^{1/2}\|_2^2\} &\leqslant E(\|\mathscr{R}_{M_iO_i}\|_\infty^2 \|\hat{\mathscr{R}}_{O_iO_i(-i)}^{(\alpha)-1}\|_\infty^2 \\
&\quad \times \|\hat{\mathscr{R}}_{O_iO_i(-i)} - \mathscr{R}_{O_iO_i}\|_2^2 \|\mathscr{R}_{O_iO_i}^{(\alpha)-1}\mathscr{R}_{O_iO_i}^{1/2}\|_\infty^2) \\
&\leqslant \|\mathscr{R}_{M_iO_i}\|_\infty^2 \alpha^{-2} E(\|\hat{\mathscr{R}}_{O_iO_i(-i)} - \mathscr{R}_{O_iO_i}\|_2^2 \alpha^{-1}) \\
&= \alpha^{-3} O(n^{-1}).
\end{aligned}
$$

References

Antoniadis, A. and Sapatinas, T. (2003) Wavelet methods for continuous-time prediction using Hilbert-valued autoregressive processes. *J. Multiv. Anal.*, **87**, 133–158.

Aston, J. A. D. and Kirch, C. (2012) Detecting and estimating changes in dependent functional data. *J. Multiv. Anal.*, **109**, 204–220.

Benko, M., Härdle, W. and Kneip, A. (2009) Common functional principal components. *Ann. Statist.*, **37**, 1–34.

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

Bugni, F. A. (2012) Specification test for missing functional data. *Econmetr. Theor.*, **28**, 959–1002.

Cai, T. T. and Hall, P. (2006) Prediction in functional linear regression. *Ann. Statist.*, **34**, 2159–2179.

Cardot, H., Ferraty, F. and Sarda, P. (1999) Functional linear model. *Statist. Probab. Lett.*, **45**, 11–22.

Cardot, H., Mas, A. and Sarda, P. (2007) CLT in functional linear regression models. *Probab. Theor. Reltd Flds*, **138**, 325–361.

Dauxois, J., Pousse, A. and Romain, Y. (1982) Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *J. Multiv. Anal.*, **12**, 136–154.

Delaigle, A. and Hall, P. (2013) Classification using censored functional data. *J. Am. Statist. Ass.*, **108**, 1269–1283.

Didericksen, D., Kokoszka, P. and Zhang, X. (2012) Empirical properties of forecasts with the functional autoregressive model. *Computnl Statist.*, **27**, 285–298.

Ferraty, F. and Romain, Y. (eds) (2011) *The Oxford Handbook of Functional Data Analysis*. Oxford: Oxford University Press.

Ferraty, F. and Vieu, P. (2006) *Nonparametric Functional Data Analysis*. New York: Springer.

Goldberg, Y., Ritov, Y. and Mandelbaum, A. (2014) Predicting the continuation of a function with applications to call center data. *J. Statist. Planng Inf.*, **147**, 53–65.

Groetsch, C. W. (1993) *Inverse Problems in the Mathematical Sciences*. Braunschweig: Vieweg.
Hall, P. and Horowitz, J. L. (2007) Methodology and convergence rates for functional linear regression. *Ann. Statist.*, **35**, 70–91.
Hall, P. and Hosseini-Nasab, M. (2006) On properties of functional principal components analysis. *J. R. Statist. Soc.* B, **68**, 109–126.
Hall, P., Müller, H.-G. and Wang, J.-L. (2006) Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.*, **34**, 1493–1517.
He, G., Müller, H.-G. and Wang, J.-L. (2003) Functional canonical analysis for square integrable stochastic processes. *J. Multiv. Anal.*, **85**, 54–77.
He, G., Müller, H.-G., Wang, J.-L. and Yang, W. (2010) Functional linear regression via canonical analysis. *Bernoulli*, **16**, 705–729.
Horváth, L., Hušková, M. and Kokoszka, P. (2010) Testing the stability of the functional autoregressive process. *J. Multiv. Anal.*, **101**, 352–367.
Horváth, L. and Kokoszka, P. (2012) *Inference for Functional Data with Applications*. New York: Springer.
Horváth, L., Kokoszka, P. and Reeder, R. (2013) Estimation of the mean of functional time series and a two-sample problem. *J. R. Statist. Soc.* B, **75**, 103–122.
James, G. M. and Hastie, T. J. (2001) Functional linear discriminant analysis for irregularly sampled curves. *J. R. Statist. Soc.* B, **63**, 533–550.
James, G. M., Hastie, T. J. and Sugar, C. A. (2000) Principal component models for sparse functional data. *Biometrika*, **87**, 587–602.
Jarušková, D. (2013) Testing for a change in covariance operator. *J. Statist. Planng Inf.*, **143**, 1500–1511.
Jolliffe, I. T. (2002) *Principal Component Analysis*. New York: Springer.
Kallenberg, O. (2002) *Foundations of Modern Probability*. New York: Springer.
Kargin, V. and Onatski, A. (2008) Curve forecasting by functional autoregression. *J. Multiv. Anal.*, **99**, 2508–2526.
Kraus, D. and Panaretos, V. M. (2012) Dispersion operators and resistant second-order functional data analysis. *Biometrika*, **99**, 813–832.
Krzanowski, W. J. (2000) *Principles of Multivariate Analysis*. Oxford: Oxford University Press.
Liebl, D. (2013) Modeling and forecasting electricity spot prices: a functional data perspective. *Ann. Appl. Statist*, **7**, 1562–1592.
Liu, B. and Müller, H.-G. (2009) Estimating derivatives for samples of sparsely observed functions, with application to online auction dynamics. *J. Am. Statist. Ass.*, **104**, 704–717.
Mas, A. (2007) Testing for the mean of random curves: a penalization approach. *Statist. Inf. Stoch. Processes*, **10**, 147–163.
Müller, H.-G. and Stadtmüller, U. (2005) Generalized functional linear models. *Ann. Statist.*, **33**, 774–805.
Panaretos, V. M., Kraus, D. and Maddocks, J. H. (2010) Second-order comparison of Gaussian random functions and the geometry of DNA minicircles. *J. Am. Statist. Ass.*, **105**, 670–682.
Pruijm, M., Ponte, B., Ackermann, D., Vuistiner, P., Paccaud, F., Guessous, I., Ehret, G., Eisenberger, U., Mohaupt, M., Burnier, M., Martin, P.-Y. and Bochud, M. (2013) Heritability, determinants and reference values of renal length: a family-based population study. *Eur. Radiol.*, **23**, 2899–2905.
Ramsay, J. O., Hooker, G. and Graves, S. (2009) *Functional Data Analysis with R and MATLAB*. New York: Springer.
Ramsay, J. O. and Silverman, B. W. (2002) *Applied Functional Data Analysis*. New York: Springer.
Ramsay, J. O. and Silverman, B. W. (2005) *Functional Data Analysis*. New York: Springer.
Sangalli, L. M., Secchi, P., Vantini, S. and Veneziani, A. (2009) A case study in exploratory functional data analysis: geometrical features of the internal carotid artery. *J. Am. Statist. Ass.*, **104**, 37–48.
Yao, F., Müller, H.-G. and Wang, J.-L. (2005a) Functional data analysis for sparse longitudinal data. *J. Am. Statist. Ass.*, **100**, 577–590.
Yao, F., Müller, H.-G. and Wang, J.-L. (2005b) Functional linear regression analysis for longitudinal data. *Ann. Statist.*, **33**, 2873–2903.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary document: Components and completion of partially observed functional data'.

Supplementary document: Components and completion of partially observed functional data

David Kraus

Institute of Social and Preventive Medicine, University Hospital Lausanne, Switzerland

Summary. This supplementary document describes computational details of the proposed methods and provides proofs of Propositions 1, 2, 3 and 4.

1. Computation

1.1. Preliminary steps

In most applications, functional data are observed at discrete time points and are possibly subject to measurement error, so it is necessary to preprocess the raw data using smoothing techniques to obtain functions or their derivatives. In the context of partially observed functional data, the measurement time points are located only in observation periods O_i, while there are no measurements in missing periods M_i. We assume that the measurement points are dense in the observation periods, so that it is possible to apply smoothing techniques to obtain the functional values of the ith curve from the measured values of this curve. We use spline smoothing with a roughness penalty, as described in Ramsay and Silverman (2005, Chapter 5), but other methods like kernel smoothing can be used as well. In our experience, a simple approach works well: we apply the smoothing procedure to all values measured for the ith curve but use the computed smooth curve only for $t \in O_i$ (ignoring it on M_i where measurements are not available to make it reliable).

In practice, the observation and missing periods are typically not given (because they are not designed) and one needs to define them. For instance, one can define M_i to consist of the periods before the first and after the last measurement time and of all gaps between two consecutive measurement times that are larger than a certain threshold g. The value of g is the largest length of intervals without measurements over which we are willing to smooth. The choice of g depends on the particular setting; in general, if, for example, one considers K equidistant points in $[0, 1]$ (e.g., $K = 10$) as the minimum reliable design for smoothing on the whole domain $[0, 1]$, then $g = 1/K$ seems reasonable.

Sometimes, registration of functional data is needed. Shift registration (Ramsay and Silverman, 2005, Section 7.2) is easy to implement for incomplete functions: in the registration criterion the sample mean of partially observed functions is computed by the method described in the next subsection and the distance of each shifted curve from the sample mean is computed by numerical integration over the observed period of the curve; the criterion is minimised by the Procrustes method as usual. Methods based on warping can be modified similarly but further investigation of their performance is needed.

1.2. Principal component analysis, functional reconstruction

For practical computation we must use finite dimensional representations of functions and operators. Two traditional approaches exist: we can use either basis expansions or evaluation on a grid

of points. It is difficult to use the basis approach in our situation because incompletely observed functions are available on different subsets of the time domain. The grid approach is more suited for this type of data since it works directly with time arguments. Let $t_k = (k-0.5)/d$, $k = 1, \dots, d$ be a fine grid of equidistant points on which all functions and kernels of integral operators will be evaluated. Denote by $\mathbf{x}_i$ the d-dimensional vector of values of X_i at points t_k; this vector contains missing values on components corresponding to $t_k \in M_i$ while for $t_k \in O_i$, its values are obtained by evaluation of the spline representation of X_i. Denote by $\mathbf{X}$ the $(n \times d)$-dimensional data matrix with $\mathbf{x}_i$, $i = 1, \dots, n$ in rows.

The vector $\mathbf{m}$ of values of the mean function μ on the grid is estimated by $\hat{\mathbf{m}}$ equal to the vector of column means of $\mathbf{X}$ computed from available (not missing) data in each column. The covariance kernel ρ of the operator $\mathscr{R}$ evaluated on the grid corresponds to the $(d \times d)$-matrix $\mathbf{R}$ with entries $R_{kl} = \rho(t_k, t_l)$ and is estimated by the sample covariance matrix $\hat{\mathbf{R}}$ with entry $\hat{R}_{kl}$ computed from the data matrix $\mathbf{X}$ using all complete pairs of observations in columns k, l.

To estimate the eigenvalues and eigenfunctions, one performs eigen-decomposition of the matrix $\hat{\mathbf{R}}$. Denote $\Delta = 1/d$, the distance between the points of the grid. If the eigenvalues and eigenvectors of $\hat{\mathbf{R}}$ are $\hat{\kappa}_j$ and $\hat{\mathbf{u}}_j$, $j = 1, \dots, d$, then the eigenvalues of the operator $\hat{\mathscr{R}}$ are $\hat{\lambda}_j = \hat{\kappa}_j \Delta$ and the corresponding eigenfunctions $\hat{\varphi}_j$ evaluated on the grid are $\hat{\mathbf{f}}_j = \hat{\mathbf{u}}_j \Delta^{-1/2}$. The observed part $\hat{\beta}_{ijO_i} = \langle X_{iO_i} - \hat{\mu}_{O_i}, \hat{\varphi}_{jO_i} \rangle$ of the jth principal score of the ith curve is computed by numerical quadrature as $\hat{\beta}_{ijO_i} = \langle \mathbf{x}_{iO_i} - \hat{\mathbf{m}}_{O_i}, \hat{\mathbf{f}}_{jO_i} \rangle \Delta$, where the latter inner product is the usual inner product of vectors and the vectors with subscript O_i are subvectors of the original vectors consisting of elements with indices k such that $t_k \in O_i$.

Within the grid representation, the evaluation of an integral operator $\mathscr{B}$ in the sense of numerical integration corresponds to matrix multiplication: for a function h, $\mathscr{B}h$ is computed as $\mathbf{B}\mathbf{h}\Delta$, where the vector $\mathbf{h}$ and the matrix $\mathbf{B}$ are the values of h and of the kernel of $\mathscr{B}$ on the grid. From a purely computational point of view, even linear operators that have no integral representation may be represented by matrices. In particular, the identity operator $\mathscr{I}$ used in ridge regularisation is represented by the matrix $\mathbf{I}$ equal to the identity matrix divided by Δ; indeed, its value at $\mathbf{h}$ is $\mathbf{I}\mathbf{h}\Delta = \mathbf{h}$, thus it maps the argument on itself. The regularised operator $\hat{\mathscr{R}}^{(\alpha)}_{O_iO_i}$ is represented by the matrix $\hat{\mathbf{R}}^{(\alpha)}_{O_iO_i} = \hat{\mathbf{R}}_{O_iO_i} + \alpha \mathbf{I}_{O_i}$, where the subscript O_i denotes the submatrix corresponding to grid points in O_i. Analogously, the operators $\hat{\mathscr{R}}_{M_iM_i}$, $\hat{\mathscr{R}}_{M_iO_i}$ etc. are given by the corresponding submatrices of $\hat{\mathbf{R}}$. Then the matrix representation of the prediction operator $\hat{\mathscr{A}}^{(\alpha)}_i$ is computed as $\hat{\mathbf{A}}^{(\alpha)}_i = \hat{\mathbf{R}}_{O_iM_i} \hat{\mathbf{R}}^{(\alpha)-1}_{O_iO_i} \Delta^{-1}$. The regularised prediction of the missing part of the principal score and of the missing part of the trajectory can be computed as

$$\hat{\beta}^{(\alpha)}_{ij} = \langle \hat{\mathbf{A}}^{(\alpha)}_i (\mathbf{x}_{iO_i} - \hat{\mathbf{m}}_{O_i}) \Delta, \hat{\mathbf{f}}_{jM_i} \rangle \Delta, \quad \hat{\mathbf{x}}^{(\alpha)}_{iM_i} = \hat{\mathbf{A}}^{(\alpha)}_i (\mathbf{x}_{iO_i} - \hat{\mathbf{m}}_{O_i}) \Delta + \hat{\mathbf{m}}_{M_i}.$$

The covariance operator $\hat{\mathscr{V}}_i$ for the missing trajectory is obtained as

$$\hat{\mathbf{V}}_i = \hat{\mathbf{R}}_{M_iM_i} - \hat{\mathbf{A}}^{(\alpha)}_i \hat{\mathbf{R}}_{O_iO_i} \hat{\mathbf{A}}^{(\alpha)\mathrm{T}}_i \Delta^2$$

and the variance for the score is $\hat{v}^2_{ij} = \langle \hat{\mathbf{f}}_{jM_i}, \hat{\mathbf{V}}_i \hat{\mathbf{f}}_{jM_i} \rangle \Delta^2$.

The effective degrees of freedom can be computed directly using the series in (9) truncated at d terms, with the eigenvalues $\hat{\lambda}_{O_iO_ik}$ of $\hat{\mathscr{R}}_{O_iO_i}$ obtained from the eigenvalues of the matrix $\hat{\mathbf{R}}_{O_iO_i}$ like in the case of those of $\hat{\mathscr{R}}$ discussed above. Alternatively, one can use the matrix trace formula $\mathrm{trace}(\hat{\mathbf{R}}^{(\alpha)-1}_{O_iO_i} \hat{\mathbf{R}}_{O_iO_i} \Delta^{-1}) \Delta$. The computation of the residual sum of squares for scores

is straightforward; in the case of trajectories, the squared norms of functions are computed as the squared norms of vectors, multiplied by Δ.

The generalised cross-validation score can be minimised numerically by a Newton-type iterative procedure. In particular, we use the method "L-BFGS-B" available in the function *optim* in the R package (R Core Team, 2013). For the reliability of the optimisation procedure, we found it useful to scale the input parameters: the minimisation is run with $(\mathbf{x}_i - \mathbf{m})/s$ in place of $\mathbf{x}_i$ (and, consequently, with $\hat{\mathbf{R}}/s^2$ in place of $\hat{\mathbf{R}}$, $\hat{\lambda}_{O_iO_ij}/s^2$ in place of $\hat{\lambda}_{O_iO_ij}$ etc.); once the optimal value of α is found, it is multiplied by s^2 to return to the original scale and perform other computations with original data. The value $s^2 = \hat{\lambda}_{O_iO_i1}$ works well. The evaluation of the generalised cross-validation score can be unstable for very small values of α. Therefore, we run the minimisation routine with a lower limit for α, namely with $\alpha_0 = \max(\varepsilon^{1/2}, \alpha_*)$, where ε is the value of machine epsilon and α_* is such that the effective degrees of freedom equal $n/4$ (which is a reasonable upper bound for the number of free parameters). We initialise the iterative procedure with α equal to $\max(\bar{\lambda}_{O_iO_i}, \alpha_0)$ where $\bar{\lambda}_{O_iO_i}$ is the average of the eigenvalues $\hat{\lambda}_{O_iO_ij}$.

2. Proofs

2.1. Proof of Proposition 1

We use the notation $Z_i = X_i - \mu$.

For part (a), denote $\bar{\mu}(t) = J(t)\mu(t)$ and write

$$\mathsf{E}\,\|\hat{\mu} - \mu\|^2 \le \mathsf{E}(\|\hat{\mu} - \bar{\mu}\| + \|\bar{\mu} - \mu\|)^2 = \mathsf{E}\,\|\hat{\mu} - \bar{\mu}\|^2 + 2\,\mathsf{E}(\|\hat{\mu} - \bar{\mu}\|\|\bar{\mu} - \mu\|) + \mathsf{E}\,\|\bar{\mu} - \mu\|^2. \quad (1)$$

The first term on the right-hand side of (1) equals

$$\begin{aligned}\mathsf{E}\left\|\frac{J}{\sum_{i=1}^n O_i}\sum_{i=1}^n O_iZ_i\right\|^2 &= n^{-2}\int_0^1 \sum_{j=1}^n\sum_{k=1}^n \mathsf{E}\left(\frac{n^2J(t)}{(\sum_{i=1}^n O_i(t))^2}O_j(t)Z_j(t)O_k(t)Z_k(t)\right)dt \\ &= n^{-2}\int_0^1 \sum_{j=1}^n \mathsf{E}\left(\frac{n^2J(t)O_j(t)}{(\sum_{i=1}^n O_i(t))^2}\right)\mathsf{E}\,Z_j(t)^2dt,\end{aligned}$$

where the last equality follows from the independence of $(O_1, \dots, O_n)$ and $(Z_1, \dots, Z_n)$, and from the independence of Z_j and Z_k for $j \ne k$. Rewrite the first expectation in the integrand as

$$\mathsf{E}\left(\frac{n^2J(t)O_j(t)}{(\sum_{i=1}^n O_i(t))^2}1_{[n^{-1}\sum_{i=1}^n O_i(t) > \delta_1]}\right) + \mathsf{E}\left(\frac{n^2J(t)O_j(t)}{(\sum_{i=1}^n O_i(t))^2}1_{[n^{-1}\sum_{i=1}^n O_i(t) \le \delta_1]}\right).$$

For all $t \in [0,1]$, the first summand is bounded from above by δ_1^{-2} while the second summand is dominated by $n^2 \sup_{t\in[0,1]} P(n^{-1}\sum_{i=1}^n O_i(t) \le \delta_1)$. Hence we see that

$$\mathsf{E}\,\|\hat{\mu} - \bar{\mu}\|^2 \le n^{-1}\left\{\delta_1^{-2} + n^2 \sup_{t\in[0,1]} P\left(n^{-1}\sum_{i=1}^n O_i(t) \le \delta_1\right)\right\}\mathsf{E}\,\|Z_1\|^2 = O(n^{-1}).$$

For the last term in (1), we obtain

$$\int_0^1 \mathsf{E}(J(t) - 1)\mu(t)^2dt = \int_0^1 P\left(\sum_{i=1}^n O_i(t) = 0\right)\mu(t)^2dt$$

$$\leq \sup_{t\in[0,1]} P\biggl(n^{-1}\sum_{i=1}^n O_i(t) \leq \delta_1\biggr)\|\mu\|^2$$
$$= O(n^{-2}).$$

The second term on the right-hand side of (1) is dominated by $2(\mathsf{E}\,\|\hat{\mu}-\bar{\mu}\|^2)^{1/2}(\mathsf{E}\,\|\bar{\mu}-\mu\|^2)^{1/2} \leq O(n^{-1})$. Putting these results together completes the proof of part (a).

The proof of part (b) is similar. Rewrite

$$\hat{\mathscr{R}} - \mathscr{R} = (\hat{\mathscr{R}} - \check{\mathscr{R}}) + (\check{\mathscr{R}} - \bar{\mathscr{R}}) + (\bar{\mathscr{R}} - \mathscr{R}), \tag{2}$$

where $\check{\mathscr{R}}$ and $\bar{\mathscr{R}}$ are integral operators with kernels

$$\check{\rho}(s,t) = \frac{I(s,t)}{\sum_{i=1}^n U_i(s,t)} \sum_{i=1}^n U_i(s,t)Z_i(s)Z_i(t),$$

and $\bar{\rho}(s,t) = I(s,t)r(s,t)$. The first term on the right-hand side of (2) reflects the effect of estimation of the mean. By direct computation, we see that

$$\mathsf{E}\,\|\hat{\mathscr{R}} - \check{\mathscr{R}}\|_2^2 = \mathsf{E}\int_{[0,1]^2} I(s,t)\{\hat{\mu}_{st}(s)-\mu(s)\}^2\{\hat{\mu}_{st}(t)-\mu(t)\}^2 dsdt$$
$$= \mathsf{E}\int_{[0,1]^2} \frac{I(s,t)}{(\sum_{i=1}^n U_i(s,t))^4}\biggl(\sum_{i=1}^n U_i(s,t)Z_i(s)\biggr)^2\biggl(\sum_{i=1}^n U_i(s,t)Z_i(t)\biggr)^2 dsdt.$$

Developing the sums in the integrand and using the independence of the functions and observation indicators and the Cauchy–Schwarz inequality, we can show that the above quantity is dominated by

$$n^{-2}\int_{[0,1]^2} \mathsf{E}\biggl(\frac{n^2 I(s,t)}{(\sum_{i=1}^n U_i(s,t))^2}\biggr)\{(\mathsf{E}\,Z_1(s)^4\,\mathsf{E}\,Z_1(t)^4)^{1/2} + \rho(s,t)^2\}dsdt \leq O(n^{-2}),$$

where the last inequality is due to the fact that the first expectation in the integrand is bounded by $\delta_2^{-2} + n^2 \sup_{(s,t)\in[0,1]^2} P(n^{-1}\sum_{i=1}^n U_i(s,t) \leq \delta_2)$, which can be shown by manipulations similar to those in part (a). Next, analogously to part (a) we obtain for the second and third term on the right-hand side of (2) that

$$\mathsf{E}\,\|\check{\mathscr{R}} - \bar{\mathscr{R}}\|_2^2 \leq n^{-1}\biggl\{\delta_2^{-2} + n^2 \sup_{(s,t)\in[0,1]^2} P\biggl(n^{-1}\sum_{i=1}^n U_i(s,t) \leq \delta_2\biggr)\biggr\}\,\mathsf{E}\,\|Z_1\otimes Z_1 - \mathscr{R}\|_2^2$$
$$= O(n^{-1})$$

(here $\otimes$ denotes the tensor product) and $\mathsf{E}\,\|\bar{\mathscr{R}} - \mathscr{R}\|_2^2 \leq O(n^{-2})$. Combining these bounds we obtain the assertion of part (b).

2.2. Proof of Proposition 2

Lemma 4.2 of Bosq (2000) and the inequality between the operator norm and Hilbert–Schmidt norm yield that $|\hat{\lambda}_j - \lambda_j| \leq \|\hat{\mathscr{R}} - \mathscr{R}\|_\infty \leq \|\hat{\mathscr{R}} - \mathscr{R}\|_2$ for all j. The first result then follows from part (b) of Proposition 1. For the second part, Lemma 4.3 of Bosq (2000) gives the inequality

$\|\hat{\varphi}_j - \hat{s}_j\varphi_j\| \leq a_j\|\hat{\mathscr{R}} - \mathscr{R}\|_\infty$, where a_j is a constant depending on the eigenvalue spacings. Note that this lemma is formulated in Bosq (2000) for fully observed functions but an inspection of the proof shows that the inequality holds for any two compact linear operators in place of $\hat{\mathscr{R}}, \mathscr{R}$. This inequality, the dominance of the Hilbert–Schmidt norm over the operator norm and part (b) of Proposition 1 complete the proof.

2.3. Proof of Proposition 3

Rewrite

$$\hat{\beta}_{ijM_i}^{(\alpha_n)} - \beta_{ijM_i} = (\hat{\beta}_{ijM_i}^{(\alpha_n)} - \tilde{\beta}_{ijM_i}) + (\tilde{\beta}_{ijM_i} - \beta_{ijM_i})$$

and use Theorem 1 to obtain the first part of the proposition. Compute

$$v_{ij}^2 = \operatorname{var}(\tilde{\beta}_{ijM_i} - \beta_{ijM_i}) = \langle \varphi_{jM_i}, \mathscr{R}_{M_iM_i}\varphi_{jM_i}\rangle - \langle \varphi_{jM_i}, \mathscr{R}_{M_iO_i}\mathscr{R}_{O_iO_i}^{-1}\mathscr{R}_{O_iM_i}\varphi_{jM_i}\rangle.$$

The convergence in probability of $\langle \hat{\varphi}_{jM_i}, \hat{\mathscr{R}}_{M_iM_i}\hat{\varphi}_{jM_i}\rangle$ to $\langle \varphi_{jM_i}, \mathscr{R}_{M_iM_i}\varphi_{jM_i}\rangle$ is a direct consequence of Propositions 1 and 2. The last term in the expression for v_{ij}^2 and the corresponding term in the estimator $\hat{v}_{ij}^2$ equal $\langle \tilde{a}_{ij}, \mathscr{R}_{O_iO_i}\tilde{a}_{ij}\rangle$, $\langle \hat{a}_{ij}^{(\alpha_n)}, \hat{\mathscr{R}}_{O_iO_i}\hat{a}_{ij}^{(\alpha_n)}\rangle$, respectively. In their difference

$$\langle \hat{a}_{ij}^{(\alpha_n)}, (\hat{\mathscr{R}}_{O_iO_i} - \mathscr{R}_{O_iO_i})\hat{a}_{ij}^{(\alpha_n)}\rangle + (\langle \hat{a}_{ij}^{(\alpha_n)}, \mathscr{R}_{O_iO_i}\hat{a}_{ij}^{(\alpha_n)}\rangle - \langle \tilde{a}_{ij}, \mathscr{R}_{O_iO_i}\tilde{a}_{ij}\rangle),$$

the convergence of the second term to zero was shown in the proof of Theorem 1. For the first term we compute

$$\begin{aligned} |\langle \hat{a}_{ij}^{(\alpha_n)}, (\hat{\mathscr{R}}_{O_iO_i} - \mathscr{R}_{O_iO_i})\hat{a}_{ij}^{(\alpha_n)}\rangle| &\leq \|\hat{\mathscr{R}}_{O_iO_i} - \mathscr{R}_{O_iO_i}\|_\infty \|\hat{a}_{ij}^{(\alpha_n)}\|^2 \\ &\leq O_P(n^{-1/2})\alpha_n^{-2}\|\hat{\mathscr{R}}_{O_iM_i}\|_\infty^2 \\ &\to 0. \end{aligned}$$

This completes the proof of the consistency of $\hat{v}_{ij}^2$. The remaining assertions are obvious.

2.4. Proof of Proposition 4

We can rewrite $\hat{X}_{iM_i}^{(\alpha_n)} - X_{iM_i} = (\hat{X}_{iM_i}^{(\alpha_n)} - \tilde{X}_{iM_i}) + (\tilde{X}_{iM_i} - X_{iM_i})$. Due to Theorem 2, the L^2-norm of the first term on the right-hand side converges to 0 in probability. The second term is the limiting stochastic process. The consistency of the covariance estimator can be proven like in the proof of Proposition 3. The assertion for the Gaussian case follows immediately from the fact that the limiting process is a linear function of X_i.

References

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

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer, New York.

D. Classification of functional fragments by regularized linear classifiers with domain selection

By David Kraus and Marco Stefanucci

Biometrika, 106(1):161–180, 2019

DOI: 10.1093/biomet/asy060

Biometrika (2019), **106**, 1, *pp.* 161–180
Printed in Great Britain
doi: 10.1093/biomet/asy060
Advance Access publication 17 December 2018

Classification of functional fragments by regularized linear classifiers with domain selection

By DAVID KRAUS

Department of Mathematics and Statistics, Masaryk University, Kotlářská 2, 611 37 Brno, Czech Republic
david.kraus@mail.muni.cz

and MARCO STEFANUCCI

Department of Statistical Sciences, Sapienza University of Rome, Piazzale Aldo Moro 5, 00185 Roma, Italy
marco.stefanucci@uniroma1.it

Summary

We consider classification of functional data into two groups by linear classifiers based on one-dimensional projections of functions. We reformulate the task of finding the best classifier as an optimization problem and solve it by the conjugate gradient method with early stopping, the principal component method, and the ridge method. We study the empirical version with finite training samples consisting of incomplete functions observed on different subsets of the domain and show that the optimal, possibly zero, misclassification probability can be achieved in the limit along a possibly nonconvergent empirical regularization path. We propose a domain extension and selection procedure that finds the best domain beyond the common observation domain of all curves. In a simulation study we compare the different regularization methods and investigate the performance of domain selection. Our method is illustrated on a medical dataset, where we observe a substantial improvement of classification accuracy due to domain extension.

Some key words: Classification; Conjugate gradient; Domain selection; Functional data; Partial observation; Regularization; Ridge method.

1. Introduction

We consider classification of a functional observation into one of two groups. Classification of functional data is a rich, longstanding topic and is comprehensively surveyed in Baíllo et al. (2011b). Delaigle & Hall (2012a) showed that depending on the relative geometric positions of the difference of the group means, representing the signal, and the covariance operator, summarizing the structure of the noise, certain classifiers can have zero misclassification probability. This remarkable phenomenon, called perfect classification, is a special property of the infinite-dimensional setting and cannot occur in the multivariate context, except in degenerate cases. Delaigle & Hall (2012a) showed that a particularly simple class of linear classifiers, based on a carefully chosen one-dimensional projection of the function to be classified, can achieve this optimal error rate either exactly or in the limit along a sequence of approximations. Berrendero et al. (2018) further elucidated the perfect classification phenomenon from the point

of view of the Feldman–Hájek dichotomy between mutual singularity and absolute continuity of two Gaussian measures on abstract spaces with respect to each other.

Motivated by these findings, we reformulate the problem of determining the best classifier as a quadratic optimization problem on a function space or, equivalently, a linear inverse problem. These problems are ill-posed; however, unlike with most inverse problems, this is not a complication but rather an advantage in the sense that the more ill-posed the problem is, the better the optimal misclassification probability. We use regularization techniques, such as the method of conjugate gradients with early stopping and ridge regularization, to solve the optimization problem, obtaining a class of regularized linear classifiers. The optimal misclassification rate is the limit along the regularization path of solutions which themselves may not converge.

We study the empirical version of the problem, where the objective function in the constrained minimization must be estimated from finite training data, and make two contributions. First, we show that it is possible to construct an empirical regularization path towards the possibly nonexistent unconstrained solution such that the classification error converges to its best value, possibly zero. We do this for conjugate gradient, principal component and ridge classification in a truly infinite-dimensional manner, in the sense that the convergence takes place along a path with decreasing regularization and holds without restrictions on the mean difference between classes. Second, all our methods and theory are developed in the setting of partially observed functional data, where trajectories are observed only on subsets of the domain. This type of incomplete data, also called functional fragments, is increasingly common in applications; see, for example, Bugni (2012), Delaigle & Hall (2013), Liebl (2013), Goldberg et al. (2014), Kraus (2015), Delaigle & Hall (2016) and Gromenko et al. (2017). The principal difficulty for inference with fragments is that temporal averaging is precluded by the incompleteness of the observed functions. Our formulation as an optimization problem enables us to overcome this issue under certain assumptions, because only averaging across individuals in the training data is needed, and not individual curves.

Since the observation domains may vary in the training sample and the new curve to be classified may be observed on a different subset, it is natural to ask which domain should be used. We propose a domain selection strategy that looks for the best classifier with domain ranging from a minimum common domain to the entire domain of the function to be classified. For various methods of selecting the best observation points, see Ferraty et al. (2010), Delaigle et al. (2012), Pini & Vantini (2016), Berrendero et al. (2018) and Stefanucci et al. (2018).

Our simulation study confirms that domain selection can considerably reduce the misclassification rate. Further simulations compare the performances of the three types of regularization. Among other findings, this study shows that the principal component and conjugate gradient classifiers often achieve comparable error rates but that the latter usually needs a lower dimension of the regularization subspace, in agreement with a theoretical result we provide.

Application to a dataset on the geometric features of the internal carotid artery in patients with and without aneurysm demonstrates the utility of our proposed approach. These data consist of trajectories observed on intervals of different lengths. Previous analyses of the data used the common domain of all curves in classification. With our results we can include information beyond this minimum domain, which leads to a substantial drop in the error rate of discrimination between risk groups.

General references on functional data analysis include Ramsay & Silverman (2005) and Horváth & Kokoszka (2012). Further relevant references are Cuesta-Albertos et al. (2007) for other methods based on one-dimensional projections, Berrendero et al. (2016) for variable selection in classification, Bongiorno & Goia (2016) and Dai et al. (2017) for classification beyond the Gaussian setting, and Cuevas (2014) for an overview.

2. Regularized linear classification

2.1. *Projection classifiers*

We regard functional observations as random elements of the separable Hilbert space $L^2(\mathcal{I})$ of square-integrable functions on a compact domain $\mathcal{I}$ equipped with inner product $\langle f,g\rangle = \int_{\mathcal{I}} f(t)g(t)\,\mathrm{d}t$ and norm $\|f\| = \langle f,f\rangle^{1/2}$. In most applications $\mathcal{I}$ is an interval and the observations are curves, but our results can be extended to other objects, such as surfaces or images. We consider classification of a Gaussian random function, X, into one of two groups of Gaussian random functions: group 0 has mean μ_0; group 1 has mean μ_1. Both groups have covariance operator $\mathscr{R}$ defined as the integral operator

$$(\mathscr{R}f)(\cdot) = \int_{\mathcal{I}} \rho(\cdot\,,t)f(t)\,\mathrm{d}t$$

with kernel $\rho(s,t) = \mathrm{cov}\{X(s),X(t)\}$. In this section we assume that μ_0, μ_1 and $\mathscr{R}$ are known, which corresponds to the asymptotic situation with an infinite training sample. To simplify the presentation we assume throughout the paper that the new observation to be classified may come from either of the two classes with equal prior probability. The general case is treated in the Supplementary Material.

Like Delaigle & Hall (2012a) we consider the class of centroid classifiers that are based on one-dimensional projections of the form $\langle X,\psi\rangle$, where ψ is a function in $L^2(\mathcal{I})$. If X belongs to group j ($j = 0,1$), the distribution of $\langle X,\psi\rangle$ is normal with mean $\langle \mu_j,\psi\rangle$ and variance $\langle \psi,\mathscr{R}\psi\rangle$. Denote the corresponding Gaussian densities by $f_{\psi,j}$. The optimal classifier based on $\langle X,\psi\rangle$ assigns X to the class $C_\psi(X)$ given by

$$C_\psi(X) = 1_{\{f_{\psi,1}(\langle X,\psi\rangle)/f_{\psi,0}(\langle X,\psi\rangle)>1\}} = 1_{\{\langle X-\mu_0,\psi\rangle^2-\langle X-\mu_1,\psi\rangle^2>0\}} = 1_{\{T_\psi(X)>0\}},$$

where $T_\psi(X) = \langle X-\bar\mu,\psi\rangle\langle \mu,\psi\rangle$ with $\bar\mu = (\mu_0+\mu_1)/2$ and $\mu = \mu_1-\mu_0$. The misclassification probability of this classifier is

$$\begin{aligned} D(\psi) &= P_0\{C_\psi(X)=1\}/2 + P_1\{C_\psi(X)=0\}/2 = P_0(\langle X-\bar\mu,\psi\rangle\langle \mu,\psi\rangle > 0) \\ &= P_0(\langle X-\mu_0,\psi\rangle > |\langle \mu,\psi\rangle|/2) = 1-\Phi\left(\frac{|\langle \mu,\psi\rangle|}{2\langle \psi,\mathscr{R}\psi\rangle^{1/2}}\right), \end{aligned} \tag{1}$$

where P_j is the distribution of curves in group j and Φ is the standard normal cumulative distribution function.

To find the best function ψ, one would ideally like to maximize $|Z(\psi)|$, where

$$Z(\psi) = \frac{\langle \mu,\psi\rangle}{\langle \psi,\mathscr{R}\psi\rangle^{1/2}}.$$

Similarly to Delaigle & Hall (2012a) and Berrendero et al. (2018), we see that if $\|\mathscr{R}^{-1/2}\mu\| < \infty$, then by the Cauchy–Schwarz inequality,

$$\frac{|\langle \mu,\psi\rangle|}{\langle \psi,\mathscr{R}\psi\rangle^{1/2}} = \frac{|\langle \mathscr{R}^{-1/2}\mu,\mathscr{R}^{1/2}\psi\rangle|}{\langle \psi,\mathscr{R}\psi\rangle^{1/2}} \leqslant \frac{\|\mathscr{R}^{-1/2}\mu\|\|\mathscr{R}^{1/2}\psi\|}{\langle \psi,\mathscr{R}\psi\rangle^{1/2}} = \|\mathscr{R}^{-1/2}\mu\|. \tag{2}$$

If, moreover, $\|\mathscr{R}^{-1}\mu\| < \infty$, then the equality is achieved for $\psi = \mathscr{R}^{-1}\mu$. For this choice of ψ, or any multiple of it, the probability of misclassification is $1-\Phi(\|\mathscr{R}^{-1/2}\mu\|/2)$, which is positive due

to the finiteness of $\|\mathscr{R}^{-1/2}\mu\|$, which can be seen as the signal-to-noise ratio. If $\|\mathscr{R}^{-1/2}\mu\| < \infty$, then regardless of whether $\|\mathscr{R}^{-1}\mu\| < \infty$ or not, two Gaussian measures with mean difference μ and covariances $\mathscr{R}$ are mutually absolutely continuous and $1-\Phi(\|\mathscr{R}^{-1/2}\mu\|/2)$ is the Bayes error for distinguishing them, i.e., the lowest possible misclassification probability for this problem among all possible classifiers (Berrendero et al., 2018). If $\|\mathscr{R}^{-1/2}\mu\| < \infty$ but $\|\mathscr{R}^{-1}\mu\| = \infty$, then the Bayes risk cannot be achieved by a projection classifier based on a bounded linear functional of the form $\langle X, \psi\rangle$ for some $\psi \in L^2(\mathcal{I})$. One can, however, use the theory of reproducing kernel Hilbert spaces to define a linear classifier that achieves the Bayes risk. We do not pursue this line of development here because, as will be seen in § 2.2, approximations in the form of projections can asymptotically achieve the Bayes risk.

The maximization of $|Z(\psi)|$ can be solved as the task of maximizing $\langle \mu, \psi\rangle$ subject to $\langle \psi, \mathscr{R}\psi\rangle = 1$. Using Lagrange multipliers $\langle \mu, \psi\rangle + \lambda(1 - \langle \psi, \mathscr{R}\psi\rangle)$ and taking the Fréchet derivative with respect to ψ, one obtains the equation $2\lambda\mathscr{R}\psi = \mu$. Solutions for all $\lambda > 0$, if they exist, i.e., if $\|\mathscr{R}^{-1}\mu\| < \infty$, yield the same optimal misclassification probability. Without loss of generality we take $\lambda = 1/2$. Thus, minimizing the error rate translates into the unconstrained quadratic optimization problem to maximize $\langle \mu, \psi\rangle - \langle \psi, \mathscr{R}\psi\rangle/2$, or

$$\text{minimize } \langle \psi, \mathscr{R}\psi\rangle/2 - \langle \mu, \psi\rangle, \tag{3}$$

i.e., into the linear problem $\mathscr{R}\psi = \mu$.

2.2. *Regularization*

If $\psi = \mathscr{R}^{-1}\mu$ does not exist in $L^2(\mathcal{I})$, i.e., $\|\mathscr{R}^{-1}\mu\| = \infty$, there is no maximizer of $|Z(\psi)|$. One can instead consider an approximating, regularized problem that can be solved. Regularization is typically used to solve, in a stable way, ill-posed inverse problems for which a solution exists. In such contexts, the path of regularized solutions converges to the solution to the problem of interest. Here it may be that no solution exists, but paths of regularized solutions towards the possibly nonexistent solution still turn out to be useful, since the misclassification probability converges to the optimal value along these paths.

If a solution exists, one can approximate it by an iterative numerical method. This approach can also be used when no solution exists. The idea is to construct a sequence of iterations of an appropriate numerical optimization method. The number of steps taken along this divergent sequence towards the nonexistent solution can be seen as a regularization parameter. The conjugate gradient method is particularly suitable for this situation.

The first m steps of the conjugate gradient method applied to the linear inverse problem $\mathscr{R}\psi = \mu$, or equivalently to the minimization of the quadratic functional $\langle \psi, \mathscr{R}\psi\rangle/2 - \langle \mu, \psi\rangle$, are described in Algorithm 1. This formulation is based on the multivariate version in Phatak & de Hoog (2002, § 5), where one can find further references and details on how applying the conjugate gradient method to the normal equations in linear regression leads to partial least squares regression. The functions v_j are conjugate directions in the sense that $\langle v_j, \mathscr{R}v_k\rangle = 0$ for $j \neq k$, and the functions ζ_j are called residuals in numerical analysis and are orthogonal, i.e., $\langle \zeta_j, \zeta_k\rangle = 0$ for $j \neq k$. In step j, the algorithm moves from the current approximate solution $\hat{\psi}_j^{\mathrm{CG}}$ along the conjugate direction v_j with step length h_j that minimizes the quadratic objective. The residual is then updated to ζ_{j+1}. The new conjugate direction v_{j+1} is obtained by projecting the residual ζ_{j+1} onto the orthogonal complement of the span of the previous conjugate directions, where orthogonality is in the sense of the inner product $\langle \cdot, \mathscr{R}(\cdot)\rangle$.

Algorithm 1. Conjugate gradient regularized classification direction.

Initialize $\psi_0^{\mathrm{CG}} = 0$, $\nu_0 = \zeta_0 = \mu$
Repeat for $j = 0, \ldots, m-1$
$h_j = \langle \nu_j, \zeta_j \rangle / \langle \nu_j, \mathscr{R}\nu_j \rangle$
$\psi_{j+1}^{\mathrm{CG}} = \psi_j^{\mathrm{CG}} + h_j \nu_j$
$\zeta_{j+1} = \mu - \mathscr{R}\psi_{j+1}^{\mathrm{CG}}$ $(= \zeta_j - h_j \mathscr{R}\nu_j)$
$g_j = -\langle \zeta_{j+1}, \mathscr{R}\nu_j \rangle / \langle \nu_j, \mathscr{R}\nu_j \rangle$
$\nu_{j+1} = \zeta_{j+1} + g_j \nu_j$
Output ψ_m^{CG}

The conjugate gradient approach is an example of dimension reduction regularization. The method solves the minimization problem (3) with ψ restricted to the Krylov subspace $K_m(\mathscr{R}, \mu)$ spanned by $\mu, \mathscr{R}\mu, \ldots, \mathscr{R}^{m-1}\mu$ and also by the first m conjugate directions ν_j or the first m residuals ζ_j; that is, it seeks to minimize $\langle \psi, \mathscr{R}\psi \rangle / 2 - \langle \mu, \psi \rangle$ subject to $\psi \in K_m(\mathscr{R}, \mu)$. The projection direction that solves this minimization is ψ_m^{CG}.

Another popular choice is to minimize $\langle \psi, \mathscr{R}\psi \rangle / 2 - \langle \mu, \psi \rangle$ subject to $\psi \in E_m(\mathscr{R})$, where $E_m(\mathscr{R})$ is the subspace spanned by the first m eigenfunctions, $\varphi_1, \ldots, \varphi_m$, of $\mathscr{R}$ in the spectral decomposition

$$\mathscr{R} = \sum_{j=1}^{\infty} \lambda_j \varphi_j \otimes \varphi_j,$$

with $\lambda_1 \geqslant \lambda_2 \geqslant \cdots > 0$ being the eigenvalues. The solution $\psi_m^{\mathrm{PC}} = \sum_{j=1}^{m} \lambda_j^{-1} \langle \mu, \varphi_j \rangle \varphi_j$ gives the principal component classifier of Delaigle & Hall (2012a).

In general one can minimize $\langle \psi, \mathscr{R}\psi \rangle / 2 - \langle \mu, \psi \rangle$ subject to $\psi \in S_m$, where S_m is the m-dimensional subspace generated by some functions $s_1, \ldots, s_m$ such that the s_j $(j = 1, 2, \ldots)$ generate the range of $\mathscr{R}$. Let $\mathscr{P}_m$ be the projection operator that projects onto S_m, and let $\mathscr{R}_m = \mathscr{P}_m \mathscr{R} \mathscr{P}_m$ and $\mathscr{R}_m^- = \mathscr{P}_m \mathscr{R}^{-1} \mathscr{P}_m$. Then the solution of the regularized minimization problem is $\psi_m = \mathscr{R}_m^- \mu$. More explicitly, considering solutions of the form $\psi_m = \sum_{j=1}^{m} c_j s_j$ leads to the m-variate minimization of $c^{\mathrm{T}} Q c / 2 - u^{\mathrm{T}} c$ where the matrix Q is such that $Q_{jk} = \langle s_j, \mathscr{R} s_k \rangle$ and the vector u has components $u_j = \langle \mu, s_j \rangle$, i.e., to the solution with coefficients $c = Q^{-1} u$. In the case of the Krylov subspace, the iterative conjugate gradient method given in Algorithm 1 is, however, preferred because the matrix Q is ill-conditioned.

We can also take another approach to regularization, based on ridge regression. Optimizing the misclassification probability in a ball with radius $\theta^{1/2}$ leads to the task of minimizing $\langle \psi, \mathscr{R}\psi \rangle / 2 - \langle \mu, \psi \rangle$ subject to $\|\psi\|^2 \leqslant \theta$ or, equivalently, minimizing $\langle \psi, \mathscr{R}\psi \rangle / 2 - \langle \mu, \psi \rangle + \alpha \|\psi\|^2 / 2$, where $\alpha \geqslant 0$ is a regularization parameter. The solution is $\psi_\alpha^{\mathrm{R}} = \mathscr{R}_\alpha^{-1} \mu$, where $\mathscr{R}_\alpha = \mathscr{R} + \alpha \mathscr{I}$ and $\mathscr{I}$ denotes the identity operator. Despite its practical performance and amenability to theoretical analysis, the functional ridge classifier does not seem to have been considered before.

There is an important difference between the conjugate gradient method and the other approaches. While the principal component and ridge methods regularize the problem without the main goal in mind, the conjugate gradient approach greedily follows the goal of optimal classification. Indeed, the conjugate gradient method as an iterative optimization procedure constructs the regularization path focusing on the minimization of the misclassification probability, whereas the other approaches regularize by modifying the operator to be inverted regardless of the goal.

From a computational point of view the conjugate gradient method is simplest because it does not require inversion or eigendecomposition.

2.3. *Properties of regularization paths*

While ψ_m, the solution regularized by a subspace constraint, in general need not converge as $m \to \infty$ since a solution to the unconstrained minimization problem may not exist, the misclassification probability associated with the linear classifier given by ψ_m converges along the regularization path. The following and all other results are proved in the Appendix.

PROPOSITION 1. *The misclassification probability of the regularized linear classifier based on* $\psi_m = \mathscr{R}_m^- \mu$ *converges to* $1 - \Phi(\|\mathscr{R}^{-1/2}\mu\|/2)$ *as* $m \to \infty$.

This result holds regardless of whether the unconstrained minimization problem (3) has a solution, i.e., regardless of whether $\|\mathscr{R}^{-1}\mu\| < \infty$. The limiting misclassification probability is positive if $\|\mathscr{R}^{-1/2}\mu\| < \infty$ or zero if $\|\mathscr{R}^{-1/2}\mu\| = \infty$. As discussed earlier, the optimal error is achieved exactly by the one-dimensional projection onto $\psi = \mathscr{R}^{-1}\mu$, when $\|\mathscr{R}^{-1}\mu\| < \infty$. Even when $\|\mathscr{R}^{-1}\mu\| = \infty$, both of the dimension reduction techniques, namely the conjugate gradient and principal component methods, and also ridge regularization as we will soon see, achieve the optimal limiting error rate along a possibly nonconvergent path of one-dimensional projection directions.

It is natural to investigate and compare how quickly the misclassification rate approaches the limit for the two main types of subspace regularization. It turns out that the conjugate gradient classifier, being a greedy, goal-oriented procedure, performs as well as or better than the principal component classifier with the same dimension.

PROPOSITION 2. *Regardless of whether the optimal misclassification probability can be achieved exactly or along a regularization path, i.e., whether* $\|\mathscr{R}^{-1}\mu\| < \infty$ *or* $\|\mathscr{R}^{-1}\mu\| = \infty$, *and regardless of whether the optimal misclassification probability is zero or positive, i.e., whether* $\|\mathscr{R}^{-1/2}\mu\| = \infty$ *or* $\|\mathscr{R}^{-1/2}\mu\| < \infty$, *the misclassification probability of the principal component classifier using m components is higher than or equal to the misclassification probability of the m-step conjugate gradient classifier.*

Phatak & de Hoog (2002, § 6.2) showed in the multivariate setting that 'PLS fits closer than PCR'. In infinite dimensions, in the context of kernel partial least squares, Blanchard & Krämer (2010, Theorem 1) showed that the partial least squares solution is closer to the true solution of the inverse problem than is the principal component solution with the same number of components. Unlike these results, our Proposition 2 does not assume the existence of a solution and instead focuses on the values of the misclassification probability.

Although Proposition 2 suggests that the conjugate gradient method will typically use fewer components than the principal component method to achieve the best result, the resulting misclassification probability with the best number of components need not be better. We address this in the simulation study. A similar phenomenon was previously studied in the literature on partial least squares in finite dimensions and in the functional setting by Febrero-Bande et al. (2017).

As in the case of subspace regularization, below we obtain the convergence of the error probability of the ridge classifier, whether or not the unconstrained minimization problem (3) has a solution, i.e., regardless of whether $\|\mathscr{R}^{-1}\mu\| < \infty$. The limiting misclassification probability is positive if $\|\mathscr{R}^{-1/2}\mu\| < \infty$ or zero if $\|\mathscr{R}^{-1/2}\mu\| = \infty$.

PROPOSITION 3. *The misclassification probability of the regularized linear classifier based on* $\psi_\alpha^{\mathrm{R}} = \mathscr{R}_\alpha^{-1}\mu$ *converges to* $1 - \Phi(\|\mathscr{R}^{-1/2}\mu\|/2)$ *as* $\alpha \to 0+$.

3. EMPIRICAL CLASSIFIERS FOR FRAGMENTARY FUNCTIONS

3.1. *Construction of classifiers with incomplete training samples*

So far we have assumed that the parameters of each group are known. We now present the empirical version with a finite training dataset, and show that under regularity conditions such classifiers can achieve asymptotically the same optimal error rate as if there were infinite training data. We aim to do this not only in the case of fully observed functions but also in the case of incomplete curves. Incompleteness can occur in the training data, with each curve possibly observed on a different domain, as well as in the new curve that we wish to classify. One strategy would be to consider all curves on the intersection of their observation domains, if it is nonempty. However, such a restriction can be too severe and is unnecessary. We will construct classifiers that use the observed new curve on a set $\mathcal{I}$, which may be its entire observation set or a subset thereof, without requiring that all training curves be completely observed on $\mathcal{I}$.

For group j let there be a training sample consisting of n_j curves, $X_{j1}, \ldots, X_{jn_j}$. The training data are assumed to be mutually independent. Curves may be observed incompletely, with values known only on a subset O_{ji} of the domain and with no information about the values on the complement. The observation domains are assumed to be independent of the curves and consist of a finite union of intervals. We let $O_{ji}(t)$ denote the indicator of the curve X_{ji} being observed at time t. Similarly, let $U_{ji}(s,t)$ indicate observation at times s and t, i.e., $U_{ji}(s,t) = O_{ji}(s)O_{ji}(t)$.

The mean μ_j of group j can be estimated by the cross-sectional average

$$\hat{\mu}_j(t) = \frac{1_{\{N_j(t)>0\}}}{N_j(t)} \sum_{i=1}^{n_j} O_{ji}(t)X_{ji}(t) \quad (j = 0, 1),$$

where $N_j(t) = \sum_{i=1}^{n_j} O_{ji}(t)$ is the total number of observed curves in group j at time t. The covariance kernel $\rho(s,t)$ can be estimated by the empirical covariance using pairwise complete observations of groupwise centred curves. Formally, the estimator is

$$\hat{\rho}(s,t) = \frac{M_1(s,t)\hat{\rho}_1(s,t) + M_2(s,t)\hat{\rho}_2(s,t)}{M_1(s,t) + M_2(s,t)},$$

where $M_j(s,t) = \sum_{i=1}^{n_j} U_{ji}(s,t)$ and

$$\hat{\rho}_j(s,t) = \frac{1_{\{M_j(s,t)>0\}}}{M_j(s,t)} \sum_{i=1}^{n_j} U_{ji}(s,t)\{X_{ji}(s) - \hat{\mu}_{jst}(s)\}\{X_{ji}(t) - \hat{\mu}_{jst}(t)\}$$

with $\hat{\mu}_{jst}(s) = 1_{\{M_j(s,t)>0\}} M_j(s,t)^{-1} \sum_{i=1}^{n_j} U_{ji}(s,t)X_{ji}(s)$. If $N_j(t) = 0$ or $M_j(s,t) = 0$, the estimators are defined as $\hat{\mu}_j(t) = 0$ or $\hat{\rho}_j(s,t) = 0$, respectively. This happens with asymptotically vanishing probability under Assumption 1 below.

Suppose that the new independent curve to be classified, X_{new}, is observed on the domain O_{new}. Let us fix the target domain $\mathcal{I} \subseteq O_{\mathrm{new}}$ on which we aim to apply the classifier to X_{new}. The empirical classifier $\hat{C}_{\hat{\psi}}$ trained on partially observed curves is defined like the theoretical one, with unknown quantities replaced by their estimators. It assigns X_{new} restricted to $\mathcal{I}$ to the class

$\hat{C}_{\hat{\psi}}(X_{\text{new}}) = 1_{\{\hat{T}_{\hat{\psi}}(X_{\text{new}})>0\}}$, where $\hat{T}_{\hat{\psi}}(X_{\text{new}}) = \langle X_{\text{new}} - \tilde{\mu}, \hat{\psi}\rangle\langle \hat{\mu}, \hat{\psi}\rangle$. Here $\tilde{\mu} = (\hat{\mu}_0 + \hat{\mu}_1)/2$ and $\hat{\mu} = \hat{\mu}_1 - \hat{\mu}_0$, with $\hat{\mu}_j$ being the estimators defined above restricted to $\mathcal{I}$. The projection direction $\hat{\psi}$ is one of $\hat{\psi}_m^{\text{CG}}$, $\hat{\psi}_m^{\text{PC}}$ or $\hat{\psi}_\alpha^{\text{R}}$, constructed respectively by conjugate gradient, principal component or ridge regularization applied to $\hat{\mu}$ and $\hat{\mathscr{R}}$, where $\hat{\mathscr{R}}$ is the integral operator with kernel $\hat{\rho}(s,t)$ introduced above, restricted to $\mathcal{I} \times \mathcal{I}$.

All methods discussed in the previous section can be formulated in terms of the population parameters, i.e., the mean difference and covariance operator, and not in terms of individual observations in the training set. The population parameters can be consistently estimated by averaging individual observations, whereas temporal averaging of individual curves, for example in inner products, is impossible due the incompleteness of the observed functions. In particular, the conjugate gradient method can be applied to fragmentary training data, whereas the usual algorithms for multivariate or functional partial least squares, such as those in De Jong (1993), Hastie et al. (2009, Algorithm 3.3) and Delaigle & Hall (2012b, § 4.2 and Appendix A.2), involve the computation of certain scores, i.e., inner products, for individual curves.

3.2. *Asymptotic behaviour along the empirical regularization path*

We aim to study the behaviour of classifiers on incomplete training samples of increasing size with decreasing amounts of regularization. Previous asymptotic results in related settings include those of Delaigle & Hall (2013), who established the consistency of empirical principal component classifiers based on partially observed training data. In the setting of complete curves, Berrendero et al. (2018) used dimension reduction regularization by evaluation of curves at a finite set of arguments; they proved consistency of the empirical version but did not study the asymptotics for decreasing amounts of regularization, i.e., they did not consider letting the dimension grow. Baíllo et al. (2011a) studied optimal classifiers for Gaussian measures based on Radon–Nikodym derivatives and investigated the performance of their empirical version in the special class of processes with triangular covariance functions. In contrast, all of our methods, including the ridge approach not considered previously, have been developed for fragmentary training samples and shown to achieve the Bayes error rate for general Gaussian processes along the empirical regularization path, as we now explain.

The following assumptions will be needed for the derivation of asymptotic properties of empirically trained regularized linear classifiers.

Assumption 1. The distributions in groups $j = 0, 1$ satisfy $E_{P_j}(\|X\|^4) < \infty$.

Assumption 2. For a domain $\mathcal{I}$, there exists $\delta > 0$ such that the observation patterns in training samples $j = 0, 1$ satisfy, as $n_j \to \infty$,

$$\sup_{(s,t)\in\mathcal{I}\times\mathcal{I}} \text{pr}\{n_j^{-1} M_j(s,t) > \delta\} = O(n_j^{-2}).$$

Assumption 1 guarantees the consistency of the empirical mean and covariance operator for samples of completely observed curves; see, for example, Bosq (2000) or Horváth & Kokoszka (2012). Kraus (2015, Proposition 1) showed, under the additional Assumption 2 with $\mathcal{I}$ equal to the entire domain of the curves, that the root-n consistency of the sample mean and covariance restricted to $\mathcal{I}$ continues to hold in the fragmentary setting. In particular, it follows that $\|\hat{\mu}_j - \mu_j\| = O_{\text{p}}(n_j^{-1/2})$ and hence $\|\hat{\mu} - \mu\| = O_{\text{p}}(n^{-1/2})$ for $n = \min(n_0, n_1) \to \infty$, and also that $\|\hat{\mathscr{R}} - \mathscr{R}\|_\infty = O_{\text{p}}\{(n_0 + n_1)^{-1/2}\}$, where $\|\cdot\|_\infty$ is the operator norm. When $\mathcal{I}$ is a subset of

the domain, analogous results hold for the restrictions of the functions and integral kernels to $\mathcal{I}$. Assumption 2 means that at all pairs of time-points there is an asymptotically nonnegligible fraction of observed values. Assumption 2 is less restrictive than the requirement that there be complete curves in the sample. It can be satisfied, for example, in situations where the observed curves consist of several shorter fragments. If the assumption is not satisfied because the data contain only one short fragment per curve, other estimation methods can be used; see, for example, Delaigle & Hall (2016) and Descary & Panaretos (2019).

We now study the asymptotic behaviour of the empirical classifier when the number m_n of steps of the conjugate gradient algorithm grows as the training sample size grows. Under certain conditions on the regularization path, we establish the convergence of the misclassification probability of the conjugate gradient classifier trained on collections of functional fragments to the same optimal limit as for the theoretical conjugate gradient classifier with an infinite training sample, regardless of whether the limiting error rate is zero or positive and regardless of whether the limit can be theoretically achieved exactly or along the path.

THEOREM 1. *Suppose that Assumption 1 holds. Assume that* $n = \min(n_0, n_1) \to \infty$ *and* $m_n \to \infty$ *in such a way that* $m_n \leqslant Cn^{1/2}$ *for some* $C > 0$ *and*

$$n^{-1/2}\omega_{m_n}^{-1}\|\gamma^{(m_n)}\| + n^{-1}\omega_{m_n}^{-3} \to 0, \tag{4}$$

where ω_{m_n} *is the smallest eigenvalue of the* $m_n \times m_n$ *matrix* H *with entries* $h_{jk} = \langle \kappa_j, \mathscr{R}\kappa_k \rangle$ *for* $\kappa_j = \mathscr{R}^{j-1}\mu$ *and the* m_n*-vector* $\gamma^{(m_n)}$ *is defined as* $\gamma^{(m_n)} = H^{-1}d$ *with* d *being the* m_n*-vector having components* $d_j = \langle \mu, \kappa_j \rangle$. *Then the misclassification probability of the empirical regularized linear classifier based on* $\hat{\psi}_{m_n}^{\mathrm{CG}}$ *converges in probability to the optimal misclassification probability* $1 - \Phi(\|\mathscr{R}^{-1/2}\mu\|/2)$.

Condition (4) guarantees that the number of components does not grow too fast in relation to the growing number of training observations and to the increased ill-conditioning of the theoretical problem. Condition (4) is analogous to (5.10) in Delaigle & Hall (2012b) for partial least squares. The vector $\gamma^{(m_n)}$ contains the coefficients of the theoretical regularized solution $\psi_{m_n}^{\mathrm{CG}}$ with respect to the non-orthogonal basis $\kappa_1, \ldots, \kappa_{m_n}$ of the Krylov subspace $K_{m_n}(\mathscr{R}, \mu)$, i.e., $\psi_{m_n} = \sum_{j=1}^{m_n} \gamma_j^{(m_n)} \kappa_j$. The eigenvalues of H are called the Ritz values in numerical analysis. For details on connections with partial least squares see Lingjærde & Christophersen (2000).

In the proof given in the Appendix we use the results of Delaigle & Hall (2012b) on the consistency of partial least squares regression for functional data. These results were obtained for situations that differ from our setting in several ways. In particular, we work with functional fragments instead of complete curves, the conjugate gradient path differs from partial least squares regression, e.g., in the group centring in the estimation of the covariance, and we do not require that the population inverse problem, $\mathscr{R}\psi = \mu$ in our context, have a solution. However, inspection of the underlying technical arguments in Delaigle & Hall (2012b) shows that appropriate analogous results can be obtained and used in our setting, as we explain in the proof.

Next, we show that the empirically trained principal component classifier with an increasing number of components asymptotically achieves the optimal misclassification probability.

THEOREM 2. *Suppose that Assumption 1 holds. Assume that* $n = \min(n_0, n_1) \to \infty$ *and* $m_n \to \infty$ *in such a way that* $\lambda_{m_n}^4 n \to \infty$ *and* $\lambda_{m_n}^2 n (\sum_{j=1}^{m_n} a_j)^{-2} \to \infty$, *where* $a_1 = 2^{3/2}(\lambda_1 - \lambda_2)^{-1}$ *and* $a_j = 2^{3/2}\max\{(\lambda_{j-1} - \lambda_j)^{-1}, (\lambda_j - \lambda_{j+1})^{-1}\}$ *for* $j = 2, 3, \ldots$. *Then the misclassification*

probability of the empirical regularized linear classifier based on $\hat{\psi}^{\mathrm{PC}}_{m_n}$ *converges in probability to the optimal misclassification probability* $1 - \Phi(\|\mathscr{R}^{-1/2}\mu\|/2)$.

The conditions on the principal component regularization path are the same as in the case of functional principal component regression (Cardot et al., 1999). Unlike in the functional linear model, it is not assumed that the inverse problem has a solution, since the goal is not to estimate the possibly nonexistent bounded linear regression functional.

Finally, the empirical ridge classifier with finite training data asymptotically attains the same optimal error rate as its theoretical counterpart. Unlike for the conjugate gradient and principal component classifiers, the conditions on the ridge path classifier do not involve parameters of the distributions because no subspace is constructed.

THEOREM 3. *Suppose that Assumption 1 holds. Assume that* $n = \min(n_0, n_1) \to \infty$ *and* $\alpha_n \to 0+$ *in such a way that* $\alpha_n^4 n \to \infty$. *Then the misclassification probability of the empirical regularized linear classifier based on* $\hat{\psi}^{\mathrm{R}}_{\alpha_n}$ *converges in probability to the optimal misclassification probability* $1 - \Phi(\|\mathscr{R}^{-1/2}\mu\|/2)$.

3.3. *Selection of the regularization parameter*

The regularization parameter can be selected by minimizing an estimate of the misclassification probability. We use leave-one-out crossvalidation. The Supplementary Material provides details of crossvalidation in the presence of incomplete curves. The best value of the regularization parameter is searched for over a grid of values, such as the values corresponding to integer degrees of freedom up to some maximum value. The number of degrees of freedom for the subspace methods is the dimension of the subspace, and for the ridge method it is defined as the trace of $(\hat{\mathscr{R}} + \alpha\mathscr{I})^{-1}\hat{\mathscr{R}}$, i.e., $\sum_{j=1}^{n_0+n_1} \hat{\lambda}_j/(\hat{\lambda}_j + \alpha)$ where $\hat{\lambda}_j$ are the eigenvalues of $\hat{\mathscr{R}}$. The maximum number of degrees of freedom we use is one fifth of the number of curves.

4. DOMAIN SELECTION

To classify the new curve X_{new} observed on O_{new}, we apply the classifier on the target domain $\mathcal{I} \subseteq O_{\mathrm{new}}$, the choice of which we now consider. One possibility would be to restrict attention to the intersection of the observation domains of all curves, say $\mathcal{I}_0$, if it is nonempty. An obvious drawback of this approach is that one can lose discriminatory power because any differences between the classes may be more pronounced outside $\mathcal{I}_0$. An advantage of our approach is its capability of working with incomplete curves, since the empirical construction of the projection direction requires only the estimation of μ and $\mathscr{R}$ on the target domain. Hence one can look at a domain larger than $\mathcal{I}_0$. A natural choice is the largest subset of O_{new} that contains enough data for estimation of the classifier, i.e., satisfies Assumption 2, and contains enough functions for validation in the crossvalidation procedure, i.e., has a sufficiently large set V. In this way one hopes to capture the widest range of shapes of the group difference. On the other hand, it could be that not even this maximal domain, $\mathcal{I}^{\max}$, will lead to the best classification accuracy, because one includes more uncertainty in the estimation due to the missing values and because the mean difference may not be important in the added part of the domain. Therefore, it seems reasonable to also consider intermediate choices between $\mathcal{I}_0$ and $\mathcal{I}^{\max}$.

Here we present a domain selection strategy for the most common case of interval observation sets. The idea, worked out in detail in Stefanucci et al. (2018), is to construct the classifier on a series of intervals that range from the common domain $\mathcal{I}_0$ to the maximal domain $\mathcal{I}^{\max}$, extending the working interval by a fixed percentage at each step. More formally, we consider a sequence

of nested intervals $\mathcal{I}_0 \subset \mathcal{I}_1 \subset \cdots \subset \mathcal{I}_k \subset \cdots \subset \mathcal{I}_K = \mathcal{I}^{\max}$, starting from $\mathcal{I}_0$ and ending in $\mathcal{I}_K = \mathcal{I}^{\max}$, and build the classifier on each interval. The regularization parameter for the kth domain is selected by crossvalidation as described in the Supplementary Material. Among these $K+1$ candidates we select the one that minimizes the crossvalidation estimate of error.

The search strategy can be extended by considering larger systems of candidate domains; for example, one could vary the two endpoints independently. The idea can be generalized to other situations, such as non-interval observation sets, multivariate functional data or functions indexed by multivariate arguments. In each situation one needs to define a meaningful system of domains and optimize the crossvalidation score over the system.

5. Simulations

5.1. *Behaviour of regularized classifiers on complete data*

In this section we illustrate the behaviour of the three estimators of ψ in different settings. We consider Gaussian processes on $[0, 1]$ with covariance kernel $\rho(s,t) = \exp(-|s-t|^2/0.01)$ and mean function depending on the group label. Group 0 has mean $\mu_0(t) = 0$ in each setting. Group 1 has mean $\mu_1(t) = \mu(t)$, for which we consider eight different forms: (i) ct, (ii) $c(t-0.5)^2$, (iii) $c(t-0.5)^3$, (iv) $c\sin(20t)$, (v) $c\varphi_1(t)$, (vi) $c\varphi_{10}(t)$, (vii) $cb(t;5,5)$, and (viii) $cb(t;2,6)$, where φ_j is the jth eigenfunction of the kernel ρ and $b(t;\alpha,\beta) = t^{\alpha-1}(1-t)^{\beta-1}$ is the beta density. In each case the parameter c is selected to yield a reasonable misclassification rate.

In each of 5000 repetitions we generated 50 curves from each group and evaluated them on a grid of 100 equispaced points in $[0, 1]$. We also generated a new observation that could arise from group 0 or group 1 with equal probability. Then we constructed the regularized classification direction by the principal component, conjugate gradient and ridge methods with m degrees of freedom and predicted the label of the new observation. We considered $m = 1, \ldots, 20$, corresponding to a reasonable minimum of five observations per degree of freedom.

Figure 1 shows the misclassification proportion over the 5000 repetitions as a function of m for the eight different choices of $\mu(t)$. As expected, the conjugate gradient method performs well in all settings and is not much affected by the shape of $\mu(t)$. By contrast, the performance of the principal component classifier depends strongly on $\mu(t)$. To see this, consider the two extreme situations in settings (v) and (vi). The classification error of the principal component approach is close to that of the conjugate gradient method in case (v), where $\mu(t)$ is the first eigenfunction, but is much higher at lower dimensions in case (vi), where $\mu(t)$ is the tenth eigenfunction. In the latter case, the principal component method reaches the same level of error as the conjugate gradient method only when $m = 10$ or more. These findings agree with Proposition 2 and with the conclusions of Delaigle & Hall (2012a) and Febrero-Bande et al. (2017), who pointed out that principal components need more degrees of freedom than partial least squares to achieve good performance. In this regard ridge regularization seems to lie between the two subspace methods, but is more similar to the conjugate gradient method in most cases. In particular, it does not completely fail at low degrees of freedom in case (vi), because it does not construct a subspace that could miss the important information; however, it also suffers in this situation, where $\mu(t)$ is on the tail of the spectrum, because ridge penalization shrinks higher-index spectral components more than lower-index components. Nevertheless, with sufficiently many degrees of freedom, the three methods behave similarly.

Additional simulation results, reported in the Supplementary Material, show that similar conclusions can be drawn when functions have nonsmooth trajectories and that the capability to discriminate between two groups with different means is robust with respect to the assumption of equal covariances. Results for increased training sample size are also provided in the Supplementary Material.

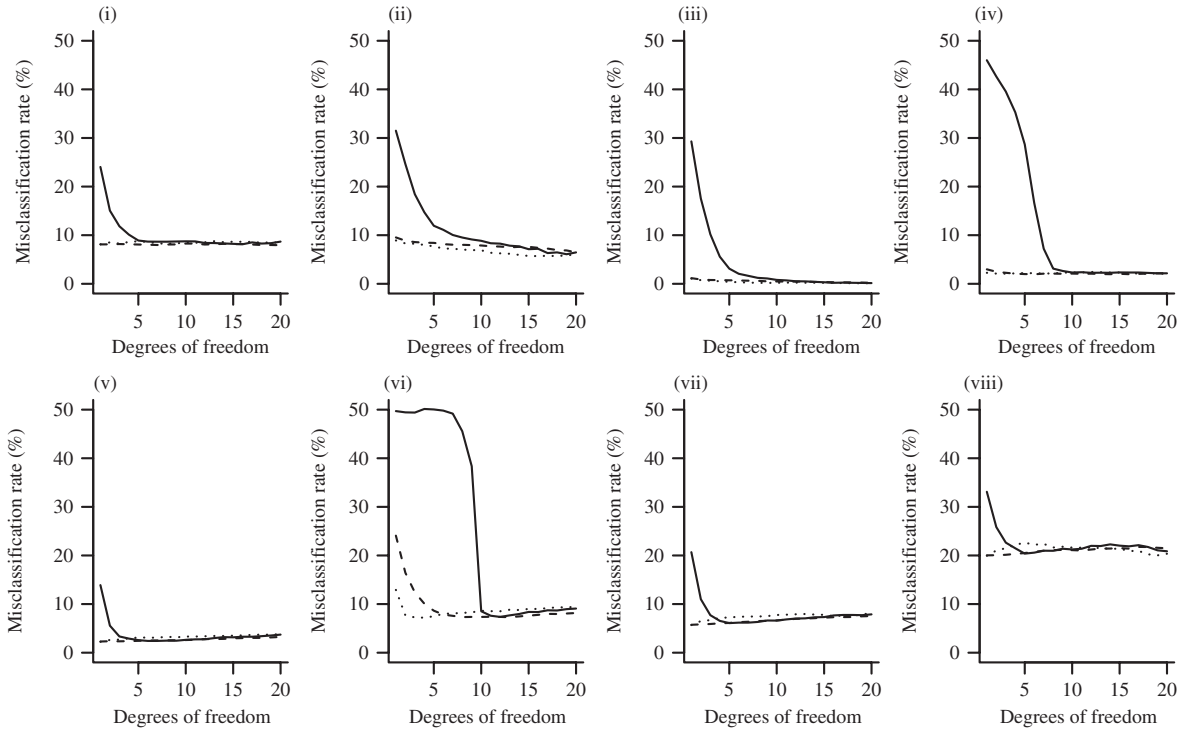

Fig. 1. Misclassification rate (%) versus degrees of freedom for different forms of $\mu(t)$: (i) linear, (ii) quadratic, (iii) cubic, (iv) sinusoidal, (v) first eigenfunction, (vi) tenth eigenfunction, (vii) symmetric beta, and (viii) asymmetric beta. The different curves represent the principal component (solid), conjugate gradient (dotted) and ridge (dashed) classifiers.

Table 1. *Misclassification rates* (%)*, with standard errors in parentheses, achieved by classifiers with degrees of freedom selected by crossvalidation in the different settings; for each classifier the numbers in the second row are the minimum misclassification rates*

	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)
PC	13.0 (0.34)	8.3 (0.28)	1.3 (0.11)	2.5 (0.16)	7.2 (0.26)	7.6 (0.27)	10.7 (0.31)	26.2 (0.44)
	8.1	6.1	0.1	2.2	2.4	7.4	6.1	20.4
CG	8.6 (0.28)	6.5 (0.25)	0.7 (0.09)	2.1 (0.14)	2.6 (0.16)	7.8 (0.27)	6.1 (0.24)	20.9 (0.41)
	8.1	5.7	0.1	2.1	2.2	7.2	5.7	19.9
R	8.4 (0.28)	7.7 (0.27)	0.7 (0.09)	2.2 (0.15)	2.4 (0.15)	7.9 (0.27)	6.1 (0.24)	20.8 (0.41)
	7.9	6.5	0.2	2.0	2.3	7.3	5.7	20.0

PC, principal component classifier; CG, conjugate gradient classifier; R, ridge classifier.

5.2. *Performance of crossvalidation for selection of degrees of freedom*

We used simulation to investigate the performance of leave-one-out crossvalidation in choosing the correct level of regularization. The settings were the same as in § 5.1, but classification was done using the number of degrees of freedom selected by leave-one-out crossvalidation. We summarize the classification errors in Table 1. Crossvalidation performs well as a selector of the best level of regularization since the misclassification rate in Table 1 is in each case close to the corresponding minimum error in Fig. 1. The principal component method appears to perform worst, while the conjugate gradient and ridge methods have comparable performance. The latter two methods nearly achieve the respective minimum errors. Table 2 reports the mean and median selected degrees of freedom. The principal component method often uses considerably more degrees of freedom than the other methods. This is particularly interesting in case (v), where the

Table 2. *Mean and median (in parentheses) degrees of freedom selected by crossvalidation*

	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)
PC	8.2 (7)	14.3 (15)	9.9 (9)	10.9 (10)	4.6 (4)	11.9 (11)	5.3 (4)	8.6 (6)
CG	5.4 (3)	10.7 (11)	3.4 (2)	4.5 (2)	2.4 (1)	4.9 (3)	2.7 (1)	8.6 (7)
R	6.4 (3)	11.6 (13)	6.0 (3)	6.1 (4)	2.7 (1)	9.3 (8)	3.4 (1)	6.7 (3)

PC, principal component classifier; CG, conjugate gradient classifier; R, ridge classifier.

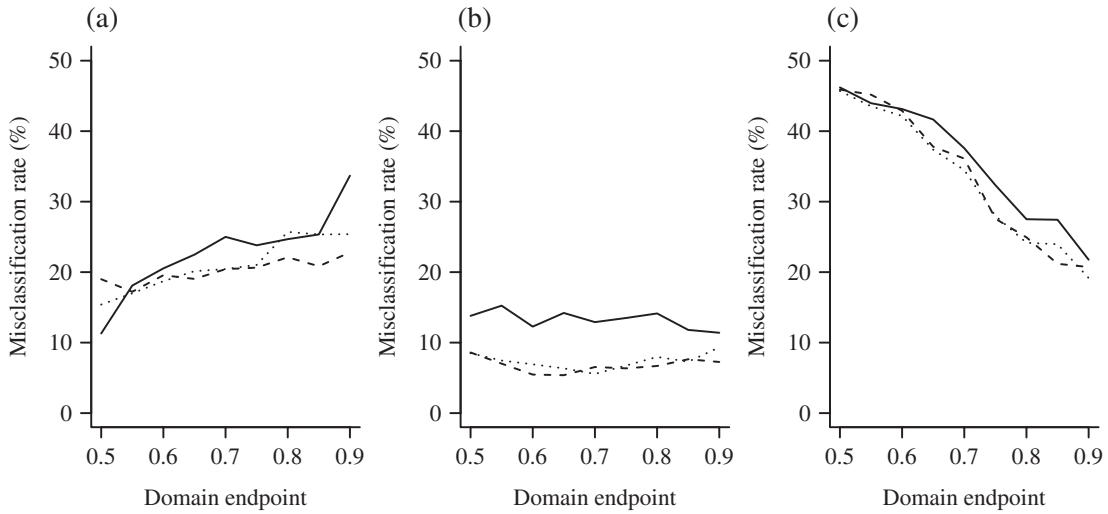

Fig. 2. Misclassification rate (%) plotted as a function of the domain extension, for $\mu(t)$ being the (a) Be(2, 6), (b) Be(5, 5) or (c) Be(6, 2) density for the principal component (solid), conjugate gradient (dotted) and ridge (dashed) classifiers with selected degrees of freedom. Classification is performed on the domains $[0, u]$ with $u \in [0.5, 0.9]$, and the error values are plotted against u.

mean difference equals the first eigenfunction and so one component should be the best choice in theory. These results again illustrate the general phenomenon that the principal component approach is inappropriate for inference about means due to the possible lack of informativeness of the principal components about the mean and the extra uncertainty associated with their estimation.

5.3. *Missing data and domain extension*

We now demonstrate the usefulness of the domain extension approach presented in § 4, using Gaussian processes on [0, 1] with the same covariance as in § 5.1 and considering three scenarios for the mean difference in the form of a multiple of a beta density, (a) $b(t; 2, 6)$, (b) $b(t; 5, 5)$ and (c) $b(t; 6, 2)$, which reflect situations where discrimination due to a peak is in the left, central and right parts of the domain, respectively. We sampled 50 curves from each group on a sequence of 100 equispaced points in [0, 1]. Then we generated endpoints of the observation interval for each curve from the uniform distribution on (0.5, 1); that is, each curve was observed between 0 and the endpoint and treated as missing beyond the endpoint. The new observation had an endpoint sampled between 0.5 and 1. So the first half of [0, 1], $\mathcal{I}_0 = [0, 0.5]$, was the common observation domain of all curves. We considered extensions of $\mathcal{I}_0$ to $\mathcal{I}_k = [0, 0.5 + 0.05k]$ $(k = 0, \ldots, 8)$. For each interval of this form that was contained in the observation domain of the curve to be classified, we estimated the classifiers, choosing the best degrees of freedom via crossvalidation, and classified the new curve. This procedure was repeated 1000 times. We plot the behaviour of the resulting classification error as a function of the endpoint of the extended domain in Fig. 2.

When the peak of the mean difference is in the left part of [0, 1], extending the domain does not lead to better classification. In this case the interval where the means mainly differ corresponds to the part of the domain where all the data are available, and inflating the domain only increases

Table 3. *Misclassification rates* (%)*, with standard errors in parentheses, achieved by classifiers with domain and degrees of freedom selected by crossvalidation in the different settings; the minimum and maximum misclassification rates are given in square brackets*

	(a)	(b)	(c)
PC	18.1 (0.38) [11.3, 33.7]	11.9 (0.32) [11.4, 15.2]	31.1 (0.46) [21.8, 46.0]
CG	19.6 (0.39) [15.4, 25.7]	7.4 (0.26) [5.6, 9.3]	30.4 (0.46) [19.2, 45.7]
R	22.4 (0.42) [17.2, 22.8]	6.9 (0.25) [5.4, 8.6]	28.4 (0.45) [20.7, 45.9]

PC, principal component classifier; CG, conjugate gradient classifier; R, ridge classifier.

the uncertainty due to missing data. In the second case, the peak of the mean difference is exactly at 0.5, and extending the domain leads to little improvement. The third scenario is the opposite of the first, as the discrimination is mainly in the right part of $[0, 1]$. In this case, extending the domain reduces the error considerably because good classification is only possible by employing the right part of the domain. The classification error is about 45% when using only $\mathcal{I}_0$, but drops to about 20% when using also the part of the interval where the data are partially observed.

5.4. *Performance with selected domain*

Domain extension may or may not improve the performance of classifiers, depending on the interplay between the form of the mean difference, the covariance structure and the missingness pattern. In practice, the user is not an oracle with access to misclassification errors for candidate subsets whose estimates are plotted in Fig. 2, and hence would select the best domain by crossvalidation. In Table 3 we report simulation results for classifiers with both domain and degrees of freedom selected by crossvalidation, for the same configurations as in § 5.3. Selection of the domain leads to a considerable improvement of the error rate compared with the worst-performing domain. On the other hand, this improvement has some limitations and a gap remains between the achieved value and the best value; this can be explained by the fact that crossvalidation provides only an estimate of the error, not the true value.

6. ANEURISK DATA EXAMPLE

We apply the proposed method to the AneuRisk dataset from an interdisciplinary project aimed at investigating the effects of blood vessel morphology, blood fluid dynamics and biomechanical properties of the vascular wall on the pathogenesis of cerebral aneurysms. An introduction to the data can be found in Sangalli et al. (2014b). This dataset has previously been analysed in several works that focused on different methodological aspects, such as function and derivative estimation (Sangalli et al., 2009b), exploratory analysis and classification (Sangalli et al., 2009a), and alignment and clustering (Sangalli et al., 2014a), among others.

The data consist of measurements of the radius and curvature of the internal carotid artery in a sample of 65 patients, 33 of which have an aneurysm at the bifurcation of the vessel or after it, while the other 32 either have an aneurysm before the bifurcation, which is much less dangerous, or are healthy. The goal is to classify the patients based on the morphology of their internal carotid artery. In this example we work with only one of the observed variables, the radius. The data have previously been pre-processed, registered and smoothed, and are observed on a grid of 2000 points in the interval $[-100.3, 5.1]$, where the argument represents the distance between the observation point and the terminal bifurcation of the internal carotid artery, with positive values indicating points inside the skull. As we can see in Fig. 3, the data are partially observed because

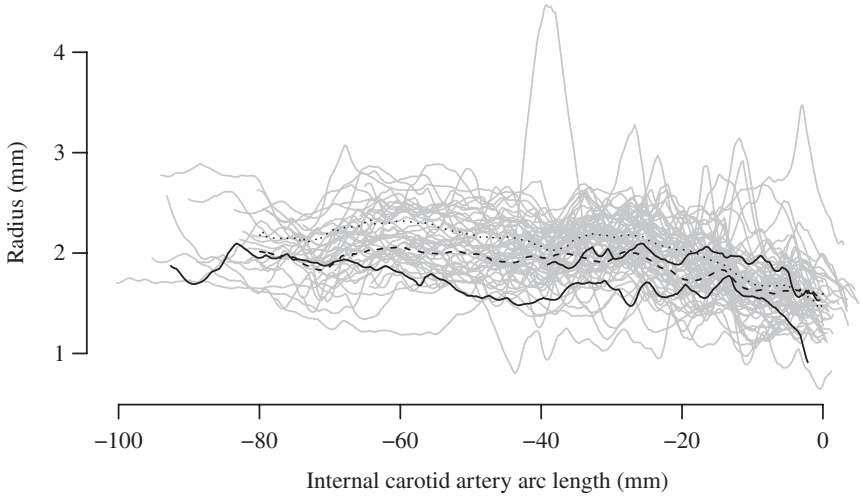

Fig. 3. Radius along the carotid artery from the AneuRisk dataset, along with the mean of the group of subjects with an aneurysm after the bifurcation (dotted) and the mean of the group of subjects with an aneurysm before the bifurcation or without an aneurysm (dashed). Curves for two example subjects are highlighted as solid lines. Note the different start and end points for different subjects in the study.

the start and end points are different from subject to subject. All subjects are observed on the subset $\mathcal{I}_0 = [-32.9, -7.4]$, which corresponds to 24.3% of the whole domain.

We first apply the regularized linear classifiers to curves restricted to the common domain $\mathcal{I}_0$. The classification error estimated by crossvalidation is 29.2% for the principal component method, 29.2% for the conjugate gradient method, and 32.3% for ridge regularized classification.

We compare the above procedure with a different approach consisting of a multivariate classification method applied to principal component scores. The covariance kernel is estimated from observations centred to their respective group means, its eigenfunctions are computed, and quadratic discriminant analysis is applied to the inner products of the uncentred curves with the eigenfunctions. This procedure is similar to that in Sangalli et al. (2009a). The best classifier of this type turns out to exhibit a misclassification error of 32.3%, obtained with two eigenfunctions.

These values show that in this dataset, when attention is restricted to the common domain $\mathcal{I}_0$, our proposed method is comparable to the more standard multivariate technique.

Next, we consider classification on extended domains including observed values outside the common domain $\mathcal{I}_0$. We build the sequence of domains $\mathcal{I}_0, \ldots, \mathcal{I}_K$ by enlarging the domain at each step by 1.25% of the complement of $\mathcal{I}_0$. This step size is a compromise between the fineness of the grid and the computational cost. We consider extended domains up to $K = 40$, corresponding to $\mathcal{I}_{40} = [-66.6, -1.2]$, because not enough subjects have observed values outside this interval for reliable estimation and crossvalidation. All regularized linear classification methods benefit from the domain extension; in particular, the error rate for the principal component method drops from 29.2% to 23.2%, for the conjugate gradient method from 29.2% to 25.8%, and for ridge regularization from 32.3% to 25%. The best domain is $\mathcal{I}_{10} = [-41.3, -5.8]$ for the conjugate gradient method and $\mathcal{I}_{11} = [-42.2, -5.7]$ for the other two methods.

The alternative method based on multivariate classification of scores cannot be applied on extended domains since the individual scores of incomplete curves cannot be computed, although they can be predicted (Kraus, 2015). By contrast, the proposed methods are entirely formulated in terms of distributional parameters, which can be consistently estimated from incomplete data, unlike individual quantities.

ACKNOWLEDGEMENT

The AneuRisk data and useful comments were kindly provided by Laura Sangalli. The work of David Kraus was supported by the Czech Science Foundation. We are grateful to two referees, an associate editor and the editor for helpful suggestions and corrections.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes the derivation of classifiers under unequal prior class probabilities, algorithmic details of crossvalidation, and additional simulation and real-data results.

APPENDIX

Proof of Proposition 1

The misclassification probability for ψ_m is $D(\psi_m)$ given in (1). Since $\psi_m \in S_m$, we compute

$$\frac{|\langle \mu, \psi_m \rangle|}{\langle \psi_m, \mathscr{R}\psi_m \rangle^{1/2}} = \frac{\langle \mu, \mathscr{R}_m^- \mu \rangle}{\langle \mu, \mathscr{R}_m^- \mathscr{R} \mathscr{R}_m^- \mu \rangle^{1/2}} = \|(\mathscr{R}_m^-)^{1/2}\mu\|.$$

By Lebesgue's monotone convergence theorem, the right-hand side converges to $\|\mathscr{R}^{-1/2}\mu\|$, finite or infinite, and therefore the limiting misclassification probability that is attained along the regularization path ψ_m, as $m \to \infty$, is $1 - \Phi(\|\mathscr{R}^{-1/2}\mu\|/2)$.

Proof of Proposition 2

The conjugate gradient method minimizes the quadratic objective function in the Krylov subspace $K_m(\mathscr{R}, \mu)$ whose elements are in the form $\eta = \sum_{k=0}^{m-1} c_k \mathscr{R}^k \mu = p(\mathscr{R})\mu$, where p is a polynomial of order lower than m. Then $\eta \in K_m(\mathscr{R}, \mu)$ can be written as $\eta = \sum_{j=1}^{\infty} p(\lambda_j) b_j \varphi_j$ with $b_j = \langle \mu, \varphi_j \rangle$. The objective function at η equals

$$\begin{aligned} \langle \eta, \mathscr{R}\eta \rangle/2 - \langle \mu, \eta \rangle &= \langle p(\mathscr{R})\mu, \mathscr{R}p(\mathscr{R})\mu \rangle/2 - \langle \mu, p(\mathscr{R})\mu \rangle \\ &= \sum_{j=1}^{\infty} b_j^2 \{p(\lambda_j)^2 \lambda_j/2 - p(\lambda_j)\} \\ &= \sum_{j=1}^{\infty} \frac{b_j^2}{2\lambda_j} q(\lambda_j)\{q(\lambda_j) - 2\}, \end{aligned} \tag{A1}$$

where $q(\lambda) = p(\lambda)\lambda$ is a polynomial of degree at most m such that $q(0) = 0$. The conjugate gradient method seeks the polynomial with these properties that minimizes the objective function. To prove the proposition we shall find a polynomial q with the required properties such that the objective function above is smaller than or equal to the objective function for the principal component classifier. The principal component classifier uses $\psi_m^{\mathrm{PC}} = \sum_{j=1}^{m} \lambda_j^{-1} b_j \varphi_j$, and the objective function at ψ_m^{PC} is

$$\langle \psi_m^{\mathrm{PC}}, \mathscr{R}\psi_m^{\mathrm{PC}} \rangle/2 - \langle \mu, \psi_m^{\mathrm{PC}} \rangle = -\sum_{j=1}^{m} \frac{b_j^2}{2\lambda_j}. \tag{A2}$$

Consider the polynomial of degree m,

$$q(\lambda) = 1 - (-1)^m \frac{\lambda - \lambda_1}{\lambda_1} \cdots \frac{\lambda - \lambda_m}{\lambda_m},$$

with $q(0) = 0$. We see that $q(\lambda_j) = 1$ for $j = 1, \ldots, m$, so the first m summands in the series (A1) and (A2) are equal. For $j > m$ we have that $0 \leqslant q(\lambda_j) \leqslant 2$ due to the properties of the eigenvalue sequence; so $q(\lambda_j)\{q(\lambda_j) - 2\} \leqslant 0$ and therefore the corresponding summands in the series (A1) are negative, whereas they are zero in the series (A2). Hence, for this polynomial,

$$\sum_{j=1}^{\infty} \frac{b_j^2}{2\lambda_j} q(\lambda_i)\{q(\lambda_i) - 2\} \leqslant -\sum_{j=1}^{m} \frac{b_j^2}{2\lambda_j},$$

and so the objective at the conjugate gradient solution must be smaller than or equal to the objective at the principal component solution. The inequality between the minima of the quadratic objective function implies the inequality between the misclassification probabilities stated in the proposition.

Proof of Proposition 3

Proceeding as in the proof of Proposition 1, we need to show that

$$\frac{\langle \mu, \mathscr{R}_\alpha^{-1}\mu\rangle}{\langle \mu, \mathscr{R}_\alpha^{-1}\mathscr{R}\mathscr{R}_\alpha^{-1}\mu\rangle^{1/2}} = \frac{\sum_{j=1}^{\infty} \frac{b_j^2}{\lambda_j+\alpha}}{\left\{\sum_{j=1}^{\infty} \frac{\lambda_j b_j^2}{(\lambda_j+\alpha)^2}\right\}^{1/2}} \xrightarrow[\alpha\to 0+]{} \left(\sum_{j=1}^{\infty} \frac{b_j^2}{\lambda_j}\right)^{1/2} = \|\mathscr{R}^{-1/2}\mu\|,$$

where $b_j = \langle \mu, \varphi_j\rangle$ is the coefficient of μ in the eigenbasis. If $\sum_{j=1}^{\infty} b_j^2/\lambda_j < \infty$, the convergence follows from Lebesgue's monotone convergence theorem. Otherwise, we use the inequality $\sum_{j=1}^{\infty} \lambda_j b_j^2/(\lambda_j + \alpha)^2 \leqslant \sum_{j=1}^{\infty} b_j^2/(\lambda_j + \alpha)$ to bound the left-hand side expression from below by $\{\sum_{j=1}^{\infty} b_j^2/(\lambda_j + \alpha)\}^{1/2}$, which diverges to infinity again by Lebesgue's theorem.

Proof of Theorem 1

The probability of misclassifying a new observation using the conjugate gradient classifier based on $\hat{\psi}_{m_n}^{\mathrm{CG}}$ is $D(\hat{\psi}_{m_n}^{\mathrm{CG}}) = 1 - \Phi\{|Z(\hat{\psi}_{m_n}^{\mathrm{CG}})|/2\}$. We need to show that the fraction in $Z(\hat{\psi}_{m_n}^{\mathrm{CG}})$ converges in probability to $\|\mathscr{R}^{-1/2}\mu\|/2$ along the regularization path satisfying the assumptions of the theorem. To deal with the numerator in $Z(\hat{\psi}_{m_n}^{\mathrm{CG}})$, one can show that

$$\langle \mu, \hat{\psi}_{m_n}^{\mathrm{CG}}\rangle - \langle \mu, \psi_{m_n}^{\mathrm{CG}}\rangle = O_{\mathrm{p}}\big(n^{-1/2}\omega_{m_n}^{-1}\|\gamma^{(m_n)}\| + n^{-1}\omega_{m_n}^{-3}\big). \tag{A3}$$

This result follows from an analogue of (5.9) in Theorem 5.3 of Delaigle & Hall (2012b) and intermediate results in the proof of that theorem which can be established in our context. The necessary modifications of the proofs of Theorems 5.1, 5.2 and 5.3 in Delaigle & Hall (2012b) are as follows. All results remain valid for incomplete instead of complete curves, because the proofs depend only on the root-n consistency of the covariance estimators, which holds also for functional fragments (Kraus, 2015, Proposition 1). Moreover, the derivations in Delaigle & Hall (2012b) can be repeated without assuming that the theoretical solution $\psi = \mathscr{R}^{-1}\mu$ exists as an element of $L^2(\mathcal{I})$. Indeed, the proofs in Delaigle & Hall (2012b) are based on stochastic expansions of $\hat{\mathscr{R}}^j\psi = \hat{\mathscr{R}}^j\mathscr{R}^{-1}\mu$, in our notation, about $\mathscr{R}^j\psi = \mathscr{R}^j\mathscr{R}^{-1}\mu = \mathscr{R}^{j-1}\mu$ and derived quantities, but the same steps can be followed for $\hat{\mathscr{R}}^{j-1}\hat{\mu}$ about $\mathscr{R}^{j-1}\mu$ in our setting. In other words, it can be shown that $\hat{\psi}_{m_n}^{\mathrm{CG}}$ and $\psi_{m_n}^{\mathrm{CG}}$ converge to each other without assuming that $\psi_{m_n}^{\mathrm{CG}}$ converges. Similarly, for the denominator in $Z(\hat{\psi}_{m_n}^{\mathrm{CG}})$ we have that

$$\langle \hat{\psi}_{m_n}^{\mathrm{CG}}, \mathscr{R}\hat{\psi}_{m_n}^{\mathrm{CG}}\rangle - \langle \psi_{m_n}^{\mathrm{CG}}, \mathscr{R}\psi_{m_n}^{\mathrm{CG}}\rangle = O_{\mathrm{p}}\big(n^{-1/2}\omega_{m_n}^{-1}\|\gamma^{(m_n)}\| + n^{-1}\omega_{m_n}^{-3}\big). \tag{A4}$$

This last result is analogous to (7.27) of Delaigle & Hall (2012b), whose proof can be repeated with the same modifications for our situation as before. Therefore, regardless of whether $\|\mathscr{R}^{-1}\mu\|$ or $\|\mathscr{R}^{-1/2}\mu\|$ is finite or infinite, the theoretical and empirical regularized quantities approach each other at the rates given in (A3) and (A4). The result on $D(\hat{\psi}_{m_n}^{\mathrm{CG}})$ then follows as in the proof of Proposition 1.

Proof of Theorem 2

We show that $D(\hat{\psi}^{\mathrm{PC}}_{m_n}) = 1-\Phi\{|Z(\hat{\psi}^{\mathrm{PC}}_{m_n})|/2\}$ converges in probability to $1-\Phi(\|\mathscr{R}^{-1/2}\|/2)$. The strategy of the proof is similar to that of Theorem 3.1 in Cardot et al. (1999) for the principal component approach to the functional linear model. The difference lies in the incompleteness of the functional data and in that we do not assume that the underlying theoretical inverse problem has a solution. We write

$$\|\hat{\psi}^{\mathrm{PC}}_{m_n} - \psi^{\mathrm{PC}}_{m_n}\| \leqslant \|\hat{\mathscr{R}}^{-}_{m_n} - \mathscr{R}^{-}_{m_n}\|_\infty \|\hat{\mu}\| + \|\mathscr{R}^{-}_{m_n}\|_\infty \|\hat{\mu} - \mu\|.$$

Proceeding as in the proof of Lemma 5.1 in Cardot et al. (1999), we can show that

$$\|\hat{\mathscr{R}}^{-}_{m_n} - \mathscr{R}^{-}_{m_n}\|_\infty \leqslant \hat{\lambda}^{-1}_{m_n}\lambda^{-1}_{m_n}\|\hat{\mathscr{R}} - \mathscr{R}\|_\infty + 2\lambda^{-1}_{m_n}\|\hat{\mathscr{R}} - \mathscr{R}\|_\infty \sum_{j=1}^{m_n} a_j.$$

Here $\hat{\lambda}_j$ are the eigenvalues of $\hat{\mathscr{R}}$ in descending order and $\hat{\varphi}_j$ are the corresponding eigenfunctions. In establishing the above inequality one uses the facts that $|\hat{\lambda}_j - \lambda_j| \leqslant \|\hat{\mathscr{R}} - \mathscr{R}\|_\infty$ and $\|\hat{\varphi}_j - \mathrm{sign}\langle \hat{\varphi}_j, \varphi_j\rangle \varphi_j\| \leqslant a_j\|\hat{\mathscr{R}} - \mathscr{R}\|_\infty$, which are known from Bosq (2000, Lemmas 4.2 and 4.3) for the empirical covariance operator from complete curves but hold also for functional fragments; see the proof of Proposition 2 in the supplementary document for Kraus (2015). Since $\|\hat{\mathscr{R}} - \mathscr{R}\|_\infty = O_{\mathrm{p}}(n^{-1/2})$, we see that $\hat{\lambda}^{-1}_{m_n}\lambda^{-1}_{m_n}\|\hat{\mathscr{R}} - \mathscr{R}\|_\infty 1_{[\hat{\lambda}_{m_n} > \lambda_{m_n}/2]} \leqslant 2\lambda^{-2}_{m_n}\|\hat{\mathscr{R}} - \mathscr{R}\|_\infty = \lambda^{-2}_{m_n} O_{\mathrm{p}}(n^{-1/2})$. Since the probability of the event $[\hat{\lambda}_{m_n} < \lambda_{m_n}/2]$ is bounded by $\lambda^{-2}_{m_n} O(n^{-1})$ and hence converges to 0, it follows that $\hat{\lambda}^{-1}_{m_n}\lambda^{-1}_{m_n}\|\hat{\mathscr{R}} - \mathscr{R}\|_\infty = \lambda^{-2}_{m_n} O_{\mathrm{p}}(n^{-1/2})$. Combining this with the facts that $\|\hat{\mu}\| = O_{\mathrm{p}}(1)$, $\|\mathscr{R}^{-}_{m_n}\| = \lambda^{-1}_{m_n}$ and $\|\hat{\mu} - \mu\| = O_{\mathrm{p}}(n^{-1/2})$ gives

$$\|\hat{\psi}^{\mathrm{PC}}_{m_n} - \psi^{\mathrm{PC}}_{m_n}\| \leqslant \lambda^{-2}_{m_n} O_{\mathrm{p}}(n^{-1/2}) + \lambda^{-1}_{m_n} O_{\mathrm{p}}(n^{-1/2}) \sum_{j=1}^{m_n} a_j.$$

Similar arguments can be used in the analysis of the denominator in $Z(\hat{\psi}^{\mathrm{PC}}_{m_n})$. In conclusion, we obtain that the estimation errors for the quantities in the numerator and denominator converge to zero at the rates

$$\langle \mu, \hat{\psi}^{\mathrm{PC}}_{m_n}\rangle - \langle \mu, \psi^{\mathrm{PC}}_{m_n}\rangle = \lambda^{-2}_{m_n} O_{\mathrm{p}}(n^{-1/2}) + \lambda^{-1}_{m_n} O_{\mathrm{p}}(n^{-1/2}) \sum_{j=1}^{m_n} a_j, \tag{A5}$$

$$\langle \hat{\psi}^{\mathrm{PC}}_{m_n}, \mathscr{R}\hat{\psi}^{\mathrm{PC}}_{m_n}\rangle - \langle \psi^{\mathrm{PC}}_{m_n}, \mathscr{R}\psi^{\mathrm{PC}}_{m_n}\rangle = \lambda^{-2}_{m_n} O_{\mathrm{p}}(n^{-1/2}) + \lambda^{-1}_{m_n} O_{\mathrm{p}}(n^{-1/2}) \sum_{j=1}^{m_n} a_j. \tag{A6}$$

In light of (A5) and (A6), the asymptotic behaviour of the misclassification probability is driven by the behaviour of the theoretical classifier addressed in Proposition 1.

Proof of Theorem 3

We show that the fraction $|Z(\hat{\psi}^{\mathrm{R}}_{m_n})|$ converges in probability to $\|\mathscr{R}^{-1/2}\mu\|/2$ as $n \to \infty$. For the numerator we write

$$\langle \mu, \hat{\psi}^{\mathrm{R}}_{\alpha_n}\rangle - \langle \mu, \mathscr{R}^{-1}_{\alpha_n}\mu\rangle = \langle \mu, (\hat{\mathscr{R}}^{-1}_{\alpha_n} - \mathscr{R}^{-1}_{\alpha_n})\hat{\mu}\rangle + \langle \mu, \mathscr{R}^{-1}_{\alpha_n}(\hat{\mu} - \mu)\rangle. \tag{A7}$$

For the first term on the right we find that

$$\begin{aligned} |\langle \mu, (\hat{\mathscr{R}}^{-1}_{\alpha_n} - \mathscr{R}^{-1}_{\alpha_n})\hat{\mu}\rangle| &\leqslant \|\mu\| \|\hat{\mathscr{R}}^{-1}_{\alpha_n} - \mathscr{R}^{-1}_{\alpha_n}\|_\infty \|\hat{\mu}\| \\ &= \|\mu\| \|\hat{\mathscr{R}}^{-1}_{\alpha_n}(\hat{\mathscr{R}}_{\alpha_n} - \mathscr{R}_{\alpha_n})\mathscr{R}^{-1}_{\alpha_n}\|_\infty \|\hat{\mu}\| \end{aligned}$$

$$\leqslant \|\mu\|\|\hat{\mathscr{R}}_{\alpha_n}^{-1}\|_\infty\|\hat{\mathscr{R}}_{\alpha_n} - \mathscr{R}_{\alpha_n}\|_\infty\|\mathscr{R}_{\alpha_n}^{-1}\|_\infty\|\hat{\mu}\|$$
$$\leqslant \alpha_n^{-2}O_{\mathrm{p}}(n^{-1/2}),$$

since $\|\hat{\mathscr{R}}_{\alpha_n}^{-1}\|_\infty \leqslant \alpha_n^{-1}$, $\|\mathscr{R}_{\alpha_n}^{-1}\|_\infty \leqslant \alpha_n^{-1}$, $\|\hat{\mu}\| = O_{\mathrm{p}}(1)$ and $\|\hat{\mathscr{R}}_{\alpha_n} - \mathscr{R}_{\alpha_n}\|_\infty = \|\hat{\mathscr{R}} - \mathscr{R}\|_\infty = O_{\mathrm{p}}\{(n_0 + n_1)^{-1/2}\}$ (Kraus, 2015, Proposition 1). For the second term on the right-hand side of (A7), we obtain

$$|\langle \mu, \mathscr{R}_{\alpha_n}^{-1}(\hat{\mu} - \mu)\rangle| \leqslant \|\mu\|\|\mathscr{R}_{\alpha_n}^{-1}\|_\infty\|\hat{\mu} - \mu\| \leqslant \alpha_n^{-1}O_{\mathrm{p}}(n^{-1/2}).$$

The quantity in the denominator of $Z(\hat{\psi}_{m_n}^{\mathrm{R}})$ can be rewritten as

$$\langle \hat{\psi}_{\alpha_n}^{\mathrm{R}}, \mathscr{R}\hat{\psi}_{\alpha_n}^{\mathrm{R}}\rangle - \langle \psi_{\alpha_n}^{\mathrm{R}}, \mathscr{R}\psi_{\alpha_n}^{\mathrm{R}}\rangle = \langle \hat{\psi}_{\alpha_n}^{\mathrm{R}} - \psi_{\alpha_n}^{\mathrm{R}}, \mathscr{R}\hat{\psi}_{\alpha_n}^{\mathrm{R}}\rangle + \langle \psi_{\alpha_n}^{\mathrm{R}}, \mathscr{R}(\hat{\psi}_{\alpha_n}^{\mathrm{R}} - \psi_{\alpha_n}^{\mathrm{R}})\rangle. \tag{A8}$$

The first term on the right is

$$\begin{aligned}\langle \hat{\psi}_{\alpha_n}^{\mathrm{R}} - \psi_{\alpha_n}^{\mathrm{R}}, \mathscr{R}\hat{\psi}_{\alpha_n}^{\mathrm{R}}\rangle &= \langle \hat{\mathscr{R}}_{\alpha_n}^{-1}\hat{\mu} - \mathscr{R}_{\alpha_n}^{-1}\mu, \mathscr{R}\hat{\mathscr{R}}_{\alpha_n}^{-1}\hat{\mu}\rangle \\ &= \langle \mathscr{R}_{\alpha_n}^{-1}(\mathscr{R}_{\alpha_n} - \hat{\mathscr{R}}_{\alpha_n})\hat{\mathscr{R}}_{\alpha_n}^{-1}\hat{\mu}, \mathscr{R}\hat{\mathscr{R}}_{\alpha_n}^{-1}\hat{\mu}\rangle + \langle \mathscr{R}_{\alpha_n}^{-1}(\hat{\mu} - \mu), \mathscr{R}\hat{\mathscr{R}}_{\alpha_n}^{-1}\hat{\mu}\rangle. \end{aligned} \tag{A9}$$

For the first summand in (A9) we have

$$\begin{aligned}|\langle \mathscr{R}_{\alpha_n}^{-1}(\mathscr{R}_{\alpha_n} - \hat{\mathscr{R}}_{\alpha_n})\hat{\mathscr{R}}_{\alpha_n}^{-1}\hat{\mu}, \mathscr{R}\hat{\mathscr{R}}_{\alpha_n}^{-1}\hat{\mu}\rangle| &\leqslant \|\hat{\mu}\|^2\|\hat{\mathscr{R}}_{\alpha_n}^{-1}\|_\infty^2\|\mathscr{R}\mathscr{R}_{\alpha_n}^{-1}\|_\infty\|\hat{\mathscr{R}} - \mathscr{R}\|_\infty \\ &\leqslant \alpha_n^{-2}O_{\mathrm{p}}(n^{-1/2}),\end{aligned}$$

using properties mentioned previously and the fact that $\|\mathscr{R}\mathscr{R}_{\alpha_n}^{-1}\|_\infty \leqslant 1$, and for the second summand we have

$$|\langle \mathscr{R}_{\alpha_n}^{-1}(\hat{\mu} - \mu), \mathscr{R}\hat{\mathscr{R}}_{\alpha_n}^{-1}\hat{\mu}\rangle| \leqslant \|\mathscr{R}\mathscr{R}_{\alpha_n}^{-1}\|_\infty\|\hat{\mathscr{R}}_{\alpha_n}^{-1}\|_\infty\|\hat{\mu} - \mu\| \leqslant \alpha_n^{-1}O_{\mathrm{p}}(n^{-1/2}).$$

Putting these results together, we see that the absolute value of the first term on the right-hand side of (A8) is dominated by $\alpha_n^{-2}O_{\mathrm{p}}(n^{-1/2})$. The second term on the right-hand side of (A8) can be analysed in a similar way to the first two terms on the right-hand side of (A7) with $\mathscr{R}\mathscr{R}_{\alpha_n}^{-1}\mu$ in place of μ. Thus we bound the absolute value from above by $\alpha_n^{-2}O_{\mathrm{p}}(n^{-1/2})$. These results imply that the estimation errors vanish at rates

$$\langle \mu, \hat{\psi}_{\alpha_n}^{\mathrm{R}}\rangle - \langle \mu, \psi_{\alpha_n}^{\mathrm{R}}\rangle = \alpha_n^{-2}O_{\mathrm{p}}(n^{-1/2}),$$
$$\langle \hat{\psi}_{\alpha_n}^{\mathrm{R}}, \mathscr{R}\hat{\psi}_{\alpha_n}^{\mathrm{R}}\rangle - \langle \psi_{\alpha_n}^{\mathrm{R}}, \mathscr{R}\psi_{\alpha_n}^{\mathrm{R}}\rangle = \alpha_n^{-2}O_{\mathrm{p}}(n^{-1/2}).$$

Hence the empirical classifier has the same limiting error as the theoretical one addressed in Proposition 3.

References

Baíllo, A., Cuevas, A. & Cuesta-Albertos, J. A. (2011a). Supervised classification for a family of Gaussian functional models. *Scand. J. Statist.* **38**, 480–98.

Baíllo, A., Cuevas, A. & Fraiman, R. (2011b). Classification methods for functional data. In *The Oxford Handbook of Functional Data Analysis*, F. Ferraty & Y. Romain, eds. Oxford: Oxford University Press, pp. 259–97.

Berrendero, J. R., Cuevas, A. & Torrecilla, J. L. (2016). Variable selection in functional data classification: A maxima-hunting proposal. *Statist. Sinica* **26**, 619–38.

Berrendero, J. R., Cuevas, A. & Torrecilla, J. L. (2018). On the use of reproducing kernel Hilbert spaces in functional classification. *J. Am. Statist. Assoc.* **113**, 1210–8.

Blanchard, G. & Krämer, N. (2010). Kernel partial least squares is universally consistent. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Y. W. Teh & M. Titterington, eds., vol. 9 of *Proceedings of Machine Learning Research*. International Joint Conferences on Artificial Intelligence (IJCAI) Organization, pp. 57–64.

BONGIORNO, E. G. & GOIA, A. (2016). Classification methods for Hilbert data based on surrogate density. *Comp. Statist. Data Anal.* **99**, 204–22.
BOSQ, D. (2000). *Linear Processes in Function Spaces*. New York: Springer.
BUGNI, F. A. (2012). Specification test for missing functional data. *Economet. Theory* **28**, 959–1002.
CARDOT, H., FERRATY, F. & SARDA, P. (1999). Functional linear model. *Statist. Prob. Lett.* **45**, 11–22.
CUESTA-ALBERTOS, J. A., DEL BARRIO, E., FRAIMAN, R. & MATRÁN, C. (2007). The random projection method in goodness of fit for functional data. *Comp. Statist. Data Anal.* **51**, 4814–31.
CUEVAS, A. (2014). A partial overview of the theory of statistics with functional data. *J. Statist. Plan. Infer.* **147**, 1–23.
DAI, X., MÜLLER, H.-G. & YAO, F. (2017). Optimal Bayes classifiers for functional data and density ratios. *Biometrika* **104**, 545–60.
DE JONG, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemomet. Intel. Lab. Syst.* **18**, 251–63.
DELAIGLE, A. & HALL, P. (2012a). Achieving near perfect classification for functional data. *J. R. Statist. Soc.* B **74**, 267–86.
DELAIGLE, A. & HALL, P. (2012b). Methodology and theory for partial least squares applied to functional data. *Ann. Statist.* **40**, 322–52.
DELAIGLE, A. & HALL, P. (2013). Classification using censored functional data. *J. Am. Statist. Assoc.* **108**, 1269–83.
DELAIGLE, A. & HALL, P. (2016). Approximating fragmented functional data by segments of Markov chains. *Biometrika* **103**, 779–99.
DELAIGLE, A., HALL, P. & BATHIA, N. (2012). Componentwise classification and clustering of functional data. *Biometrika* **99**, 299–313.
DESCARY, M.-H. & PANARETOS, V. M. (2019). Recovering covariance from functional fragments. *Biometrika* **106**, 145–60.
FEBRERO-BANDE, M., GALEANO, P. & GONZÁLEZ-MANTEIGA, W. (2017). Functional principal component regression and functional partial least-squares regression: An overview and a comparative study. *Int. Statist. Rev.* **85**, 61–83.
FERRATY, F., HALL, P. & VIEU, P. (2010). Most-predictive design points for functional data predictors. *Biometrika* **97**, 807–24.
GOLDBERG, Y., RITOV, Y. & MANDELBAUM, A. (2014). Predicting the continuation of a function with applications to call center data. *J. Statist. Plan. Infer.* **147**, 53–65.
GROMENKO, O., KOKOSZKA, P. & SOJKA, J. (2017). Evaluation of the cooling trend in the ionosphere using functional regression with incomplete curves. *Ann. Appl. Statist.* **11**, 898–918.
HASTIE, T. J., TIBSHIRANI, R. J. & FRIEDMAN, J. H. (2009). *The Elements of Statistical Learning*. New York: Springer, 2nd ed.
HORVÁTH, L. & KOKOSZKA, P. (2012). *Inference for Functional Data with Applications*. New York: Springer.
KRAUS, D. (2015). Components and completion of partially observed functional data. *J. R. Statist. Soc.* B **77**, 777–801.
LIEBL, D. (2013). Modeling and forecasting electricity spot prices: A functional data perspective. *Ann. Appl. Statist.* **7**, 1562–92.
LINGJÆRDE, O. C. & CHRISTOPHERSEN, N. (2000). Shrinkage structure of partial least squares. *Scand. J. Statist.* **27**, 459–73.
PHATAK, A. & DE HOOG, F. (2002). Exploiting the connection between PLS, Lanczos methods and conjugate gradients: Alternative proofs of some properties of PLS. *J. Chemomet.* **16**, 361–7.
PINI, A. & VANTINI, S. (2016). The interval testing procedure: A general framework for inference in functional data analysis. *Biometrics* **72**, 835–45.
RAMSAY, J. O. & SILVERMAN, B. W. (2005). *Functional Data Analysis*. New York: Springer, 2nd ed.
SANGALLI, L. M., SECCHI, P. & VANTINI, S. (2014a). Analysis of AneuRisk65 data: k-mean alignment. *Electron. J. Statist.* **8**, 1891–904.
SANGALLI, L. M., SECCHI, P. & VANTINI, S. (2014b). AneuRisk65: A dataset of three-dimensional cerebral vascular geometries. *Electron. J. Statist.* **8**, 1879–90.
SANGALLI, L. M., SECCHI, P., VANTINI, S. & VENEZIANI, A. (2009a). A case study in exploratory functional data analysis: Geometrical features of the internal carotid artery. *J. Am. Statist. Assoc.* **104**, 37–48.
SANGALLI, L. M., SECCHI, P., VANTINI, S. & VENEZIANI, A. (2009b). Efficient estimation of three-dimensional curves and their derivatives by free-knot regression splines, applied to the analysis of inner carotid artery centrelines. *J. R. Statist. Soc.* C **58**, 285–306.
STEFANUCCI, M., SANGALLI, L. M. & BRUTTI, P. (2018). PCA-based discrimination of partially observed functional data, with an application to AneuRisk65 data set. *Statist. Neer.* **72**, 246–64.

[*Received on* 22 *August* 2017. *Editorial decision on* 2 *August* 2018]

Biometrika, *pp*. 1–5
Printed in Great Britain

Supplementary material for Classification of functional fragments by regularized linear classifiers with domain selection

BY DAVID KRAUS

Department of Mathematics and Statistics, Masaryk University, Kotlářská 2, 611 37 Brno, Czech Republic
david.kraus@mail.muni.cz

AND MARCO STEFANUCCI

Department of Statistical Sciences, Sapienza University of Rome, Piazzale Aldo Moro 5, 00185 Roma, Italy
marco.stefanucci@uniroma1.it

SUMMARY

The Supplementary Material provides the derivation of classifiers under unequal prior class probabilities, algorithmic details of cross-validation and additional simulation and real data results.

Some key words: Classification; Conjugate gradients; Domain selection; Functional data; Partial observation; Regularization; Ridge method.

S1. DERIVATIONS UNDER UNEQUAL PRIOR CLASS PROBABILITIES

Let π_j be the prior probability of class j ($j = 0, 1$). The optimal classifier based on the one-dimensional projection $\langle X, \psi \rangle$ assigns X to the class $C_\psi(X)$ given by

$$\begin{aligned} C_\psi(X) &= 1_{\{\pi_1 f_{\psi,1}(\langle X,\psi\rangle) > \pi_0 f_{\psi,0}(\langle X,\psi\rangle)\}} \\ &= 1_{\{\langle X-\mu_0,\psi\rangle^2 - \langle X-\mu_1,\psi\rangle^2 > 2\langle \psi, \mathscr{R}\psi\rangle \log(\pi_0/\pi_1)\}} \\ &= 1_{\{\langle X-\bar{\mu},\psi\rangle\langle \mu,\psi\rangle > \langle \psi,\mathscr{R}\psi\rangle \log(\pi_0/\pi_1)\}}, \end{aligned}$$

where $\bar{\mu} = (\mu_0 + \mu_1)/2$ and $\mu = \mu_1 - \mu_0$. The effect of unequal prior class probabilities is a shift of the decision boundary and the classifier is invariant with respect to multiplication of ψ by a non-zero constant.

Due to the fact that $\langle X - \bar{\mu}, \psi \rangle = \langle X - \mu_0, \psi \rangle - \langle \mu, \psi \rangle/2 = \langle X - \mu_1, \psi \rangle + \langle \mu, \psi \rangle/2$, the misclassification probability for an observation coming from class 0 or 1 with probabilities π_0,

π_1 is

$$\begin{aligned}
&\pi_0 P_0\{C_\psi(X)=1\}+\pi_1 P_1\{C_\psi(X)=0\}\\
&\quad=\pi_0 P_0\{(\langle X-\mu_0,\psi\rangle-\langle\mu,\psi\rangle/2)\langle\mu,\psi\rangle>\langle\psi,\mathscr{R}\psi\rangle\log(\pi_0/\pi_1)\}\\
&\qquad+\pi_1 P_1\{(\langle X-\mu_1,\psi\rangle+\langle\mu,\psi\rangle/2)\langle\mu,\psi\rangle<\langle\psi,\mathscr{R}\psi\rangle\log(\pi_0/\pi_1)\}\\
&\quad=\pi_0 P_0\left\{\frac{\langle X-\mu_0,\psi\rangle}{\langle\psi,\mathscr{R}\psi\rangle^{1/2}}>\frac{\langle\psi,\mathscr{R}\psi\rangle^{1/2}}{|\langle\mu,\psi\rangle|}\log(\pi_0/\pi_1)+\frac{|\langle\mu,\psi\rangle|}{2\langle\psi,\mathscr{R}\psi\rangle^{1/2}}\right\}\\
&\qquad+\pi_1 P_1\left\{\frac{\langle X-\mu_1,\psi\rangle}{\langle\psi,\mathscr{R}\psi\rangle^{1/2}}<\frac{\langle\psi,\mathscr{R}\psi\rangle^{1/2}}{|\langle\mu,\psi\rangle|}\log(\pi_0/\pi_1)-\frac{|\langle\mu,\psi\rangle|}{2\langle\psi,\mathscr{R}\psi\rangle^{1/2}}\right\}\\
&\quad=\pi_0\left[1-\Phi\left\{\frac{\langle\psi,\mathscr{R}\psi\rangle^{1/2}}{|\langle\mu,\psi\rangle|}\log(\pi_0/\pi_1)+\frac{|\langle\mu,\psi\rangle|}{2\langle\psi,\mathscr{R}\psi\rangle^{1/2}}\right\}\right]\\
&\qquad+\pi_1\Phi\left\{\frac{\langle\psi,\mathscr{R}\psi\rangle^{1/2}}{|\langle\mu,\psi\rangle|}\log(\pi_0/\pi_1)-\frac{|\langle\mu,\psi\rangle|}{2\langle\psi,\mathscr{R}\psi\rangle^{1/2}}\right\}.
\end{aligned}$$

Since the function

$$\pi_0[1-\Phi\{z^{-1}\log(\pi_0/\pi_1)+z/2\}]+\pi_1\Phi\{z^{-1}\log(\pi_0/\pi_1)-z/2\}$$

is decreasing in $z>0$, the minimization of the misclassification probability is equivalent to the maximization of

$$\frac{|\langle\mu,\psi\rangle|}{\langle\psi,\mathscr{R}\psi\rangle^{1/2}}$$

like in the case of equal prior probabilities discussed in the main body of the paper. If $\|\mathscr{R}^{-1/2}\mu\|<\infty$, the upper bound for the above fraction is $\|\mathscr{R}^{-1/2}\mu\|$ and the corresponding misclassification probability equals

$$\pi_0\left[1-\Phi\left\{\frac{\log(\pi_0/\pi_1)}{\|\mathscr{R}^{-1/2}\mu\|}+\frac{\|\mathscr{R}^{-1/2}\mu\|}{2}\right\}\right]+\pi_1\Phi\left\{\frac{\log(\pi_0/\pi_1)}{\|\mathscr{R}^{-1/2}\mu\|}-\frac{\|\mathscr{R}^{-1/2}\mu\|}{2}\right\}.$$

When $\|\mathscr{R}^{-1/2}\mu\|<\infty$, that is, when the Gaussian measures with means μ_1, μ_2 and covariance $\mathscr{R}$ are mutually absolutely continuous, this is the optimal misclassification probability among all classifiers, i.e., the Bayes error, as shown in Theorem 2 in Berrendero et al. (2018). The Bayes error is achieved by $\psi=\mathscr{R}^{-1}\mu$, if $\|\mathscr{R}^{-1}\mu\|<\infty$.

We can proceed like in the case of equal probabilities and apply regularization techniques to the inverse problem $\mathscr{R}\psi=\mu$. All theoretical results presented for the case of equal probabilities can be restated and reproved with the above form of the optimal error rate for the general case, including in the situation with $\|\mathscr{R}^{-1/2}\mu\|=\infty$, in which case the optimal error rate is zero and the two Gaussian measure are mutually singular.

In the empirical version of the problem one either estimates the prior class probabilities by $n_j/(n_0+n_1)$ if the training sample can be seen as a sample from the mixture of populations with these probabilities, or uses some fixed values.

S2. SELECTION OF THE REGULARIZATION PARAMETER AND DOMAIN BY CROSS-VALIDATION

Given the target domain $\mathcal{I}$, regularization method and regularization parameter, Algorithm S1 describes the estimation of the misclassification probability by cross-validation.

Algorithm S1. Estimation of the misclassification probability by cross-validation

Set $V = \{(j,i) : j \in \{0,1\}, i \in \{1,\ldots,n_j\}, O_{ji} \supseteq \mathcal{I}\}$
Repeat for $(j,i) \in V$
  Estimate the mean and covariance function restricted to $\mathcal{I}$
    using all training functions except X_{ji}
  Estimate the projection direction $\hat{\psi}$ using the given regularization method
    and regularization parameter
  Apply $\hat{C}_{\hat{\psi}}$ to the restriction of X_{ji} to $\mathcal{I}$ and save the predicted class label to c_{ji}
  Set the misclassification indicator $\delta_{ji} = 1_{[c_{ji} \neq j]}$
Output $\sum_{(j,i)\in V} \delta_{ji}/|V|$

The misclassification probability is estimated for a grid of values of the regularization parameter using Algorithm S1. The value that minimizes the error is selected.

When selecting the domain as well, one repeats the above process for each candidate domain in place of $\mathcal{I}$.

Once the regularization parameter and possibly domain are selected, the classifier is re-estimated using all training curves and applied to the new curve X_{new}.

S3. Additional simulation results

S3·1. *Processes with non-smooth trajectories*

Fig. S1 presents simulation results to compare the behaviour of classifiers on the conjugate gradient, principal component and ridge regularization path for Gaussian processes with non-smooth trajectories. We considered the Ornstein–Uhlenbeck process with covariance function $\rho(s,t) = \exp(-|s-t|)$. We used the same configurations for the mean difference between the classes as in Subsection 5.1 in the main body of the paper, except in cases (v) and (vi), where the mean difference now was the first and tenth eigenfunction of the Ornstein–Uhlenbeck covariance kernel.

The main conclusion from Subsection 5.1 of the paper is still valid for this situation. All three regularization methods reach about the same best error rate but the conjugate gradient method does it with less degrees than the other methods. The principal component method appears to be less stable than in the case of the smooth process of Subsection 5.1 which can probably be explained by the increased error of the estimation of the eigenfunction.

S3·2. *Behaviour under different covariance operators in groups*

The methods presented in the paper are derived under the assumption of equal covariance operators in both groups. Fig. S2 shows simulation results when this assumption is violated. We used Gaussian processes with covariance function $\exp(-|s-t|^2/0.01)$ in one group and $\exp(-|s-t|)$ in the other group. We considered the same scenarios for the mean difference as in Subsection 5.1 in the paper, except for scenarios (v) and (vi), where the mean difference was the first and tenth eigenfunction of the mixture covariance $0.5\exp(-|s-t|^2/0.01) + 0.5\exp(-|s-t|)$.

We conclude that the findings of Subsection 5.1 are robust with respect to the assumption of equal covariance operators. The principal component classifier again appears to be the least preferable method. Moreover, the error rates in this situation with different covariances are between the error rates in situations in which the two groups both have one of the considered

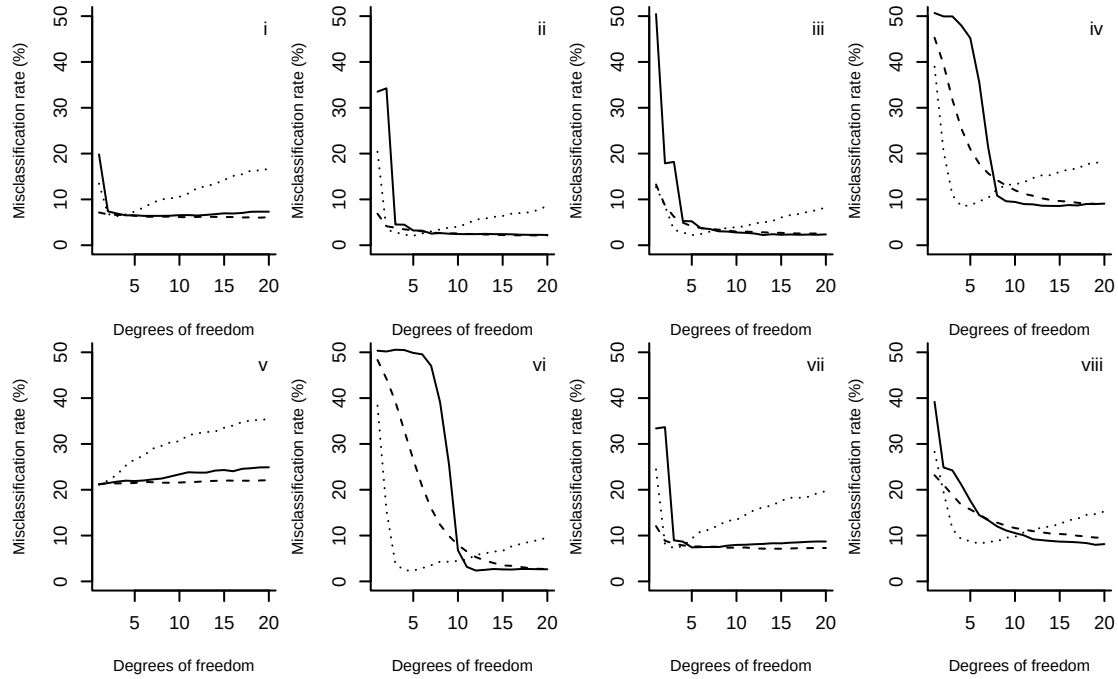

Fig. S1. Misclassification rate (%) versus degrees of freedom for non-smooth processes for different forms of $\mu(t)$, (i) linear, (ii) quadratic, (iii) cubic, (iv) sinusoidal, (v) first eigenfunction, (vi) tenth eigenfunction, (vii) symmetric beta, (viii) asymmetric beta, for principal component (solid), conjugate gradient (dotted) and ridge (dashed) classifiers.

covariance structures. Hence if there is a difference in the means, unequal covariances do not appear to have a serious negative effect on the performance of the classifiers.

S3·3. *Performance under increasing training sample size*

We performed additional simulations to study the effect of the training sample size. Fig. S3 presents results for the same settings as in Subsection 5.1 in the paper but with 100 training observations in each group, twice as many as in the paper.

Overall, the misclassification rates in Fig. S3 are slightly lower than in Fig. 1 in the paper due to the reduction of the estimation error. The difference is, however, small, suggesting that at the considered training sample sizes the estimation error is a relatively unimportant part of the total misclassification error.

S4. PERFORMANCE ON BENCHMARK DATA

We applied the proposed methods to two datasets, referred to as the wheat data and the phoneme data, on which Delaigle & Hall (2012) and Berrendero et al. (2018) previously compared functional classifiers. See these papers for references to the original sources of the data. We repeated with our classifiers their procedure which consisted of randomly splitting the data to the training set and test set, building the classifier on the training set and applying it to the test set to compute the proportion of misclassified curves, repeating this whole process two hundred times to estimate the misclassification rate. Table S1 reports the results. We can see that misclas-

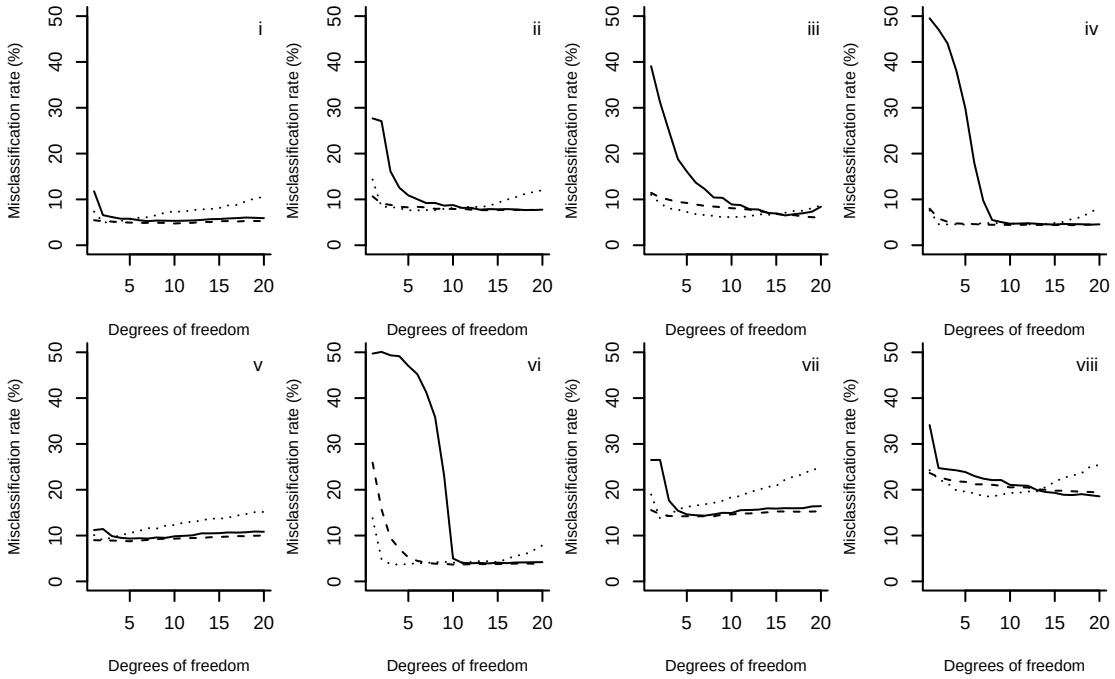

Fig. S2. Misclassification rate (%) versus degrees of freedom for processes with unequal covariance operators for different forms of $\mu(t)$, (i) linear, (ii) quadratic, (iii) cubic, (iv) sinusoidal, (v) first eigenfunction, (vi) tenth eigenfunction, (vii) symmetric beta, (viii) asymmetric beta, for principal component (solid), conjugate gradient (dotted) and ridge (dashed) classifiers.

sification rates decrease with increasing training sample size. Overall, on these data all classifiers appear to perform similarly and similarly to other methods studied in Delaigle & Hall (2012) and Berrendero et al. (2018). The ridge method might seem to perform slightly worse than the other two on the wheat data but in view of the standard errors we do not over-interpret this and other differences.

REFERENCES

BERRENDERO, J. R., CUEVAS, A. & TORRECILLA, J. L. (2018). On the use of reproducing kernel Hilbert spaces in functional classification. *Journal of the American Statistical Association* To appear.

DELAIGLE, A. & HALL, P. (2012). Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society. Series B. Statistical Methodology* **74**, 267–286.

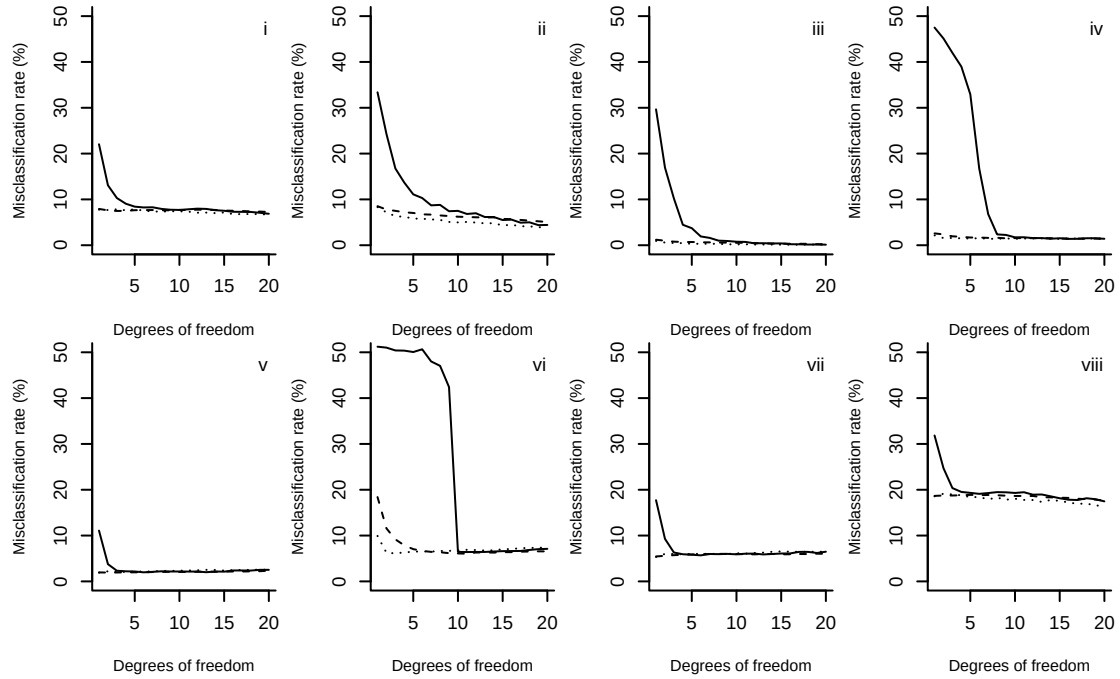

Fig. S3. Misclassification rate (%) versus degrees of freedom for 200 training observations for different forms of $\mu(t)$, (i) linear, (ii) quadratic, (iii) cubic, (iv) sinusoidal, (v) first eigenfunction, (vi) tenth eigenfunction, (vii) symmetric beta, (viii) asymmetric beta, for principal component (solid), conjugate gradient (dotted) and ridge (dashed) classifiers.

Table S1. *Misclassification rate (%) and its standard error achieved for wheat and phoneme data*

	Training sample size	PC	CG	R
Wheat	30	0.94 (1.89)	0.93 (2.06)	2.48 (2.79)
	50	0.36 (1.23)	0.58 (1.84)	1.73 (3.02)
Phoneme	30	24.1 (4.79)	23.3 (3.87)	22.1 (2.90)
	50	21.7 (2.76)	21.6 (2.12)	21.0 (2.07)
	100	20.1 (1.67)	20.1 (1.51)	20.1 (1.55)

PC, principal components; CG, conjugate gradients; R, ridge.

E. Inferential procedures for partially observed functional data

By David Kraus
Journal of Multivariate Analysis, 173:583–603, 2019
DOI: 10.1016/j.jmva.2019.05.002

Journal of Multivariate Analysis 173 (2019) 583–603

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Inferential procedures for partially observed functional data

David Kraus

Department of Mathematics and Statistics, Masaryk University, Kotlářská 2, 611 37 Brno, Czech Republic

ARTICLE INFO

Article history:
Received 19 September 2018
Received in revised form 14 May 2019
Accepted 15 May 2019
Available online 27 May 2019

AMS 2010 subject classifications:
primary 62M99
secondary 62G10

Keywords:
Bootstrap
Covariance operator
Functional data
K-sample test
Partial observation
Principal components

ABSTRACT

In functional data analysis it is usually assumed that all functions are completely, densely or sparsely observed on the same domain. Recent applications have brought attention to situations where each functional variable may be observed only on a subset of the domain while no information about the function is available on the complement. Various advanced methods for such partially observed functional data have already been developed but, interestingly, some essential methods, such as K-sample tests of equal means or covariances and confidence intervals for eigenvalues and eigenfunctions, are lacking. Without requiring any complete curves in the data, we derive asymptotic distributions of estimators of the mean function, covariance operator and eigenelements and construct hypothesis tests and confidence intervals. To overcome practical difficulties with storing large objects in computer memory, which arise due to partial observation, we use the nonparametric bootstrap approach. The proposed methods are investigated theoretically, in simulations and on a fragmentary functional data set from medical research.

1. Introduction

Functional data analysis is an established field [17,28,34,54] with well-developed methodologies for common types of observation of random curves, i.e., full (or dense) and sparse observation regimes. Due to new applications recent years have seen the emergence of a new type of observation of functional data, called functional fragments or partially observed functional data. For various examples see Bugni [6], Delaigle and Hall [14], Liebl [38], Gellar et al. [21], Goldberg et al. [23], Kraus [35], Delaigle and Hall [15], Gromenko et al. [24], Kneip and Liebl [32], Dawson and Müller [13], Mojirsheibani and Shaw [45], Stefanucci et al. [55], Descary and Panaretos [16], Kraus and Stefanucci [37] or Liebl and Rameseder [40].

Functional data are collections of observations of random elements of a function space, such as curves, images, surfaces, spatio-temporal fields. We consider random functions in a separable Hilbert space. Without loss of generality we work with the space $L^2([0,1])$ of square-integrable functions on $[0,1]$ equipped with inner product $\langle f,g\rangle = \int_0^1 f(t)g(t)dt$ and norm $\|f\| = \langle f,f\rangle^{1/2}$ but our results are applicable to more general spaces. Partially observed functional data consist of realizations of random functions that are not observed on the entire domain. Each function in the sample may be observed on a different subset of the domain and no information is available on the function values at arguments in the complement of this subset. For the ith functional variable $X_i \in L^2([0,1])$ there is a subset $O_i \subseteq [0,1]$ such that $X_i(t)$ is observed for $t \in O_i$ and not observed for $t \in [0,1]\setminus O_i$. The observation sets may be random, corresponding to data that are missing by happenstance, or non-random for designed experiments. We assume that the observation sets are mutually independent and independent of the curves. We refer to Liebl and Rameseder [40] for a study of the case of dependent missingness.

Although some advanced procedures, such as goodness-of-fit tests, regression, classification and reconstruction methods, have been developed for functional fragments, basic methods of inference about the fundamental characteristics

E-mail address: david.kraus@mail.muni.cz.

https://doi.org/10.1016/j.jmva.2019.05.002
0047-259X/

of functional variables are still missing. In particular, the asymptotic distribution of estimators of the mean function and covariance operator, K-sample tests of equal means or covariances, and confidence intervals for eigenvalues and eigenfunctions have not been studied yet in the setting of incomplete functions. Users who wish to perform these basic tasks currently have the only option: to omit the partially observed functions and apply existing procedures to the complete data only. This approach is not only clearly sub-optimal due to a possibly large loss of information and resulting decay of power and accuracy, but also hardly or totally inapplicable in situations where the data contain few or no complete curves.

In this paper, we address this deficiency of existing methodology and develop essential methods of inference about the mean and covariance structure of incomplete functional data. Random functions are characterized by the mean function $\mu = \mathrm{E}X$ and the covariance operator $\mathscr{R} : L^2([0,1]) \to L^2([0,1])$ defined as

$$(\mathscr{R}f)(\cdot) = \int_0^1 \rho(\cdot, t)f(t)dt, \quad f \in L^2([0,1]),$$

where $\rho(s,t) = \mathrm{cov}\{X(s), X(t)\}$ is the covariance function, assuming it exists. The covariance structure is best understood via principal component analysis or eigendecomposition of $\mathscr{R}$ in the form

$$\mathscr{R} = \sum_{m=1}^{\infty} \lambda_m \varphi_m \otimes \varphi_m,$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ are the eigenvalues, φ_m are the corresponding orthonormal eigenfunctions, and $(a \otimes b)f = \langle b, f \rangle a$ for $a, b, f \in L^2([0,1])$. For a theoretical background see, e.g., Bosq [5].

We find appropriate assumptions on the observation pattern that enable us to establish the asymptotic distribution of estimators of μ and $\mathscr{R}$. We develop tests for comparing the mean functions in K populations of functional data based on samples of fragments. Next, we propose several tests of equal covariance operators in K samples. We also construct confidence intervals for the eigenvalues and eigenfunctions estimated from incomplete data.

The practical implementation of methods for functional fragments is more complicated than for complete curves. The main difficulty is that temporal averaging (e.g., in inner products for dimension reduction) is impossible due to missing values. This leads to asymptotic distributions whose parameters follow rather complicated formulas. More importantly, since dimension reduction is not possible, the asymptotic distributions are, upon discretization, characterized by large objects (matrices or arrays) that are difficult or even impossible to store and manipulate in computer memory. The bootstrap turns out to be a solution to this problem. We provide specific algorithms for resampling functional fragments for mean and covariance testing and for confidence intervals for eigenelements.

In a simulation study we investigate the performance of the proposed tests, focusing in particular on the impact of missingness on the different tests and on the effect of the interplay between missingness and the form of differences between groups. The study shows that the proposed methods are superior to the currently only available approach based on omitting incomplete curves.

The proposed methodology is applied to a data set of temporal profiles of heart rate. The data consist of several hundred curves recorded by an automatic device during several hours in the evening during the transition from the day to night regime of heart activity. The profiles are not observed always available on the entire domain of interest because either the device did not measure or record measurements, or the person switched off the device. These fragmentary data were previously analysed in Kraus [35], where further details can be found.

Section 2 develops methods of inference about means in one and K samples. Section 3 deals with tests about covariance operators and with inference about principal components. Section 4 presents bootstrap approximations. Results of the simulation study and the data example are reported in Sections 5 and 6. In the Appendix we provide a central limit theorem for non-identically distributed functional variables needed in the asymptotic analysis of fragments, and proofs of all theorems. Additional simulation results and further results of the data analysis.

2. Mean inference from incomplete curves

2.1. Estimation of the mean function

In this section we focus on inference about the mean of functional data. Let us first consider estimation of the mean function μ of a homogeneous population. Let there be n independent functional observations. Each curve X_i, $i \in \{1, \ldots, n\}$ may be observed incompletely, with values known only for arguments in a subset $O_i \subseteq [0,1]$, with no information on the complement of O_i. The observation sets may be non-random or random. They are assumed to be mutually independent and independent of the curves and to consist of a finite union of intervals. We denote by $O_i(t)$ the indicator that the value of $X_i(t)$ is observed.

The mean function $\mu(t)$ can be estimated by the cross-sectional average of available observations

$$\hat{\mu}(t) = \frac{J(t)}{N(t)} \sum_{i=1}^{n} O_i(t)X_i(t),$$

where $N(t) = \sum_{i=1}^{n} O_i(t)$ is the number of available observations at time t and $J(t) = 1_{[N(t)>0]}$. The estimator is defined to be zero when $N(t) = 0$. In Kraus [35, Proposition 1] it was shown that under non-restrictive assumptions on the observation pattern the estimator $\hat{\mu}$ is consistent for the mean function μ, namely, it was proven that $\mathrm{E}\,\|\hat{\mu}-\mu\|^2 = O(n^{-1})$ as $n \to \infty$. We now aim to provide the asymptotic distribution of the estimator. The result will be essential in the derivation of the limiting distribution of the test statistics that we construct afterwards.

We denote $\pi_i(t) = \mathrm{E}\,O_i(t) = \Pr\{O_i(t) = 1\}$ and $\bar{\pi}(t) = n^{-1}\sum_{i=1}^{n}\pi_i(t)$. Furthermore, we denote by $U_i(s,t) = O_i(s)O_i(t)$ the indicator of observing the function values at the pair of arguments s and t, and define $\nu_i(s,t) = \mathrm{E}\,U_i(s,t)$, $\bar{\nu}(s,t) = n^{-1}\sum_{i=1}^{n}\nu_i(s,t)$ and $M(s,t) = \sum_{i=1}^{n} U_i(s,t)$. We need to introduce conditions on the observation pattern as follows.

Condition 1.

(a) Let there be a function $\pi(t)$ such that $\pi_0 = \inf_{t\in[0,1]}\pi(t) > 0$ and $\sup_{t\in[0,1]}|\bar{\pi}(t)-\pi(t)| \to 0$ for $n \to \infty$.
(b) Let there be a function $\nu(s,t)$ such that $\bar{\nu}(s,t) \to \nu(s,t)$ for all $s,t \in [0,1]$.
(c) Let there be a value $\nu_0 > 0$ such that for each $(s,t) \in [0,1]^2$ either $\nu(s,t) \geq \nu_0$ or $\nu(s,t) = 0$, and let the convergence $\sup_{(s,t)\in[0,1]^2}|\bar{\nu}(s,t)-\nu(s,t)| \to 0$ for $n \to \infty$ hold.

Condition (a) guarantees the consistency of the estimator $\hat{\mu}$, see Kraus [35]. Condition (b) is needed for the weak convergence of the estimator. Condition (c) is needed for consistent estimation of the covariance operator of the limiting distribution. We emphasize that no complete curves are required since these conditions may be satisfied even when the sample contains only fragments. We illustrate this attractive property in the simulation study in Section 5.

When the observation indicators $O_1,\ldots,O_n$ are identically distributed, then Condition (a) is satisfied if $\pi(t) = P\{O_i(t) = 1\}$ is bounded away from zero, Condition (b) is satisfied automatically and Condition (c) is satisfied if for each $(s,t) \in [0,1]^2$, $\nu(s,t) = P\{O_i(s) = 1, O_i(t) = 1\}$ is either bounded away from zero or equal to zero. The case of non-identically distributed observation indicators may be relevant, for example, for designed experiments in which non-random, designed observation sets may vary across subjects.

By $\|\cdot\|_2$ below we denote the Hilbert–Schmidt norm of an operator.

Theorem 1. *Assume that* $\mathrm{E}(\|X_1\|^2) < \infty$. *Let Conditions 1(a) and 1(b) hold. Then*

$$n^{1/2}\{\hat{\mu}(\cdot)-\mu(\cdot)\}, \quad N(\cdot)^{1/2}\{\hat{\mu}(\cdot)-\mu(\cdot)\}$$

are asymptotically distributed as mean zero Gaussian processes with covariance operators $\mathscr{K}'$, $\mathscr{K}$ *with kernels*

$$\kappa'(s,t) = \pi(s)^{-1}\pi(t)^{-1}\nu(s,t)\rho(s,t), \quad \kappa(s,t) = \pi(s)^{-1/2}\pi(t)^{-1/2}\nu(s,t)\rho(s,t),$$

respectively.

If, moreover, Definition 1(c) is satisfied, then $\mathscr{K}'$ *and* $\mathscr{K}$ *can be consistently estimated by the operators* $\hat{\mathscr{K}}'$ *and* $\hat{\mathscr{K}}$ *with kernels* $\hat{\kappa}'(s,t) = \hat{\pi}(s)^{-1}\hat{\pi}(t)^{-1}\hat{\nu}(s,t)\hat{\rho}(s,t)$ *and* $\hat{\kappa}(s,t) = \hat{\pi}(s)^{-1/2}\hat{\pi}(t)^{-1/2}\hat{\nu}(s,t)\hat{\rho}(s,t)$, *respectively, i.e.,* $\mathrm{E}\,\|\hat{\mathscr{K}}'-\mathscr{K}'\|_2^2 \to 0$ *and* $\mathrm{E}\,\|\hat{\mathscr{K}}-\mathscr{K}\|_2^2 \to 0$, *where* $\hat{\pi}(t) = N(t)/n$, $\hat{\nu}(s,t) = M(s,t)/n$, $\hat{\rho}(s,t)$ *is the empirical covariance based on all complete pairs of function values at* s,t, *and the value of the kernels is set to* 0 *whenever* $\hat{\pi}(s)$ *or* $\hat{\pi}(t)$ *is* 0.

The proof of this and other theorems is provided in the Appendix. Since the observable functional variables may be non-identically distributed due to possibly non-identically distributed observation indicators, the proof uses a central limit theorem for non-identically distributed functional random variables given in the Appendix.

Notice that the covariance kernels $\kappa'(s,t)$ and $\kappa(s,t)$ of the limiting distributions are zero when $\nu(s,t) = 0$ regardless of the value of $\rho(s,t)$. Therefore, it is not necessary to estimate $\rho(s,t)$ at such points. This is why Definition 1(c) does not require the function $\nu(s,t)$ to be bounded away from zero on the entire domain $[0,1]^2$ which is needed for the estimation of $\mathscr{R}$, as will be seen in Section 3, Definition 2(a). This means that the theorem applies also in the context of short fragments of curves considered, e.g., by Delaigle and Hall [15] or Descary and Panaretos [16], where each curve in the sample is observed on a short interval and no completely observed curves are available.

2.2. Tests of equality of means in several populations

Let us now consider K independent samples of functional data. Let the jth sample ($j \in \{1,\ldots,K\}$) consist of independent curves $X_{j1},\ldots,X_{jn_j}$ coming from the same distribution with mean μ_j and covariance operator $\mathscr{R}_j$. The functions may not be observed completely. It is assumed that for each function X_{ji} its values are available on a subset O_{ij}. Let the observation subsets be mutually independent and independent of the curves. Our aim is to test the null hypothesis that $\mu_1 = \cdots = \mu_K$ against the general alternative that the null does not hold. The literature on hypothesis testing for means of functional data is rich. See, for example, [2,3,8,9,18,28,39,43,49,52,53,56,57,59].

In the literature on complete functional samples there exist two main approaches to comparing mean functions. One is based on the L^2 distance between the means and one uses projections on finite dimensional subspaces.

The assessment of the hypothesis will be based on the contrasts of the group means and a null estimate of the common mean, i.e., on the differences $\hat{\mu}_j - \hat{\mu}$, $j \in \{1,\ldots,K\}$. Here we use $\hat{\mu}_j(t) = J_j(t)N_j(t)^{-1}\sum_{i=1}^{n_j} O_{ji}(t)X_{ji}(t)$, $j \in \{1,\ldots,K\}$, with

$N_j(t) = \sum_{i=1}^{n_j} O_{ji}(t)$ and $J_j(t) = 1_{[N_j(t)>0]}$. The estimator $\hat{\mu}$ is obtained as a weighted average of the group means in the form $\hat{\mu}(t) = \sum_{j=1}^{K} \hat{w}_j(t)\hat{\mu}_j(t)$ with weights

$$\hat{w}_j(t) = \frac{N_j(t)/\hat{r}_j^2}{\sum_{k=1}^{K} N_k(t)/\hat{r}_k^2},$$

where $\hat{r}_j^2 = \operatorname{tr} \hat{\mathscr{R}}_j$ is the trace of the estimated covariance operator in the jth sample (the estimators $\hat{\mathscr{R}}_j$ are discussed later). The role of the scaling by $\hat{r}_j^2$ is to account for possibly different covariance structures in the samples. This way of combining estimated means of heteroscedastic samples is inspired by the univariate case and its standard multivariate extensions. If the covariance structures are known to be the same in all samples, the factors $\hat{r}_j^2$ can be replaced by the trace of an estimator of the common covariance operator, which leads to the estimated mean based on the pooled sample of curves.

The first test we propose is inspired by the method of Cuevas et al. [9] who in the context of fully observed functional data developed an ANOVA test based on the L^2 norms of the contrasts of the group means and the pooled sample mean. A two-sample version of the test using the nonparametric bootstrap was proposed by Benko et al. [3]. Horváth et al. [29] studied a two-sample test based on the L^2 norm in the context of functional time series. The standardized contrast processes $N_j(\cdot)^{1/2}\{\hat{\mu}_j(\cdot) - \hat{\mu}(\cdot)\}/\hat{r}_j$, $j \in \{1, \dots, K\}$ can be collected into a K-dimensional vector that is a random element of the product space $\{L^2([0, 1])\}^K$ with inner product $\langle f, g\rangle = \sum_{j=1}^{K}\langle f_j, g_j\rangle$ for $f = (f_1, \dots, f_K)^\top$, $g = (g_1, \dots, g_K)^\top$. We use its L^2 norm as the test statistic, i.e., base the test on

$$T_{L^2} = \sum_{j=1}^{K} \|N_j(\cdot)^{1/2}\{\hat{\mu}_j(\cdot) - \hat{\mu}(\cdot)\}/\hat{r}_j\|^2 = \sum_{j=1}^{K} \int_0^1 N_j(t)\{\hat{\mu}_j(t) - \hat{\mu}(t)\}^2/\hat{r}_j^2 dt \tag{1}$$

and reject when the value of the statistic is significantly large.

Another main approach to curve mean testing uses dimension reduction. See, e.g., Aue et al. [2], Horváth and Kokoszka [28] or Horváth et al. [29]. The idea is to focus on a finite number of important features of the infinite-dimensional data. The functional observations are projected on a finite-dimensional subspace and multivariate ANOVA or a similar multivariate procedure is applied to the resulting vectors of Fourier scores. This strategy is not directly applicable in the situation of incompletely observed curves because, unlike in the fully observed case, Fourier scores of functional fragments cannot be computed by numerical integration as inner products of the functional variable and the basis function since the functional variable is not available on the entire domain.

Let $\hat{\psi}_1, \dots, \hat{\psi}_d$ be some linearly independent functions in $L^2([0, 1])$. Without loss of generality we assume that they are orthonormal. These functions may be either deterministic or random (estimated from the data). In the construction of our projection tests we use Fourier scores of the standardized contrast processes with respect to the basis functions $\hat{\psi}_l$. We denote these scores $Q_{jl} = \langle N_j(\cdot)\{\hat{\mu}_j(\cdot) - \hat{\mu}(\cdot)\}, \hat{\psi}_l\rangle/(\hat{r}_j n_j^{1/2})$, $j \in \{1, \dots, K\}$, $l \in \{1, \dots, d\}$ and collect them in the score vector $Q = (Q_{11}, \dots, Q_{1d}, \dots, Q_{K1}, \dots, Q_{Kd})^\top$. The score statistic is the quadratic form

$$T_d = Q^\top \hat{V}^- Q, \tag{2}$$

where $\hat{V}^-$ is the Moore–Penrose pseudoinverse of the estimated $(Kd) \times (Kd)$ covariance matrix of Q whose entry on the position with index (jl, km) is

$$\hat{V}_{jl,km} = \langle \hat{\pi}_j^{1/2}\hat{\psi}_l, \hat{\mathscr{V}}_{jk}(\hat{\pi}_k^{1/2}\hat{\psi}_m)\rangle = \int_{[0,1]^2} \hat{\pi}_j(s)^{1/2}\hat{\psi}_l(s)\hat{v}_{jk}(s, t)\hat{\psi}_m(t)\hat{\pi}_k(t)^{1/2} ds dt$$

for $j, k \in \{1, \dots, K\}$, $l, m \in \{1, \dots, d\}$. Here $\hat{\mathscr{V}}_{jk}$ is the covariance operator with kernel

$$\hat{v}_{jk}(s, t) = \sum_{l=1}^{K} \hat{r}_j^{-1}\{\delta_{jl} - N_j(s)^{1/2}\hat{w}_l(s)N_l(s)^{-1/2}\}\hat{\kappa}_l(s, t)\{\delta_{kl} - N_k(t)^{1/2}\hat{w}_l(t)N_l(t)^{-1/2}\}\hat{r}_k^{-1}, \tag{3}$$

where δ_{jk} is the Kronecker delta. The test rejects for large values of T_d.

Analogously to the case of one group considered in Section 2.1, we denote for $j \in \{1, \dots, K\}$, $i \in \{1, \dots, n_j\}$ the following quantities characterizing the observation patterns in each group, $\pi_{ji}(t) = \mathrm{E}\, O_{ji}(t) = \Pr(O_{ji}(t) = 1)$, $\bar{\pi}_j(t) = n_j^{-1}\sum_{i=1}^{n_j}\pi_{ji}(t)$, $U_{ji}(s, t) = O_{ji}(s)O_{ji}(t)$, $\nu_{ji}(s, t) = \mathrm{E}\, U_{ji}(s, t)$, $\bar{\nu}_j(s, t) = n_j^{-1}\sum_{i=1}^{n_j}\nu_{ij}(s, t)$ and $M_j(s, t) = \sum_{i=1}^{n_j} U_{ji}(s, t)$. Under mild assumptions we obtain the asymptotic distribution of both test statistics.

Theorem 2. *For $j \in \{1, \dots, K\}$ assume that $n_j \to \infty$, $n_j/(n_1 + \cdots + n_K) \to a_j > 0$ and $\mathrm{E}\,\|X_{j1}\|^2 < \infty$. Let the observation patterns in each group satisfy Definition 1. Then under the null hypothesis of equal means we obtain the following results:*

(i) The test statistic T_{L^2} is asymptotically distributed as $\sum_{k=1}^{\infty}\gamma_k C_k$, where C_k are independent chi-square distributed variables with one degree of freedom and γ_k can be consistently estimated by the eigenvalues of the operator $\hat{\mathscr{V}}$ given in (3).

(ii) Assume that there exist linearly independent non-random functions $\psi_1, \dots, \psi_d$ such that $\|\hat{\psi}_l - \psi_l\| \xrightarrow{P} 0$ for $l \in \{1, \dots, d\}$. Then the test statistic T_d is asymptotically chi-square distributed with $(K - 1)d$ degrees of freedom.

The test statistic based on the L^2 norm is not distribution-free but the critical values can be obtained straightforwardly by simulation, provided that the eigenvalues of $\hat{\mathscr{V}}$ consistently estimate γ_k. Similarly, the consistency of $\hat{\mathscr{V}}$ (and hence of $\hat{V}$) is needed for the score statistic. The consistency of $\hat{\mathscr{V}}$ is guaranteed by Definition 1(c). It may sometimes happen that $M_j(s,t)$ is low for some s,t, making the estimator $\hat{\mathscr{V}}$ less reliable. For this reason, and also for computational reasons, to avoid the estimation of the limiting covariance one can use the bootstrap method, as we describe in Section 4.

In the literature on complete functional data, the most common choice of the basis functions for the projection test is derived from principal component analysis (see Horváth and Kokoszka [28] and references therein, or Fremdt et al. [19]). The approach uses several leading eigenfunctions of the pooled sample covariance operator. The motivation for this choice is the property that the first eigenfunctions capture the principal modes of variation, the most important features of random deviations of the functional variables from the mean. Another approach is to use a fixed set of basis functions, such as several elements of the Fourier basis of sines and cosines or several orthonormal Legendre polynomials.

For several reasons we prefer deterministic bases to the basis of eigenfunctions. One drawback of the latter approach is that the principal components of variability may be only weakly related or entirely unrelated (orthogonal) to the differences between the mean functions, resulting in a test that is weak or inconsistent against this alternative. It may of course happen that the deterministic functions we choose are orthogonal to the alternative too, or that the leading eigenfunctions capture the mean differences well. However, with fixed functions it is at least possible to say before the analysis which alternatives can be detected. With principal components it is not known beforehand which departures from the null can be captured because the eigenfunctions are usually unknown. Moreover, their property of capturing the largest portion of variability, which is typically the main argument for using them, is not exactly what one wishes in mean testing. In fact, one would rather wish to maximize the signal-to-noise ratio or non-centrality, which, for example, in the case of components with equal magnitude of means would mean to minimize variability. In reality, the true interplay between the magnitude of components of the mean difference and their variability is not known, and we, therefore, prefer fixed functions.

The choice of the number of basis functions is important with projection methods. For the approach using eigenfunctions, we follow the recommendation of Horváth et al. [29] to use the smallest number of components needed to explain at least 85% of the total variability. For the method using fixed functions, in light of the above discussion of the relation of the power and variability we do not base the choice of d on the explained variability. Instead, we can specify what shape differences we wish to detect and use the corresponding basis functions. For example, using just $d = 3$ Legendre polynomials describing constant, monotonic as well as convex or concave non-monotonic differences seems to be a good choice in many applications.

3. Covariance inference under partial observation

3.1. Asymptotics for the estimated covariance operator and principal components

Given a collection of independent realizations of curves $X_1, \dots, X_n$ with mean function μ and covariance operator $\mathscr{R}$ observed on subsets $O_1, \dots, O_n$, the covariance function $\rho(s,t)$ can be estimated by the empirical covariance using pairwise complete observations, that is, by

$$\hat{\rho}(s,t) = \frac{I(s,t)}{M(s,t)} \sum_{i=1}^{n} U_i(s,t)\{X_i(s) - \hat{\mu}_{st}(s)\}\{X_i(t) - \hat{\mu}_{st}(t)\},$$

where $I(s,t) = 1_{[M(s,t)>0]}$ and

$$\hat{\mu}_{st}(s) = \frac{1_{[M(s,t)>0]}}{M(s,t)} \sum_{i=1}^{n} U_i(s,t) X_i(s).$$

If $M(s,t) = 0$, we define $\hat{\rho}(s,t) = 0$ and $\hat{\mu}_{st}(s) = 0$. Under certain assumptions on the observation pattern, the operator $\hat{\mathscr{R}}$ with kernel $\hat{\rho}(s,t)$ was shown to be a consistent estimator of $\mathscr{R}$ in Kraus [35, Proposition 1].

In the theorem below we give the asymptotic distribution under a set of conditions for which we denote $E_i(s,t,u,v) = O_i(s)O_i(t)O_i(u)O_i(v)$, the indicator that the observation of X_i at points s,t,u,v is available, and set $\theta_i(s,t,u,v) = \Pr\{E_i(s,t,u,v) = 1\}$, $\bar{\theta}(s,t,u,v) = \sum_{i=1}^{n} \theta_i(s,t,u,v)/n$ and $L(s,t,u,v) = \sum_{i=1}^{n} E_i(s,t,u,v)$.

Condition 2.

(a) Let there be a function $\nu(s,t)$ such that $\nu_0 = \inf_{(s,t)\in[0,1]^2} \nu(s,t) > 0$ and $\sup_{(s,t)\in[0,1]^2} |\bar{\nu}(s,t) - \nu(s,t)| \to 0$ for $n \to \infty$.

(b) Let there be a function $\theta(s,t,u,v)$ such that $\bar{\theta}(s,t,u,v) \to \theta(s,t,u,v)$ for all $s,t,u,v \in [0,1]$.

(c) Let there be a value $\theta_0 > 0$ such that for each $(s,t,u,v) \in [0,1]^4$ either $\theta(s,t,u,v) \geq \theta_0$ or $\theta(s,t,u,v) = 0$, and let the convergence $\sup_{(s,t,u,v)\in[0,1]^4} |\bar{\theta}(s,t,u,v) - \theta(s,t,u,v)| \to 0$ for $n \to \infty$ hold.

Condition (a) means that there are enough observations at all pairs of arguments. The condition is needed for the consistency of $\hat{\mathscr{R}}$, see Kraus [35] for a proof under an essentially equivalent condition. Condition (b) guarantees the weak convergence in the theorem below, and the additional condition (c) guarantees that the covariance of the asymptotic distribution can be estimated. We stress that these conditions do not require that the data contain any complete curves. They may be satisfied even in situations, where all functional observations are fragmentary. When the observation indicators $O_1, \dots, O_n$ are identically distributed, then Condition (a) is satisfied if $\nu(t) = P\{O_i(s) = 1, O_i(t) = 1\}$ is bounded away from zero, Condition (b) is satisfied automatically and Condition (c) is satisfied if for each $(s, t, u, v) \in [0, 1]^4$, $\theta(s, t, u, v) = P\{O_i(s) = 1, O_i(t) = 1, O_i(u) = 1, O_i(v) = 1\}$ is either bounded away from zero or equal to zero.

Theorem 3. *Assume that* $\mathrm{E}(\|X_1\|^4) < \infty$. *Let Conditions 2(a) and 2(b) hold. Then* $n^{1/2}(\hat{\mathscr{R}} - \mathscr{R})$ *and the operator with kernel* $M(\cdot,\cdot)^{1/2}\{\hat{\rho}(\cdot,\cdot) - \rho(\cdot,\cdot)\}$ *are asymptotically distributed as mean zero Gaussian operators whose covariance operators* $\mathfrak{H}'$, $\mathfrak{H}$ *have kernels*

$$\eta'(s,t,u,v) = \nu(s,t)^{-1}\nu(u,v)^{-1}\theta(s,t,u,v)\{\zeta(s,t,u,v) - \rho(s,t)\rho(u,v)\},$$
$$\eta(s,t,u,v) = \nu(s,t)^{-1/2}\nu(u,v)^{-1/2}\theta(s,t,u,v)\{\zeta(s,t,u,v) - \rho(s,t)\rho(u,v)\},$$

respectively, where $\zeta(s,t,u,v) = \mathrm{E}[\{X(s) - \mu(s)\}\{X(t) - \mu(t)\}\{X(u) - \mu(u)\}\{X(v) - \mu(v)\}]$.

If, in addition, Definition 2(c) is satisfied, then $\mathfrak{H}'$ *and* $\mathfrak{H}$ *can be consistently estimated by the operators* $\hat{\mathfrak{H}}'$ *and* $\hat{\mathfrak{H}}$ *with kernels* $\hat{\eta}'(s,t,u,v) = \hat{\nu}(s,t)^{-1}\hat{\nu}(u,v)^{-1}\hat{\theta}(s,t,u,v)\{\hat{\zeta}(s,t,u,v) - \hat{\rho}(s,t)\hat{\rho}(u,v)\}$ *and* $\hat{\eta}(s,t,u,v) = \hat{\nu}(s,t)^{-1/2}\hat{\nu}(u,v)^{-1/2}\hat{\theta}(s,t,u,v)\{\hat{\zeta}(s,t,u,v) - \hat{\rho}(s,t)\hat{\rho}(u,v)\}$, *respectively, i.e.,* $\mathrm{E}\,\|\hat{\mathfrak{H}}' - \mathfrak{H}'\|_2^2 \to 0$ *and* $\mathrm{E}\,\|\hat{\mathfrak{H}} - \mathfrak{H}\|_2^2 \to 0$, *where* $\hat{\eta}'(s,t,u,v)$ *and* $\hat{\eta}(s,t,u,v)$ *are set to* 0 *whenever* $\hat{\nu}(s,t)$ *or* $\hat{\nu}(u,v)$ *is* 0, $\hat{\theta}(s,t,u,v) = L(s,t,u,v)/n$ *and* $\hat{\zeta}(s,t,u,v)$ *is the empirical fourth central moment of the functional random variable computed using all complete quadruples of function values at arguments* s, t, u, v.

The weak convergence in the theorem above is on the separable Hilbert space of Hilbert–Schmidt operators equipped with the Hilbert–Schmidt norm $\|\cdot\|_2$. The limiting covariance operator $\mathfrak{H}$ is an operator that maps a Hilbert–Schmidt operator $\mathscr{F}$ with kernel $f(u,v)$ to an operator with kernel $\int_0^1\int_0^1 \eta(s,t,u,v)f(u,v)dudv$, similarly for other objects in the theorem.

Next, we study the estimators $\hat{\lambda}_m$ and $\hat{\varphi}_m$ of the eigenvalues and eigenfunctions of $\mathscr{R}$. The estimators are obtained by the eigendecomposition of $\hat{\mathscr{R}}$. Their root-n consistency was established by Kraus [35, Proposition 2]. Here we find the approximate distribution of the fluctuation of the estimators around their true counterparts (with appropriate sign for the eigenfunctions as usual).

Theorem 4. *Assume that* $\mathrm{E}(\|X_1\|^4) < \infty$ *and* $\mathscr{R}$ *has eigenvalues with multiplicity* 1. *Let Conditions 2(a) and 2(b) hold. Denote by* $\mathscr{H}'^{\infty}$ *a random operator following the limiting Gaussian distribution of* $n^{1/2}(\hat{\mathscr{R}} - \mathscr{R})$ *with mean zero and covariance* $\mathfrak{H}'$ *given in Theorem 3. Then, for* $n \to \infty$, *we obtain the following results:*

(i) $n^{1/2}(\hat{\lambda}_m - \lambda_m)$ *is asymptotically distributed as* $\langle \mathscr{H}'^{\infty}\varphi_m, \varphi_m\rangle$, *which is a normal variable with mean zero and variance*

$$\int_{[0,1]^4} \varphi_m(s)\varphi_m(t)\eta'(s,t,u,v)\varphi_m(u)\varphi_m(v)dsdtdudv.$$

(ii) $n^{1/2}(\hat{\varphi}_m - \hat{s}_m\varphi_m)$, *where* $\hat{s}_m = \mathrm{sign}\langle\hat{\varphi}_m, \varphi_m\rangle$, *is asymptotically distributed as the Gaussian random function* $\mathscr{Q}_m\mathscr{H}'^{\infty}\varphi_m$, *where*

$$\mathscr{Q}_m = \sum_{\substack{k=1\\k\neq m}}^{\infty} \frac{\varphi_k \otimes \varphi_k}{\lambda_m - \lambda_k}.$$

The limiting covariance operator of $n^{1/2}(\hat{\varphi}_m - \hat{s}_m\varphi_m)$ *is*

$$\sum_{\substack{k=1\\k\neq m}}^{\infty}\sum_{\substack{l=1\\l\neq m}}^{\infty} \frac{\varphi_k \otimes \varphi_l}{(\lambda_m - \lambda_k)(\lambda_m - \lambda_l)} \int_{[0,1]^4} \varphi_k(s)\varphi_m(t)\eta'(s,t,u,v)\varphi_m(u)\varphi_l(v)dsdtdudv.$$

If, additionally, Definition 2(c) is satisfied, then the limiting variance and covariance above can be consistently estimated by plugging-in estimates from Theorem 3.

The theorem is proved in the Appendix with the help of perturbation theory. The theorem generalizes the classic results of Dauxois et al. [11] who considered completely observed functions. See Kokoszka and Reimherr [33] for related results for functional time series. In the case of complete Gaussian curves Dauxois et al. [11] showed that the limiting covariance structure of the empirical covariance operator simplifies [see also 46] which eventually leads to a simpler form of the limiting variance of the empirical eigenvalue, namely to $2\lambda_m^2$. No such simplification is in general possible in the case of incomplete curves, even if they are Gaussian. Therefore, to make inference about eigenvalues or eigenfunctions, e.g., to construct confidence intervals, one possibility is to estimate the function $\eta'(s,t,u,v)$ and use the complicated expressions above for the limiting covariance structure. In Section 4 we provide an alternative approach based on the bootstrap which enables to avoid the possibly unstable estimation of η' and computer memory demanding storage and manipulation with the estimate.

3.2. Testing the equality of covariance operators

We now study tests for equality of covariance operators of several populations. Let there be K independent samples of partially observed functions with mean μ_j and covariance $\mathscr{R}_j$ in the jth sample, as described in Section 2.2. We aim to test the null hypothesis that $\mathscr{R}_1 = \cdots = \mathscr{R}_K$ against the general alternative. The general problem of hypothesis testing for covariance operators was previously studied in various contexts by various methods. See, e.g., [3,4,7,20,25,26,30,31,36,44,46,49–51,57,58].

Tests of the null hypothesis of equal covariance operators can be based on the differences between the estimators $\hat{\mathscr{R}}_j$ and the null estimator $\hat{\mathscr{R}}$ which is the pooled covariance operator with kernel

$$\hat{\rho}(s,t) = \sum_{j=1}^{K} \hat{w}_j(s,t)\hat{\rho}_j(s,t),$$

where

$$\hat{w}_j(s,t) = \frac{M_j(s,t)}{\sum_{k=1}^{K} M_k(s,t)}.$$

The differences are expressed by the contrast operators with kernels $M_j(\cdot,\cdot)^{1/2}\{\hat{\rho}_j(\cdot,\cdot) - \hat{\rho}(\cdot,\cdot)\}$. We propose two types of tests measuring the importance of the contrasts: one approach is based on the Hilbert–Schmidt norm of the contrasts and one is based on their projections on a subspace.

The first approach is inspired by methods that were previously considered in the case of fully observed functions, e.g., by Boente et al. [4]. The importance of the contrasts is expressed by the Hilbert–Schmidt norm. The test statistic takes the form

$$S_{\mathrm{HS}} = \sum_{j=1}^{K} \|M_j(\cdot,\cdot)^{1/2}\{\hat{\rho}_j(\cdot,\cdot) - \hat{\rho}(\cdot,\cdot)\}\|_2^2 = \sum_{j=1}^{K} \int_{[0,1]^2} M_j(s,t)\{\hat{\rho}_j(s,t) - \hat{\rho}(s,t)\}^2 dsdt \tag{4}$$

(in this notation we identify kernels and the corresponding operators).

The second approach uses projections of the contrasts onto a finite-dimensional subspace of the space of Hilbert–Schmidt operators. This type of tests was used for complete functions in various settings, e.g., by Horváth et al. [27], Panaretos et al. [46], Panaretos et al. [47], Kraus and Panaretos [36], Fremdt et al. [20], and Jarušková [30]. It is natural to project on the subspace generated by the leading eigenfunctions of $\hat{\mathscr{R}}$ because they carry information about the object of interest, the covariance operator (unlike in the case of mean functions where we prefer to use a fixed basis for the projection test). Let $\hat{\varphi}_1, \ldots, \hat{\varphi}_d$ be the first d eigenfunctions of $\hat{\mathscr{R}}$. Then the operators

$$\hat{\mathscr{U}}_{lm} = \begin{cases} \hat{\varphi}_l \otimes \hat{\varphi}_l, & l = m, \\ (\hat{\varphi}_l \otimes \hat{\varphi}_m + \hat{\varphi}_m \otimes \hat{\varphi}_l)/2^{1/2}, & l < m \end{cases}$$

with kernels $\hat{u}_{ll}(s,t) = \hat{\varphi}_l(s)\hat{\varphi}_l(t)$ and $\hat{u}_{lm}(s,t) = \{\hat{\varphi}_l(s)\hat{\varphi}_m(t) + \hat{\varphi}_m(s)\hat{\varphi}_l(t)\}/2^{1/2}$, $l < m$ form an orthonormal basis of a $d(d+1)/2$-dimensional subspace of $\mathrm{HS}(L^2([0,1]))$. The Fourier coefficients of the projection of the jth standardized contrast on this subspace are

$$R_{jlm} = \langle M_j(\cdot,\cdot)\{\hat{\rho}_j(\cdot,\cdot) - \hat{\rho}(\cdot,\cdot)\}/n_j^{1/2}, \hat{\mathscr{U}}_{lm}\rangle = \int_{[0,1]^2} M_j(s,t)\{\hat{\rho}_j(s,t) - \hat{\rho}(s,t)\}\hat{u}_{lm}(s,t)dsdt/n_j^{1/2}. \tag{5}$$

Denote by R the $Kd(d+1)/2$-dimensional score vector with components R_{jlm}, $j \in \{1,\ldots,K\}$, $1 \le l \le m \le d$. The test statistic measures the size of the projection of the contrast operators on the subspace. It takes the form

$$S_d = R\hat{W}^{-}R, \tag{6}$$

where $\hat{W}^{-}$ is the Moore–Penrose pseudoinverse of the estimator of the asymptotic covariance matrix whose entry with indices (jlm, kpq) is

$$\begin{aligned} \hat{W}_{jlm,kpq} &= \langle \hat{v}_j(\cdot,\cdot)^{1/2}\hat{u}_{lm}(\cdot,\cdot), \hat{\mathfrak{B}}_{jk}\{\hat{v}_k(\cdot,\cdot)^{1/2}\hat{u}_{pq}(\cdot,\cdot)\}\rangle \\ &= \int_{[0,1]^4} \hat{v}_j(s,t)^{1/2}\hat{u}_{lm}(s,t)\hat{\beta}_{jk}(s,t,u,v)\hat{u}_{pq}(u,v)\hat{v}_k(u,v)^{1/2}dsdtdudv, \end{aligned} \tag{7}$$

$j,k = 1,\ldots,K$, $1 \le l \le m \le d$, $1 \le p \le q \le d$. The kernel of $\hat{\mathfrak{B}}_{jk}$ is

$$\begin{aligned} \hat{\beta}_{jk}(s,t,u,v) = \sum_{l=1}^{K} &\{\delta_{jl} - M_j(s,t)^{1/2}\hat{w}_l(s,t)M_l(s,t)^{-1/2}\}\hat{\eta}_l(s,t,u,v) \\ &\times \{\delta_{kl} - M_k(u,v)^{1/2}\hat{w}_l(u,v)M_l(u,v)^{-1/2}\}. \end{aligned} \tag{8}$$

We now give the asymptotic distribution of the Hilbert–Schmidt and projection statistics.

Theorem 5. *For $j \in \{1, \dots, K\}$ assume that $n_j \to \infty$, $n_j/(n_1 + \cdots + n_K) \to a_j > 0$, $\mathrm{E}\,\|X_{j1}\|^4 < \infty$ and all eigenvalues of $\mathscr{R}_j$ have multiplicity 1. Let the observation patterns in each group satisfy Definition 2. Then under the null hypothesis of equal covariance operators we obtain the following results:*

(i) The test statistic S_{HS} is asymptotically distributed as $\sum_{k=1}^{\infty} \delta_k C_k$, where C_k are independent chi-square distributed variables with one degree of freedom and δ_k can be consistently estimated by the eigenvalues of the operator $\hat{\mathfrak{B}}$ given in (8).

(ii) The test statistic S_d is asymptotically chi-square distributed with $(K-1)d(d+1)/2$ degrees of freedom.

The asymptotic distribution of S_{HS} can be approximated by simulation like in Boente et al. [4]. Section 4 presents a practical bootstrap implementation of these tests in which it is not necessary to compute the operator $\hat{\mathfrak{B}}$.

Tests based directly on covariance operators are not the only option. As an alternative we explore the approach of Pigoli et al. [50] who argue that although covariance operators are contained in the Hilbert space of Hilbert–Schmidt operators, they do not form a linear subspace, and propose other distances than those based on the difference of covariances, such as the Procrustes distance and the square root distance. This direction of research was further investigated by Cabassi et al. [7] and Masarotto [44]. One of the proposals of Pigoli et al. [50] was to use the Hilbert–Schmidt distance between square root covariance operators $d_{\mathrm{sqrt}}(\mathscr{R}_1, \mathscr{R}_2) = \|\mathscr{R}_1^{1/2} - \mathscr{R}_2^{1/2}\|_2$. They report good power results for a two-sample test of equal covariances in the setting of complete functions based on this distance between estimated operators, $d_{\mathrm{sqrt}}(\hat{\mathscr{R}}_1, \hat{\mathscr{R}}_2)$. We extend this approach to K samples consisting of partially observed functions.

Since the data may contain incomplete functions, the empirical covariance operators $\hat{\mathscr{R}}_j$ used before may have negative eigenvalues. To be able to work with empirical square root covariance operators, we need to modify the covariance estimators to ensure they are non-negative definite. We use

$$\hat{\mathscr{R}}_{j+} = \sum_{l=1}^{n_j} (\hat{\lambda}_{jl})_+ \hat{\varphi}_{jl} \otimes \hat{\varphi}_{jl},$$

where $(\hat{\lambda}_{jl})_+ = \max(\hat{\lambda}_{jl}, 0)$ is the positive part of the eigenvalue $\hat{\lambda}_{jl}$ of $\hat{\mathscr{R}}_j$ and $\hat{\varphi}_{jl}$ is the corresponding eigenfunction. As discussed in Kraus [35], negative eigenvalues are typically of small magnitude in comparison with leading eigenvalues and, therefore, are negligible in practice. For a test statistic, we need to use the distance d_{sqrt} to define a null estimator of $\mathscr{R}$ and contrasts between the group estimators $\hat{\mathscr{R}}_{j+}$ and the null estimator. The common covariance operator can be estimated by

$$\hat{\mathscr{R}}_{\mathrm{sqrt}} = \left(\frac{\sum_{j=1}^{K} n_j \hat{\mathscr{R}}_{j+}^{1/2}}{\sum_{j=1}^{K} n_j} \right)^2,$$

which is the weighted Fréchet mean of the group-specific operators, i.e., the minimizer with respect to $\mathscr{R}$ of $\sum_{j=1}^{K} n_j d_{\mathrm{sqrt}}(\hat{\mathscr{R}}_{j+}, \mathscr{R})^2$.

The attained minimum of this objective function,

$$S_{\mathrm{sqrt}} = \sum_{j=1}^{K} n_j d_{\mathrm{sqrt}}(\hat{\mathscr{R}}_{j+}, \hat{\mathscr{R}}_{\mathrm{sqrt}})^2 = \sum_{j=1}^{K} \|n_j^{1/2}(\hat{\mathscr{R}}_{j+}^{1/2} - \hat{\mathscr{R}}_{\mathrm{sqrt}}^{1/2})\|_2^2, \tag{9}$$

can serve as a test statistic for comparing covariance operators in K samples. The statistic summarizes the size of the contrasts between the group and null estimators of the square root covariance operator. Following Pigoli et al. [50] we use resampling to approximate the null distribution of the statistic.

Notice that the contrasts between the group and null estimators in S_{sqrt} and S_{HS} are weighted differently. In S_{HS} we weight the contrast kernels by $M_j(s,t)^{1/2}$ which in the fragmentary setting reflects the accuracy of the estimation of the covariance kernel at each point of $[0,1]^2$ due to the number of observations available at that point. In S_{sqrt} this would not be meaningful because the square root covariance operator is a function of the entire covariance operator and thus the accuracy of the estimation of the square root covariance kernel at one point depends also on the numbers of available observations at all other points. We therefore simply weight by $n_j^{1/2}$ reflecting the overall accuracy of the square root covariance estimator. Both S_{HS} and S_{sqrt} are the attained minimum of the corresponding objective functional that defines the null estimator.

4. Practical implementation and bootstrap approximations

Functional data procedures are practically implemented by discretization. Functional observations are evaluated at q points of a grid in the domain. Functions then correspond to q-vectors (possibly with missing values), operators on the function space correspond to $(q \times q)$-matrices and operators on operators correspond to four-way arrays with all dimensions q.

To make inference (tests and confidence intervals), one can use the asymptotic distributions found in the previous section. However, the implementation of such procedures would be excessively demanding in terms of computer memory,

especially in the case of covariance inference. For example, when the evaluation grid consists of $q = 100$ points, arrays such as the one corresponding to the fourth moment kernel $\zeta(s, t, u, v)$ contain $q^4 = 10^8$ entries. To compare covariances, e.g., in $K = 3$ samples, one would have to work with an array with $K^2q^4 = 9 \times 10^8$ entries whose size already approaches the memory limits of usual computers, even if symmetry is exploited. In the case of multivariate, spatial or image data the number of evaluation points q is typically much larger than for functions of a one-dimensional argument. Aston et al. [1] give an example of acoustic phonetic data with bivariate, time–frequency argument with $q = 8100$. In conclusion, the size of objects representing the asymptotic covariance structure for tests or confidence intervals may be far beyond memory limits.

Projection covariance tests for complete functions can avoid the computation, storage and manipulation with such large arrays by computing principal scores of each function with respect to the required low number d of eigenfunctions [20,27, 46,47]. The covariance matrix of the score then depends on easy-to-handle d-dimensional four-way arrays instead of large q-dimensional four-way arrays. This dimension reduction approach is not applicable in the case of incomplete functions because the principal scores $\langle X_{ji} - \hat{\mu}_j, \hat{\varphi}_m \rangle$ cannot be computed when X_{ji} is available only on a subset of its domain [they can only be predicted, see 35]. Therefore, even the computation of the projection test statistic (6) is difficult due the large arrays the matrix $\hat{W}$ depends on.

The computation of the Hilbert–Schmidt statistic (4) and the square root covariance statistic (9) does not involve large four-way arrays. However, to use the asymptotic distribution of S_{HS} (see Theorem 5) one needs to estimate the eigenvalues of an operator on operators. Upon discretization and vectorization, this leads to a large eigenproblem of dimension $(Kq^2) \times (Kq^2)$, e.g., $30\,000 \times 30\,000$ for $K = 3$, $q = 100$. Again, dimension reduction cannot be used due to incomplete functions.

To overcome these difficulties we use the bootstrap. For completely observed functional data bootstrap tests of equal mean functions or covariance operators were studied by Benko et al. [3] and Paparoditis and Sapatinas [48,49]. In our missing data setting, all bootstrap procedures consist of appropriate resampling of fragmentary curves, which means that each bootstrap sample is again a collection of partially observed functions. The proposed procedures enable to completely avoid the computation of each entry of the large four-way covariance array and the storage and decomposition of the whole array.

The implementation of the tests of equal means is described in Algorithm 1. To correctly reproduce the limiting distribution of the group mean estimators under the null, the resampling is done separately in each group of groupwise centred fragmentary observations. The stratification guarantees that neither the missingness patterns nor distributional characteristics of the functions beyond the means need to be equal in all groups. The L^2 statistic is computed directly for each bootstrap sample and the observed value is then compared with the resampled values. The direct computation of the projection test statistic from observed or resampled data would require the estimation of the covariance functions $\hat{v}_{jk}$ in (3), which may be memory demanding and possibly unstable in regions with few complete pairs. We avoid it by estimating the covariance matrix of the score vector from the resampled score vectors, calculating the quadratic form statistic using the observed score vector and the bootstrap estimate of its covariance matrix, and comparing it with its asymptotic chi-square distribution.

Algorithm 1 Bootstrap approximation for tests of equal mean functions

1: Calculate $\hat{\mu}_j$ from observed samples of fragments $X_{j1}, \ldots, X_{jn_j}$, $j = 1, \ldots, K$, and $\hat{\mu}$
2: Calculate the test statistic T_{L^2} and the score vector Q
3: Set $X_{ji0} = X_{ji} - \hat{\mu}_j + \hat{\mu}$
4: For $b = 1, \ldots, B$
5: For each $j = 1, \ldots, K$, sample with replacement from fragments $X_{j10}, \ldots, X_{jn_j0}$ to get fragments $X^*_{j10}, \ldots, X^*_{jn_j0}$
6: Calculate the statistic $T^{*(b)}_{L^2}$ and score vector $Q^{*(b)}$ from $X^*_{j10}, \ldots, X^*_{jn_j0}$, $j = 1, \ldots, K$
7: Approximate the p-value of the L^2-test using T_{L^2} and $T^{*(1)}_{L^2}, \ldots, T^{*(B)}_{L^2}$
8: Calculate the empirical covariance matrix $\hat{V}^*$ of $Q^{*(1)}, \ldots, Q^{*(B)}$ and the statistic $T_d = Q^\top \hat{V}^{*-} Q$
9: Approximate the p-value of the projection test using T_d and the $\chi^2_{(K-1)d}$ distribution

Algorithm 2 describes the bootstrap implementation of confidence intervals for eigenelements. Resampling is applied to fragments and eigenelements are computed. The resampled eigenfunction is possibly reflected about zero so that its sign agrees with that of the observed data empirical eigenfunction. Standard methods of construction of confidence intervals can then be used. Since we again wish to avoid the calculation of variance estimates of eigenelements (see Theorem 4), we use the normal or basic bootstrap method [12, Chapter 5]. Intervals for eigenvalues are constructed on the logarithmic scale and untransformed. This is appropriate in general because in the case of completely observed Gaussian curves the asymptotic variance of $n^{1/2}(\hat{\lambda}_m - \lambda_m)$ is $2\lambda^2_m$ and thus the log-transformation approximately stabilizes variance.

Bootstrap covariance testing is described in Algorithm 3. Unlike in the case of mean testing, it is not possible to transform the data to the common null covariance structure and use stratified resampling. Bootstrap samples are instead

Algorithm 2 Bootstrap confidence intervals for eigenvalues and eigenfunctions

1: Calculate $\hat{\mathscr{R}}$ from the observed fragmentary functional data $X_1, \ldots, X_n$
2: Calculate the eigenvalues $\hat{\lambda}_m$ and eigenfunctions $\hat{\varphi}_m$ of $\hat{\mathscr{R}}$
3: For $b = 1, \ldots, B$
4: Sample with replacement from fragments $X_1, \ldots, X_n$ to get fragments $X_1^*, \ldots, X_n^*$
5: Calculate $\hat{\mathscr{R}}^*$ from $X_1^*, \ldots, X_n^*$ and its eigenvalues $\hat{\lambda}_m^{*(b)}$ and eigenfunctions $\hat{\varphi}_m^{*(b)}$
6: Replace $\hat{\varphi}_m^{*(b)}$ by $\operatorname{sign}\langle\hat{\varphi}_m^{*(b)}, \hat{\varphi}_m\rangle\hat{\varphi}_m^{*(b)}$
7: Based on $\hat{\lambda}_m^{*(b)}$, $\hat{\varphi}_m^{*(b)}$, $b = 1, \ldots, B$, calculate bootstrap confidence intervals for λ_m using log-transformation and pointwise bootstrap confidence intervals for $\varphi_m(t)$

drawn from the pooled sample of groupwise centred fragments, similarly to Paparoditis and Sapatinas [49, Subsection 2.2] for complete curves. Then, under the null hypothesis, if characteristics of observation patterns (θ_j) and fourth order moments (ζ_j) are the same in all groups, the pooled resampling asymptotically replicates the limiting distributions of interest. The Hilbert–Schmidt norm and square root covariance statistics are computed directly and the significance is decided upon by comparing the observed statistics with the resampled ones. Like in the case of mean testing, dimension reduction is impossible due to partial observation, and thus the computation of the covariance matrix of the score vector would require to compute large four-way arrays. Instead, the bootstrap is used to estimate the covariance matrix of the score and the quadratic statistic with this matrix is used.

Algorithm 3 Bootstrap approximation for tests of equal covariance operators

1: Calculate $\hat{\mu}_j$ and $\hat{\mathscr{R}}_j$ from observed samples of fragments $X_{j1}, \ldots, X_{jn_j}$, $j = 1, \ldots, K$, and $\hat{\mathscr{R}}$
2: Perform eigendecomposition of $\hat{\mathscr{R}}$, determine d and calculate $\hat{\mathscr{U}}_{lm}$, $1 \le l \le m \le d$
3: Calculate the test statistics S_{HS} and S_{sqrt} and the score vector R with respect to $\hat{\mathscr{U}}_{jm}$
4: Set $X_{ji0} = X_{ji} - \hat{\mu}_j$
5: For $b = 1, \ldots, B$
6: For each $j = 1, \ldots, K$, sample with replacement from the pooled collection of fragments X_{ji0}, $j = 1, \ldots, K$, $i = 1, \ldots, n_j$ to get fragments $X_{j10}^*, \ldots, X_{jn_j0}^*$
7: Calculate the statistics $S_{\mathrm{HS}}^{*(b)}$ and $S_{\mathrm{sqrt}}^{*(b)}$ and the score vector $R^{*(b)}$ with respect to $\hat{\mathscr{U}}_{jm}$ from $X_{j10}^*, \ldots, X_{jn_j0}^*$, $j = 1, \ldots, K$
8: Approximate the p-value of the Hilbert–Schmidt norm test using S_{HS} and $S_{\mathrm{HS}}^{*(1)}, \ldots, S_{\mathrm{HS}}^{*(B)}$ and the p-value of the square root covariance test using S_{sqrt} and $S_{\mathrm{sqrt}}^{*(1)}, \ldots, S_{\mathrm{sqrt}}^{*(B)}$
9: Calculate the empirical covariance matrix $\hat{W}^*$ of $R^{*(1)}, \ldots, R^{*(B)}$ and the statistic $S_d = R^\top \hat{W}^{*-} R$
10: Approximate the p-value of the projection test using S_d and the $\chi^2_{(K-1)d(d+1)/2}$ distribution

While we do not provide formal proofs of the validity of the bootstrap approximations, these could be obtained along the lines of the proofs in Paparoditis and Sapatinas [48] and Paparoditis and Sapatinas [49] using our asymptotic results (Theorems 1–5). Note that in our setting the observation sets might be non-identically distributed (e.g., in the case of designed experiments), and hence the bootstrap is applied to possibly non-identically distributed observed fragments. Their average characteristics, however, converge under Definitions 1 and 2. It is possible to use the bootstrap even with mildly non-identically distributed data, as discussed in the general context by Liu [41] who shows that if average moment characteristics of possibly non-identically distributed variables converge, the bootstrap is still applicable.

The use of the bootstrap for the square root covariance test is based on empirical evidence from simulation studies (Section 5 and the Supplementary Material). Its theoretical justification would require to first establish the asymptotic distribution of the estimated square root covariance operator, which is not available even in the case of completely observed curves [50].

5. Simulation results

The main goal of the study is to investigate the impact of partial observation on the performance of the different mean and covariance tests and compare the proposed tests using complete and incomplete curves with the simple approach using complete curves only.

We repeatedly generate three samples of curves of sizes $n_1 = 80$, $n_2 = 100$, $n_3 = 120$. Curves in the jth sample take the form

$$X(t) = \mu_j(t) + \lambda_{j0}^{1/2}\beta_{j0}h_j(t) + \sum_{k=1}^{20}\lambda_{jk}^{1/2}\beta_{jk}2^{1/2}\cos(k\pi t), \quad t \in [0, 1],$$

Table 1
Empirical rejection probability (in %) of the L^2 test, T_{L^2}, and projection test, T_d, of equal means. A dash indicates the same value as on the preceding row. The observation patterns (1)–(9) and mean configurations A–D are described in the text.

Observation pattern	Mean configuration							
	A		B		C		D	
	T_{L^2}	T_d	T_{L^2}	T_d	T_{L^2}	T_d	T_{L^2}	T_d
	Tests using complete and incomplete curves (proposed approach)							
(1)	5.6	6.2	69	60	49	56	52	63
(2)	5.4	6.7	59	52	28	29	38	50
(3)	–	–	–	–	50	56	44	62
(4)	4.4	6.5	66	58	51	57	51	62
(5)	–	–	–	–	44	49	50	58
(6)	5.4	7.1	58	51	50	55	42	49
(7)	–	–	–	–	28	34	37	42
(8)	5.4	5.8	55	47	34	37	42	48
(9)	5.4	7.8	37	40	20	23	26	34
	Tests using complete curves only (simple approach)							
(2), (3)	5.7	7.4	40	34	26	32	27	35
(4), (5)	3.6	7.4	28	27	18	26	19	28
(6), (7)	4.9	26.8	7	31	6	29	6	31
(8)	4.0	11.5	13	22	8	20	10	21

where $\beta_{jk}, j \in \{1, 2, 3\}, k \in \{0, \ldots, 20\}$ are mutually independent standard normal variables. Additional simulations with t_5 distributed coefficients are reported in the supplementary material. In all simulations we use 1000 repetitions of the test procedures, each based on 500 bootstrap samples. All tests are performed on the nominal level of 5%. All results have been computed in R 3.4.

The tests are applied to complete trajectories, observation pattern (1), and to fragments obtained by deleting missing periods following several random or nonrandom patterns. Observation patterns (2) and (3) are nonrandom: under pattern (2), the period $[0, 0.5]$ is removed from 50% of the curves in the first sample, 50% in the second sample and 60% in the third sample; pattern (3) is symmetric about 0.5, i.e., the period $[0.5, 1]$ instead of $[0, 0.5]$ is missing in the same subset of curves. Under patterns (4)–(7), a random missing period is generated independently for each curve and removed from the trajectory. First, we consider random missing periods taking the form $M = [C - E, C + E] \cap [0, 1]$ with $C = dU_1^{1/2}$ and $E = fU_2$, where U_1, U_2 are independent variables uniformly distributed on $[0, 1]$ and d, f are parameters. For missingness pattern (4) we set $d = 1.4$ and $f = 0.2$; this gives 39% of completely observed curves and the cross-sectional percentage of observed values decreases from 99% at time 0 to 79% at time 1. Pattern (5) is symmetric about 0.5. For pattern (6) we use the same model as for (4) and set $d = 1.2$ and $f = 0.5$; this leads to 7% of complete curves and the cross-sectional probability of observation is 94% at 0 and decreases to about 45% near 1. Pattern (7) is again obtained by reflecting pattern (6) about 0.5. Pattern (8) consists of observation periods generated independently for each curve in the form $O = [U_1, U_2] \cap [0, 1]$, where U_1, U_1 are independent variables uniformly distributed on $[a, C]$, $[C, 1 - a]$, respectively, $a = -0.3$ and C is uniformly distributed on $[0, 1]$; the percentage of complete curves in this case is 16% and the cross-sectional observation probability at 0.5 is 77% and decreases to 44% towards both endpoints of the domain. Finally, for pattern (9) curves are observed on random intervals generated as $[C - 0.2, C + 0.2] \cap [0, 1]$, where C is uniformly distributed in $[0, 1]$. This corresponds to fragments of curves of length at most 0.4, hence the datasets contain no complete curves, the median length of observed fragments is 0.3 and the cross-sectional probability of observation is 0.3 in the middle of the domain and decreases towards the endpoints, where it is 0.15.

In the study of mean tests four configurations of the mean functions are considered. Under configuration A the null hypothesis is satisfied: all mean functions are zero. Under configuration B the mean functions differ by a constant vertical shift: $\mu_1(t) = 0$, $\mu_2(t) = 0.18$, $\mu_3(t) = -0.1$. Under configuration C there are monotonic differences between the means: $\mu_1(t) = 0$, $\mu_2(t) = 0.35\exp(-4t)$, $\mu_3(t) = -0.25\exp(-3t)$. Under configuration D the means differ in a more complex, nonmonotonic way and they cross: $\mu_1(t) = 0$, $\mu_2(t) = 2t\exp(-3t)$, $\mu_3(t) = 0.1 - 8t^2\exp(-5t)$. We set $\lambda_{j0} = 0.5$, $\lambda_{jk} = 3^{-k}$ and $h_j(t) = 1$, that is, the covariance structure is the same in all three groups. Additional simulations with unequal covariance structures lead to similar results and are included in the Supplementary Material. We report in the first part of Table 1 the size and power of the L^2 test based on T_{L^2} given in (1) and of the projection test based on T_d given in (2) using $d = 3$ Legendre polynomials of order zero, one and two. Blank entries in the table correspond to situations where the true rejection probability is the same as in the entry above; such situations arise when the observation pattern is obtained by reflecting the preceding pattern and the processes $\{X(t) : t \in [0, 1]\}$ and the time-reversed processes $\{X(1 - t) : t \in [0, 1]\}$ have the same distribution.

We see in the first part of Table 1 that under the null hypothesis, configuration A, the rejection probability of the L^2 tests is close to the nominal level. The size of the projection test seems to be somewhat above the nominal level due to the sample size, especially under observation pattern (9), where the missingness rate is the highest. Our simulation study

Table 2
Empirical rejection probability (in %) of the Hilbert–Schmidt norm test, S_{HS}, projection test, S_d, and square root covariance test, S_{sqrt}, of equal covariance operators. A dash indicates the same value as on the preceding row. The observation patterns (1)–(5) and covariance configurations A–D are described in the text.

Observation pattern	Covariance configuration											
	A			B			C			D		
	S_{HS}	S_d	S_{sqrt}	S_{HS}	S_d	S_{sqrt}	S_{HS}	S_d	S_{sqrt}	S_{HS}	S_d	S_{sqrt}
	Tests using complete and incomplete curves (proposed approach)											
(1)	5.4	5.8	4.8	69	82	80	69	58	69	78	62	81
(2)	4.6	6.4	4.9	54	63	41	37	32	38	76	64	54
(3)	–	–	–	–	–	–	–	–	–	46	30	48
(4)	5.0	5.1	5.8	64	74	72	61	53	62	72	56	73
(5)	–	–	–	–	–	–	–	–	–	77	60	77
	Tests using complete curves only (simple approach)											
(2), (3)	4.1	7.3	4.6	32	38	41	33	28	34	45	30	47
(4), (5)	4.3	5.5	4.2	26	32	33	25	24	28	34	23	36

of power provides raw rejection probabilities in Table 1 and size-adjusted powers (using the method from Subsection 3.2 of Lloyd [42]) in Table S2 in the Supplementary Material. The possibility of size issues should be kept in mind in applications: especially in marginal cases, users should not simply compare p-values with a single threshold but rather carefully report them.

Under scenario B the L^2 test is more powerful than the projection method. The reason is that the projection method uses in addition to the constant basis function two other terms (linear and quadratic) that do not contribute to the detection of the constant difference between the means but on the other hand they increase the degrees of freedom and hence decrease the power. The L^2 method uses infinitely many directions in the space of alternatives but these redundant features are downweighted by the decreasing eigenvalues (the constant difference of means agrees with the constant leading eigenfunction which receives the highest weight in the L^2 statistic). Most partial observation patterns lead to a relatively small decrease of power because under this scenario the mean functions differ by a constant vertical shift which is a very simple, global feature that is easily detected even with reduced, fragmented data. The loss of power is largest under pattern (9), where also the reduction of observed data is considerably larger than under the other patterns.

Both tests have comparable power under scenario C. Both tests lose power under observation pattern (2) because a large portion of data is missing on the interval $[0, 0.5]$, where the difference between the means is the largest; on the other hand, the reflected pattern (3) does not lead to a loss of power because curves are missing only in $[0.5, 1]$, where the means do not differ much. A similar effect is seen under observation patterns (6) and (7).

Under scenario D the projection test seems to be slightly more powerful than the L^2 (even after the size adjustment in Table S2 in the Supplementary Material) because the nonmonotonic differences between the mean functions are well captured by both the first three Legendre polynomials and the first three eigenfunctions but the contribution of the latter is downweighted in the L^2 statistic whereas the projection statistic treats all three components equally.

The second part of Table 1 shows for each missingness pattern and mean configuration the performance of the tests applied to the subset of complete curves only. The complete curve approach would be the only possibility if the tests developed in this paper were not available. Results for the pairs of patterns (2) and (3), (4) and (5), (6) and (7) are presented on the same rows of the second part of the table because the subsets of complete curves are the same under both patterns in each pair. Pattern (9) is omitted because it contains no complete curves and hence inference is impossible without our methods. Under patterns (2) (or (3)) and (4) (or (5)), the use of complete curves only, which form 46% and 39%, respectively, of the whole sample, leads to a considerable loss of power in most situations. Configuration C under pattern (2) is an exception. Here removing incomplete curves does not decrease the power because they are observed on the subdomain $[0.5, 1]$, where the means do not differ much. Under patterns (6) (or (7)) and (8) there are only 7% and 16% complete curves, respectively. With such small sample sizes the projection test becomes unreliable in terms of level and the L^2 test loses almost all power.

Next, we study the behaviour of the tests for comparing covariance operators. Under all scenarios we generate mean zero trajectories. Configuration A satisfies the null hypothesis with $\lambda_{j0} = 0.5$, $\lambda_{jk} = 3^{-k}$ and $h_j(t) = 1$, $j \in \{1, 2, 3\}$. Under configuration B the same parameters are used except for the third sample where the overall scale is larger, namely $\lambda_{3,0} = 1.5 \times 0.5$ and $\lambda_{3,k} = 1.5 \times 3^{-k}$. Under scenario C the first two eigenvalues in the third sample are interchanged, i.e., $\lambda_{3,0} = 3^{-1}$, $\lambda_{3,1} = 0.5$ and $\lambda_{3,k} = 3^{-k}$, $k \in \{2, \ldots, 20\}$, otherwise the parameters are the same as in A. Scenario D differs from A in that we set $h_3(t) = 1$ for $t \in [0, 0.5]$ and $h_3(t) = 2.2^{1/2}$ for $t \in (0.5, 1]$. Table 2 shows the size and power of the Hilbert–Schmidt norm test based on S_{HS} in (4), projection test based on S_d in (6) with d selected to explain at least 85% of the total variability of the null covariance estimate, and square root covariance test based on S_{sqrt} in (9). Like before, entries where the true rejection probability equals the one above are left blank. We use only observation patterns (1)–(5). Under the other patterns the amount of missing information is too large for second order inference.

Under the null hypothesis, configuration A, the first part of Table 2 shows that the rejection probability of all tests is close to the nominal level under all missingness patterns, with the projection test being slightly above the level in some cases.

It is interesting to notice the different impact of missingness on the power in different situations. We report raw power in Table 2 and size-adjusted power in Table S4 in the Supplementary Material. While in many situations the loss of power due to missingness is similar for all three tests, in some situations the square root test appears to be more sensitive to missingness. For example under scenario B and missingness pattern (2), the square root covariance test loses almost half of its power relative to no missingness, much more than the other two tests. This can be explained by the fact that the square root covariance estimator depends on the estimator of the covariance kernel at all arguments which means that uncertainty due to missingness localized in a certain region in the domain, like under pattern (2), propagates. Similarly, under scenario D and pattern (2) the Hilbert–Schmidt and projection tests do not lose much power and the square root test does because the difference between the covariances is due to the differences of $h_j(t)$ for $t \in [0.5, 1]$ while missingness occurs for $t \in [0, 0.5]$. For these reasons, under the same scenario, pattern (3) leads to a larger loss of power than pattern (2) for the Hilbert–Schmidt and projection tests, whereas the loss of the square root covariance is not much higher than under pattern (2), where it was already high.

The second part of Table 2 shows results for tests applied to the subset of complete curves only. Like before, patterns (3) and (5) are shown on the same rows as patterns (2) and (4), respectively, because the subsets of complete curves are the same. We observe a large decrease of power in comparison with the power of the proposed tests in cases, where the neglected incomplete curves carry information on the difference between covariance operators. When the difference is mostly in the frequently missing region (e.g., configuration D, pattern (3)), removing incomplete curves affects the power much less.

These results highlight the usefulness of the proposed methods as an efficient, and often the only viable approach to testing with incomplete functions. In no situation the proposed methods behaved worse than the simple approach using complete curves only, and in many cases it behaved dramatically better. Additional results for non-Gaussian curves can be found in the Supplementary Material.

6. Application to partially observed heart rate temporal profiles

We illustrate our methods on curves describing the evolution of heart rate in 427 male participants in the period from 8 PM to 2 AM corresponding to the domain $[20, 26]$. The data come from the Swiss Kidney Project on Genes in Hypertension. There are three groups of persons according to their age: younger than 40 years (164 persons), between 40 and 65 (180), and older than 65 (83). The curves and their first derivative are plotted in Fig. 1. Although the percentage of observed values at each time or at each pair of time points is relatively high (Fig. 2), only 58% of the curves are complete.

Plots of the estimated mean functions in Fig. 1 indicate differences between the age groups both in terms the temporal profiles and their first derivative. We first compare the group means of heart rate profiles. The p-values of the L^2 test and projection test using three Legendre polynomials are 0.006 and less than 0.001, respectively, confirming the clearly visible differences. To compare the dynamics of heart rate during the transition between day and night we test whether the means of the first derivative differ. The L^2 and projection test have nearly zero p-values, meaning that the mean heart rate profiles differ between age groups more than by a vertical shift. The plots suggest it may be interesting to compare some pairs of groups. E.g., while the mean profiles of the middle and oldest group significantly differ ($p < 0.01$ for both tests), they appear to be approximately parallel. The difference between the derivatives is indeed insignificant ($p = 0.07$ for the L^2 test, $p = 0.09$ for the projection test).

Without the methods developed in this paper one would have to use complete curves only. There are 249 complete functions (43, 110 and 96 in the three age groups). The projection test still detects the differences between the three groups ($p = 0.008$) but the L^2 test loses significance ($p = 0.066$). When comparing the second and third group, the projection test now fails to detect the difference ($p = 0.13$) and the L^2 test gives a marginally significant result ($p = 0.048$). This can be explained by a loss of power seen in simulations because the removed incomplete curves are more often observed at earlier times, where also the difference between the two mean curves is more pronounced.

Estimates of the covariance function, eigenvalues and eigenfunctions of heart rate profiles and of their derivatives for each age group are plotted in Fig. 3 and Fig. 4. Further plots can be found in the supplementary document. The plots suggest some differences between the groups. The variance and covariance appears to be higher in younger participants, especially earlier in the time interval (during the day). We assess the significance of these differences using the proposed tests. For the projection test we consider up to three principal components (plotted in the supplementary document), which corresponds to the projection on a subspace of dimension six in the space covariance operators. Table 3 reports the p-values. None of the tests rejects the null hypothesis on usual significance levels. Similarly, pairwise comparisons provided no overwhelming evidence of differences. It is of course possible that there are differences between groups that may be detected with larger samples. To gain further insight into the structure of possible differences one can inspect the values of the standardized score components $R_{jlm}/\hat{W}^{1/2}_{jlm,jlm}$ (see (5) and (7)) whose graphical representation is provided in the supplement.

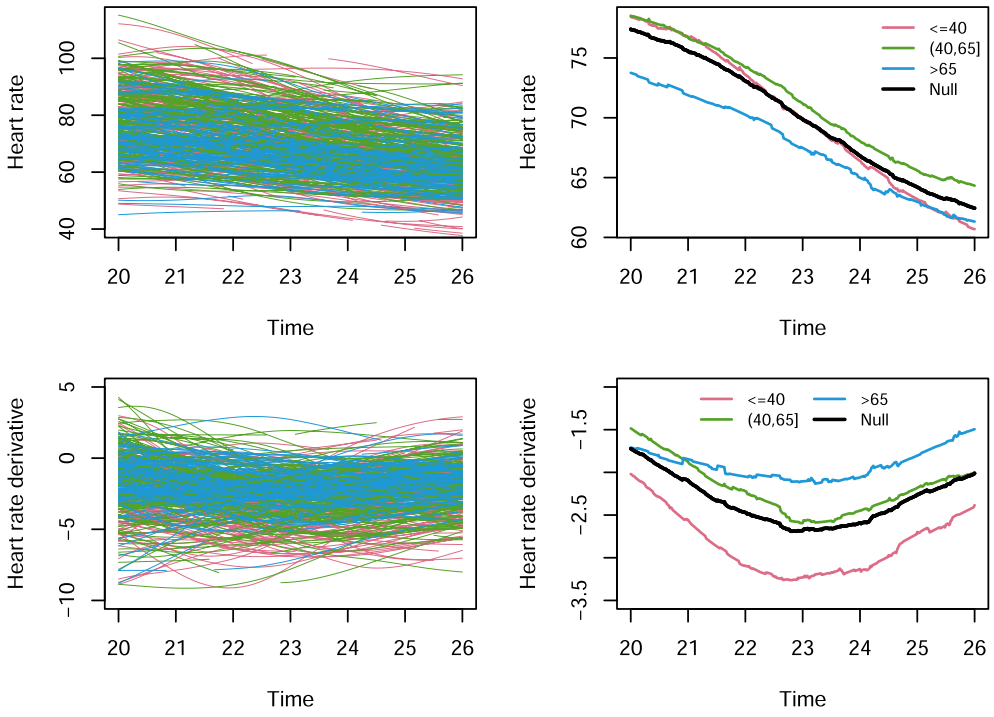

Fig. 1. Individual heart rate profiles and their first derivative (left panels) and the corresponding group-specific and null estimates of the mean (right panels).

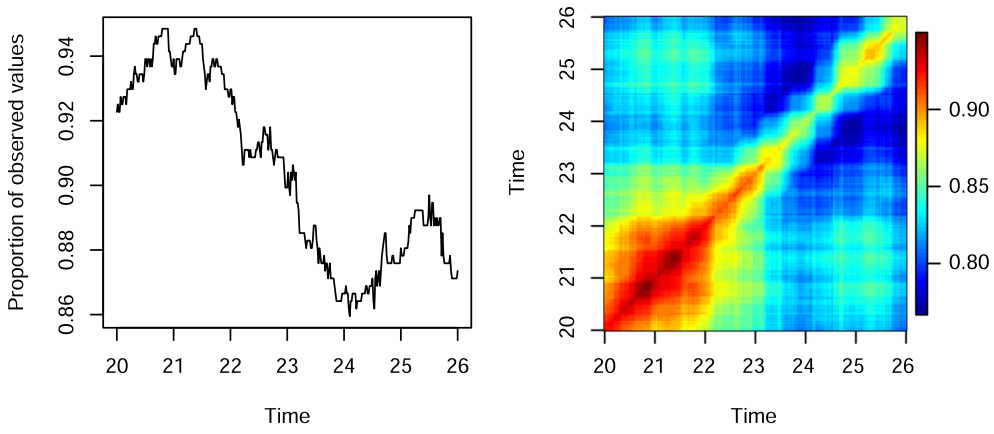

Fig. 2. Cross-sectional percentage of observed values (left) and percentage of pairwise complete observations (right).

Table 3
p-values of the Hilbert–Schmidt norm test, S_{HS}, the square root covariance test, S_{sqrt}, and the projection tests, S_d, with $d = 1, 2, 3$, for comparing covariance structures of heart rate profiles and of their first derivative in three age groups. The fraction of variance explained by the first d principal components of the null covariance estimate is indicated in parentheses.

	S_{HS}	S_{sqrt}	S_1	S_2	S_3
Curves	0.338	0.118	0.317 (88.2%)	0.439 (97.3%)	0.275 (99.1%)
First derivative	0.226	0.114	0.322 (62.6%)	0.131 (94.4%)	0.094 (98.7%)

Acknowledgments

We are grateful to all reviewers for their valuable comments and suggestions. This work was supported by the Czech Science Foundation under Grant GJ17-22950Y. Access to computing and storage facilities owned by parties and projects contributing to the MetaCentrum National Grid Infrastructure provided under the programme “Projects of Large Research, Development, and Innovations Infrastructures” (CESNET LM2015042) is greatly appreciated.

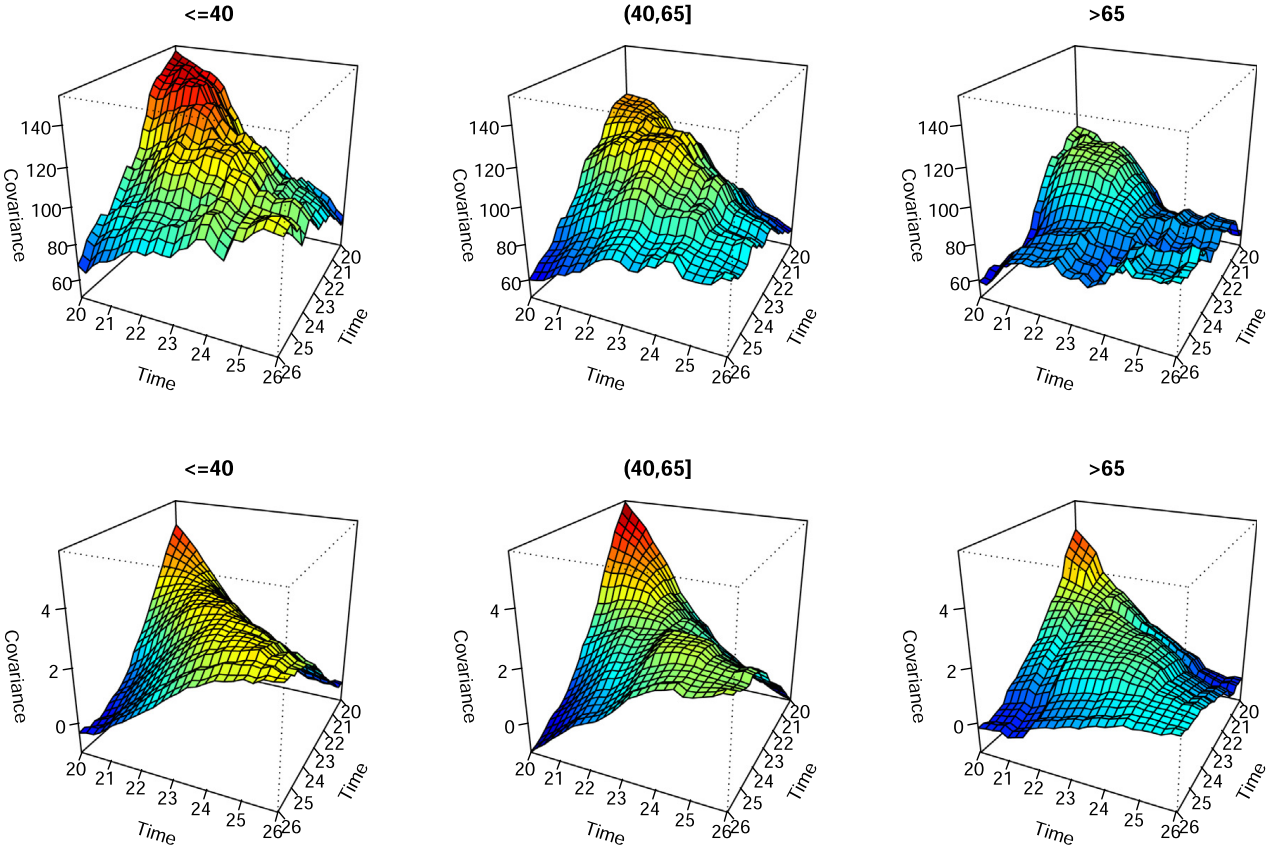

Fig. 3. Estimated covariance functions of heart rate profiles (top row) and of their derivatives (bottom row) in age groups.

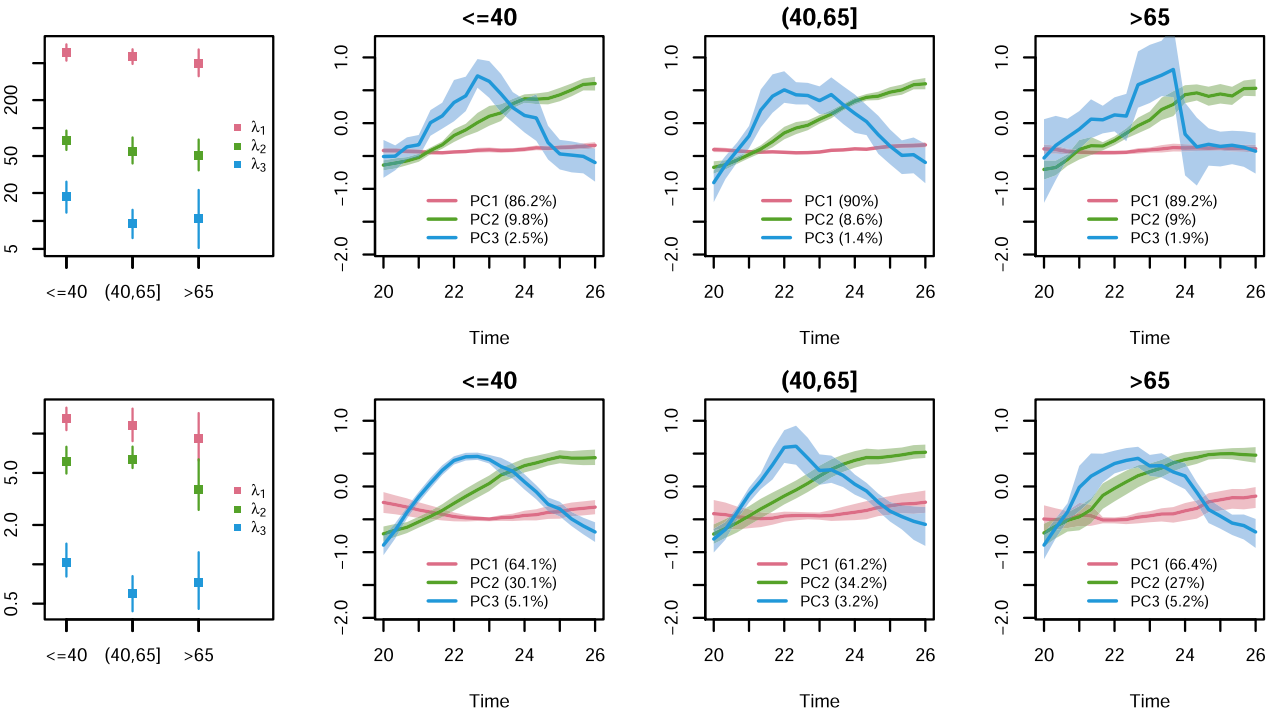

Fig. 4. Estimated eigenvalues and eigenfunctions of heart rate profiles (top row) and of their derivatives (bottom row) in age groups with pointwise 95% bootstrap confidence intervals.

Appendix A. A central limit theorem

We provide a general central limit theorem for independent but not necessarily identically distributed random elements of a separable Hilbert space. It is needed in the proofs, where non-identical distributions arise due to partial observation, but is of more general interest. It extends the standard result for independent identically distributed functional variables [5, Theorem 2.7] by relaxing the assumption of identical distributions and by considering triangular arrays. The notation $\|\cdot\|_\infty$ below means the operator norm.

Theorem 6. *Let Y_{ni}, $n \in \{1, 2, \dots\}$, $i \in \{1, \dots, n\}$ be random elements of a separable Hilbert space $\mathcal{H}$ with mean zero, $\mathrm{E}\,\|Y_{ni}\|^2 < \infty$ and covariance operators $\mathscr{C}_{ni}$. Let $Y_{n1}, \dots, Y_{nn}$ be mutually independent for each $n \in \{1, 2, \dots\}$. Denote $S_n = n^{-1/2} \sum_{i=1}^n Y_{ni}$ and $\mathscr{G}_n = n^{-1} \sum_{i=1}^n \mathscr{C}_{ni}$. Assume that*

(i) $\|\mathscr{G}_n - \mathscr{G}\|_\infty \to 0$ as $n \to \infty$ for some covariance operator $\mathscr{G}$,
(ii) for all $\varepsilon > 0$,

$$n^{-1} \sum_{i=1}^n \mathrm{E}(\|Y_{ni}\|^2 1_{[\|Y_{ni}\| > n^{1/2}\|\mathscr{G}_n\|_\infty \varepsilon]}) \to 0$$

as $n \to \infty$,
(iii) $\operatorname{tr} \mathscr{G}_n \to \operatorname{tr} \mathscr{G}$ as $n \to \infty$.

Then S_n converges in distribution to a Gaussian random element with mean zero and covariance operator $\mathscr{G}$.

Appendix B. Proofs

Proof of Theorem 1. We rewrite $N^{1/2}(\hat{\mu} - \mu) = \hat{\pi}^{1/2} n^{1/2}(\hat{\mu} - \mu)$. The main task is to establish the weak convergence of the process

$$n^{1/2}(\hat{\mu} - \mu) = \frac{1}{\pi} S_n + \left(\frac{J}{\hat{\pi}} - \frac{1}{\pi}\right) S_n + n^{1/2}(J - 1)\mu, \tag{B.1}$$

where $S_n = n^{-1/2} \sum_{i=1}^n O_i(X_i - \mu)$. We show that the first term on the right side of (B.1) converges in distribution to a mean zero Gaussian process with covariance operator with kernel $\pi(s)^{-1}\pi(t)^{-1}\nu(s,t)\rho(s,t)$ that can be consistently estimated by $\hat{\pi}(s)^{-1}\hat{\pi}(t)^{-1}\hat{\nu}(s,t)\hat{\rho}(s,t)$, and that the norms of the other two terms converge in probability to 0. The proof of the weak convergence of $N^{1/2}(\hat{\mu} - \mu)$ then follows from the convergence of $\hat{\pi}$ to π, the consistency of the estimator of its covariance kernel can be shown analogously.

The weak convergence of S_n is shown with the help of Theorem 6, a central limit theorem for independent non-identically distributed Hilbert space variables given in the Appendix. We apply the theorem with $Y_{ni} = O_i(X_i - \mu)$. The covariance operator $\mathscr{G}_n$ of S_n is given by the kernel $\bar{\nu}(s,t)\rho(s,t)$. Denote by $\mathscr{G}$ the covariance operator with kernel $\nu(s,t)\rho(s,t)$. Conditions of the central limit theorem Theorem 6 can be shown using Definition 1(b) as follows. Condition (i) of Theorem 6 is satisfied because

$$\|\mathscr{G}_n - \mathscr{G}\|_\infty^2 \le \|\mathscr{G}_n - \mathscr{G}\|_2^2 = \int_{[0,1]^2} \{\bar{\nu}(s,t) - \nu(s,t)\}^2 \rho(s,t)^2 ds dt \to 0$$

as $n \to \infty$ by the dominated convergence theorem. Condition (ii) of Theorem 6 holds because

$$\begin{aligned} n^{-1} \sum_{i=1}^n \mathrm{E}(\|Y_{ni}\|^2 1_{[\|Y_{ni}\| > n^{1/2}\|\mathscr{G}_n\|_\infty \varepsilon]}) &\le n^{-1} \sum_{i=1}^n \mathrm{E}(\|X_i - \mu\|^2 1_{[\|X_i - \mu\| > n^{1/2}\|\mathscr{G}_n\|_\infty \varepsilon]}) \\ &= \mathrm{E}(\|X_1 - \mu\|^2 1_{[\|X_1 - \mu\| > n^{1/2}\|\mathscr{G}_n\|_\infty \varepsilon]}), \end{aligned}$$

which converges to 0 by the dominated convergence theorem. Finally, $\int_0^1 \bar{\nu}(t,t)\rho(t,t)dt \to \int_0^1 \nu(t,t)\rho(t,t)dt$ by the dominated convergence theorem again, and thus condition (iii) of Theorem 6 is satisfied. Hence the process S_n is asymptotically Gaussian with covariance kernel $\nu(s,t)\rho(s,t)$.

The expectation of the squared norm of the second term on the right side of (B.1) can be rewritten as

$$\int_0^1 \mathrm{E}\left[\left\{\frac{J(t)}{\hat{\pi}(t)} - \frac{1}{\pi(t)}\right\}^2 S_n(t)^2 1_{[\hat{\pi}(t) \ge \pi_0/2]}\right] dt + \int_0^1 \mathrm{E}\left[\left\{\frac{J(t)}{\hat{\pi}(t)} - \frac{1}{\pi(t)}\right\}^2 S_n(t)^2 1_{[\hat{\pi}(t) < \pi_0/2]}\right] dt. \tag{B.2}$$

The first summand above is dominated by

$$\int_0^1 \mathrm{E}\left[\frac{\{\pi(t) - \hat{\pi}(t)\}^2}{\pi_0^4/4} S_n(t)^2\right] dt \le \int_0^1 \mathrm{E}\left[\frac{\{\pi(t) - \hat{\pi}(t)\}^2}{\pi_0^4/4}\right] \rho(t,t) dt$$

which converges to zero by the dominated convergence theorem since $\mathrm{E}[\{\pi(t)-\hat{\pi}(t)\}^2] = \{\pi(t)-\bar{\pi}(t)\}^2 + n^{-2}\sum_{i=1}^n \pi_i(t)\{1-\pi_i(t)\} \to 0$ for $n \to \infty$. Next, we first compute

$$\left\{\frac{J(t)}{\hat{\pi}(t)} - \frac{1}{\pi(t)}\right\}^2 1_{[\hat{\pi}(t)<\pi_0/2]} = \left[J(t)\left\{\frac{\pi(t)-\hat{\pi}(t)}{\hat{\pi}(t)\pi(t)}\right\}^2 + \{1-J(t)\}\frac{1}{\pi(t)^2}\right]1_{[\hat{\pi}(t)<\pi_0/2]}$$
$$\leq [J(t)n^2/\pi_0^2 + \{1-J(t)\}/\pi_0^2]1_{[\hat{\pi}(t)<\pi_0/2]} \leq n^2/\pi_0^2 1_{[\hat{\pi}(t)<\pi_0/2]}.$$

Then the second summand in (B.2) is smaller than or equal to

$$\int_0^1 \mathrm{E}\{n^2/\pi_0^2 1_{[\hat{\pi}(t)<\pi_0/2]} S_n(t)^2\}dt \leq \int_0^1 n^2/\pi_0^2 \Pr\{\hat{\pi}(t) < \pi_0/2\}\rho(t,t)dt \leq n^2 \sup_{t\in[0,1]} \Pr\{\hat{\pi}(t) < \pi_0/2\}/\pi_0^2 \operatorname{tr}\mathscr{R},$$

which converges to 0 because, in light of Hoeffding's inequality and Definition 1(a), for all $t \in [0,1]$,

$$\Pr\{\hat{\pi}(t) < \pi_0/2\} \leq \exp[-2n\{\bar{\pi}(t)-\pi_0/2\}^2] \leq \exp\left[-2n\left\{\pi_0/2 - \sup_{t\in[0,1]}|\bar{\pi}(t)-\pi(t)|\right\}^2\right] \to 0.$$

This completes the proof of the convergence in probability of the norm of the second term on the right hand side of (B.1) to zero. The last term in (B.1) can be shown to converge to zero using similar arguments based on Hoeffding's inequality.

We now turn to the proof of the consistency of the estimator of the covariance kernel. To show that

$$\mathrm{E}\int_{[0,1]^2}\left\{\frac{\hat{\nu}(s,t)\hat{\rho}(s,t)}{\hat{\pi}(s)\hat{\pi}(t)} - \frac{\nu(s,t)\rho(s,t)}{\pi(s)\pi(t)}\right\}^2 dsdt \to 0,$$

we can split the integral into the integrals over $A_0 = \{(s,t) \in [0,1]^2 : \nu(s,t) = 0\}$ and $A_1 = \{(s,t) \in [0,1]^2 : \nu(s,t) \geq \nu_0\}$ because Definition 1(c) implies that $A_0 \cup A_1 = [0,1]^2$. On A_0 we obtain

$$\mathrm{E}\int_{A_0}\left\{\frac{\hat{\nu}(s,t)\hat{\rho}(s,t)}{\hat{\pi}(s)\hat{\pi}(t)}\right\}^2 \{1_{[\min(\hat{\pi}(s),\hat{\pi}(t))\geq\pi_0/2]} + 1_{[\min\{\hat{\pi}(s),\hat{\pi}(t)\}<\pi_0/2]}\}dsdt$$
$$\leq \int_{A_0} \mathrm{E}\{\hat{\nu}(s,t)^2\}\,\mathrm{E}\{\hat{\rho}(s,t)^2\}dsdt\left((\pi_0/2)^{-4} + n^4 \sup_{(s,t)\in[0,1]^2} \Pr[\min\{\hat{\pi}(s),\hat{\pi}(t)\} < \pi_0/2]\right).$$

Here the integral converges to zero by the dominated convergence theorem as the integrand can be shown to go to 0 and the second term in the brackets asymptotically vanishes due to an exponential rate of decrease of the supremum that can be established with the help of Hoeffding's inequality as before, hence the whole quantity above converges to 0. We now focus on A_1. We rewrite

$$\frac{\hat{\nu}(s,t)\hat{\rho}(s,t)}{\hat{\pi}(s)\hat{\pi}(t)} - \frac{\nu(s,t)\rho(s,t)}{\pi(s)\pi(t)} = \frac{\hat{\nu}(s,t)}{\hat{\pi}(s)\hat{\pi}(t)}\{\hat{\rho}(s,t)-\rho(s,t)\} + \left\{\frac{\hat{\nu}(s,t)}{\hat{\pi}(s)\hat{\pi}(t)} - \frac{\nu(s,t)}{\pi(s)\pi(t)}\right\}\rho(s,t) \tag{B.3}$$

and show that the integral over A_1 of the expectation of the square of each summand converges to zero. For the first summand we compute

$$\int_{A_1}\mathrm{E}\left(\left[\frac{\hat{\nu}(s,t)}{\hat{\pi}(s)\hat{\pi}(t)}\{\hat{\rho}(s,t)-\rho(s,t)\}\right]^2\{1_{[\min(\hat{\pi}(s),\hat{\pi}(t))\geq\pi_0/2]} + 1_{[\min\{\hat{\pi}(s),\hat{\pi}(t)\}<\pi_0/2]}\}\right)dsdt$$
$$\leq \mathrm{E}\int_{A_1}\{\hat{\rho}(s,t)-\rho(s,t)\}^2 dsdt\left[(\pi_0/2)^{-4} + n^4 \sup_{(s,t)\in[0,1]^2}\Pr(\min\{\hat{\pi}(s),\hat{\pi}(t)\} < \pi_0/2)\right],$$

where the integral term converges to 0 by similar arguments to those in the proof of Proposition 1 in Kraus [35] with the help of Definition 1(c) and the second term goes to 0 by Hoeffding's inequality again. For the second summand on the right in (B.3) we can write

$$\int_{A_1}\mathrm{E}\left[I(s,t)\left\{\frac{\pi(s)\pi(t)\hat{\nu}(s,t) - \hat{\pi}(s)\hat{\pi}(t)\nu(s,t)}{\hat{\pi}(s)\hat{\pi}(t)\pi(s)\pi(t)}\right\}^2\right]\rho(s,t)^2 dsdt$$
$$+ \int_{A_1}\mathrm{E}\left[\{1-I(s,t)\}\left\{\frac{\nu(s,t)}{\pi(s)\pi(t)}\right\}^2\right]\rho(s,t)^2 dsdt. \tag{B.4}$$

Like before, we split the first term in (B.4) into two summands by writing

$$\int_{A_1}\mathrm{E}\left[I(s,t)\left\{\frac{\pi(s)\pi(t)\hat{\nu}(s,t) - \hat{\pi}(s)\hat{\pi}(t)\nu(s,t)}{\hat{\pi}(s)\hat{\pi}(t)\pi(s)\pi(t)}\right\}^2\{1_{[\min(\hat{\pi}(s),\hat{\pi}(t))\geq\pi_0/2]} + 1_{[\min\{\hat{\pi}(s),\hat{\pi}(t)\}<\pi_0/2]}\}\right]\rho(s,t)^2 dsdt.$$

The first summand is bounded by $16\pi_0^{-8}\int_{A_1}\mathrm{E}[\{\pi(s)\pi(t)\hat{\nu}(s,t) - \hat{\pi}(s)\hat{\pi}(t)\nu(s,t)\}^2]\rho(s,t)^2 dsdt$, which converges to 0 by the dominated convergence theorem since the expectation in the integrand can be shown to converge to 0; the

second summand in the displayed expression above is dominated by $n^4\pi_0^{-4}\|\mathscr{R}\|_2^2 \sup_{(s,t)\in[0,1]^2} \Pr(\min\{\hat\pi(s),\hat\pi(t)\} < \pi_0/2)$, which converges to 0 by Hoeffding's inequality. Finally, the second term in (B.4) is dominated by $\sup_{(s,t)\in A_1} \Pr(\hat\nu(s,t) < \nu_0/2)\pi_0^{-4}\|\mathscr{R}\|_2^2$, which converges to 0 again by Hoeffding's inequality.

Proof of Theorem 2. Denote $Z_j(\cdot) = N_j(\cdot)^{1/2}\{\hat\mu_j(\cdot) - \hat\mu(\cdot)\}/\hat r_j$ and $Z = (Z_1,\dots,Z_K)^\top$. Under the null hypothesis we can write $Z = \hat{\mathscr{D}}H$, where $H = (H_1,\dots,H_K)^\top$ with $H_j = N_j^{1/2}(\hat\mu_j - \mu)$ and $\hat{\mathscr{D}}$ is a bounded linear operator from $\{L^2([0,1])\}^K$ to $\{L^2([0,1])\}^K$ that maps an element f to an element g whose jth component is given by $g_j(t) = \sum_{l=1}^K(\hat{\mathscr{D}}_{jl}f_l)(t) = \sum_{l=1}^K \hat r_j^{-1}\{\delta_{jl} - N_j(t)^{1/2}\hat w_l(t)J_l(t)N_l(t)^{-1/2}\}f_l(t)$ (here δ_{jl} is the Kronecker delta and $J_l(t)N_l(t)^{-1/2}$ is zero if $J_l(t) = 1_{[N_l(t)>0]}$ is zero). From Theorem 1 we see that H converges in distribution to the random element $H^\infty = (H_1^\infty,\dots,H_K^\infty)^\top$ whose components are mutually independent Gaussian processes with mean zero and covariance operators $\mathscr{K}_j, j = 1\dots,K$ analogous to the operator $\mathscr{K}$ in Theorem 1. The operator $\hat{\mathscr{D}}$ converges in probability to the operator $\mathscr{D}$ whose elements are defined by $(\mathscr{D}_{jl}f_l)(t) = r_j^{-1}\{\delta_{jl} - \pi_j(t)^{1/2}a_j^{1/2}w_l(t)\pi_l(t)^{-1/2}a_l^{-1/2}\}f_l(t)$ with $w_l(t) = a_l\pi_l(t)/r_l^2/(\sum_{k=1}^K a_k\pi_k(t)/r_k^2)$ (the convergence is in the operator norm, i.e., $\|\hat{\mathscr{D}} - \mathscr{D}\|_\infty \xrightarrow{P} 0$). Therefore, it follows from Slutsky's and continuous mapping theorem that $Z = \hat{\mathscr{D}}H$ converges weakly to $Z^\infty = \mathscr{D}H^\infty$. This is a K-dimensional mean zero Gaussian random process with cross-covariance operator between Z_j^∞ and Z_k^∞ equal to $\mathscr{V}_{jk} = \sum_{l=1}^K \mathscr{D}_{jl}\mathscr{K}_l\mathscr{D}_{kl}^*$, $j = 1,\dots,K$, $k = 1,\dots,K$. These can be consistently estimated by plugging-in the estimators $\hat{\mathscr{D}}_{jl}$ and $\hat{\mathscr{K}}_l$. The kernel of the estimator $\hat{\mathscr{V}}_{jk}$ takes the form $\hat v_{jk}(s,t) = \sum_{l=1}^K \hat r_j^{-1}\{\delta_{jl} - N_j(s)^{1/2}\hat w_l(s)N_l(s)^{-1/2}\}\hat\kappa_l(s,t)\{\delta_{kl} - N_k(t)^{1/2}\hat w_l(t)N_l(t)^{-1/2}\}\hat r_k^{-1}$.

For (i), the continuous mapping theorem gives that the statistic $T_{L^2} = \|Z\|^2$ converges weakly to the random variable $\|Z^\infty\|^2$. The process Z^∞ is a Gaussian random element of the separable Hilbert space $\{L^2([0,1])\}^K$. Therefore, it can be expanded in a Karhunen–Loève series with Gaussian coefficients. Consequently, the distribution of its squared norm is that of the series given in the theorem. The consistency of $\hat{\mathscr{V}}$ implies the consistency of the estimated eigenvalues.

To prove (ii), notice that the components of the score vector satisfy $Q_{jl} = \langle\hat\pi_j^{1/2}Z_j, \hat\psi_l\rangle$. The continuous mapping theorem and Slutsky's theorem in conjunction with the convergence of $\hat\psi_l$ imply that Q is asymptotically distributed as a Gaussian vector with mean zero and covariance matrix with entries $V_{jl,km} = \langle\pi_j^{1/2}\psi_l, \mathscr{V}_{jk}(\pi_k^{1/2}\psi_m)\rangle$. The consistency of $\hat V_{jl,km}$ follows from the consistency of $\hat{\mathscr{V}}_{jk}$ and $\hat\pi_j$ and convergence of $\hat\psi_l$. The process $(\hat\pi_1^{1/2}Z_1,\dots,\hat\pi_K^{1/2}Z_K)$ lies in a $(K-1)$-dimensional subspace of the K-dimensional product space $\{L^2([0,1])\}^K$ and the same holds for its limit. Therefore, the score vector lies in a $(K-1)d$-dimensional subspace of $\mathbb{R}^{Kd}$, leading to $(K-1)d$ degrees of freedom of the chi-square distribution.

Proof of Theorem 3. The kernel of $n^{1/2}(\hat{\mathscr{R}} - \mathscr{R})$ is

$$n^{1/2}\{\hat\rho(s,t) - \rho(s,t)\} = n^{1/2}\{\hat\rho(s,t) - \check\rho(s,t)\} + \frac{1}{\nu(s,t)}\sigma(s,t) + \left\{\frac{I(s,t)}{\hat\nu(s,t)} - \frac{1}{\nu(s,t)}\right\}\sigma(s,t) + n^{1/2}\{I(s,t) - 1\}\rho(s,t), \tag{B.5}$$

where $\check\rho$ is defined like $\hat\rho$ with the true mean in place of the estimated mean and $\sigma(s,t) = n^{-1/2}\sum_{i=1}^n U_i(s,t)[\{X_i(s) - \mu(s)\}\{X_i(t) - \mu(t)\} - \rho(s,t)]$. Let us focus on the second summand on the right side of (B.5). All the other terms are negligible in the appropriate sense as we explain later. The kernel $\sigma(s,t)$ corresponds to the operator $\mathscr{S}_n = n^{-1/2}\sum_{i=1}^n \mathscr{Y}_{ni}$, where $\mathscr{Y}_{ni}$ are the integral operators with kernels $y_{ni}(s,t) = U_i(s,t)[\{X_i(s) - \mu(s)\}\{X_i(t) - \mu(t)\} - \rho(s,t)]$. We will apply Theorem 6 to $\mathscr{Y}_{ni}$, which is a triangular array of row-wise independent non-identically distributed zero-mean random elements of the separable Hilbert space of the Hilbert–Schmidt operators on $L^2([0,1])$. The covariance operator of $\mathscr{Y}_{ni}$ is the Hilbert–Schmidt operator $\mathfrak{C}_{ni}$ on Hilbert–Schmidt operators given by

$$\langle\mathscr{A}_1, \mathfrak{C}_{ni}\mathscr{A}_2\rangle = \operatorname{cov}(\langle\mathscr{Y}_{ni},\mathscr{A}_2\rangle, \langle\mathscr{Y}_{ni},\mathscr{A}_1\rangle) = \int_{[0,1]^4}\alpha_1(s,t)\operatorname{cov}\{y_{ni}(s,t), y_{ni}(u,v)\}\alpha_2(u,v)dsdtdudv,$$

where $\mathscr{A}_1$, $\mathscr{A}_2$ are Hilbert–Schmidt operators with kernels α_1, α_2, respectively. The kernel of $\mathfrak{C}_{ni}$ is $c_{ni}(s,t,u,v) = \operatorname{cov}\{y_{ni}(s,t), y_{ni}(u,v)\} = \theta_i(s,t,u,v)\{\zeta(s,t,u,v) - \rho(s,t)\rho(u,v)\}$. The covariance operator of $\mathscr{S}_n$ is $\mathfrak{G}_n = n^{-1}\sum_{i=1}^n \mathfrak{C}_{ni}$ with kernel $\bar\theta(s,t,u,v)\{\zeta(s,t,u,v) - \rho(s,t)\rho(u,v)\}$. Like in the proof of Theorem 1, one can use the dominated convergence theorem to show that $\|\mathfrak{G}_n - \mathfrak{G}\|_2 \to 0$, where $\mathfrak{G}$ has kernel $\theta(s,t,u,v)\{\zeta(s,t,u,v) - \rho(s,t)\rho(u,v)\}$. Thus condition (i) of Theorem 6 is verified. Condition (ii) can be verified like in the proof of Theorem 1. Next, condition (iii) is satisfied because $\operatorname{tr}\mathfrak{G}_n = \int_{[0,1]^2}\bar\theta(s,t,s,t)\{\zeta(s,t,s,t) - \rho(s,t)^2\}dsdt$ converges to $\operatorname{tr}\mathfrak{G} = \int_{[0,1]^2}\theta(s,t,s,t)\{\zeta(s,t,s,t) - \rho(s,t)^2\}dsdt$. Therefore, $\mathscr{S}_n$ is asymptotically distributed as a Gaussian random operator with mean zero and covariance operator $\mathfrak{G}$ and, consequently, by the continuous mapping theorem the second term on the right-hand side of (B.5) weakly converges to the mean zero Gaussian operator with covariance operator $\mathfrak{H}'$ given in Theorem 3.

The operators corresponding to the first and fourth summand on the right side in (B.5) were shown to converge to zero in the proof of Proposition 1 in Kraus [35] in the sense that the expectation of their squared Hilbert–Schmidt norm converges to zero. Also, the Hilbert–Schmidt norm of the third term on the right in (B.5) converges to zero in mean square which can be shown by arguments analogous to those used for the second term on the right in (B.1) in the proof of Theorem 1. Therefore, in view of Slutsky's lemma these terms are negligible for the weak convergence.

The weak convergence of the operator with kernel $M(s,t)^{1/2}\{\hat\rho(s,t)-\rho(s,t)\}$ follows from the convergence of $\hat\nu(s,t)$ to $\nu(s,t)$. The consistency of the estimators of $\mathfrak{H}'$ and $\mathfrak{H}$ can be proved along the lines of the proof for $\mathscr{K}'$ and $\mathscr{K}$ in Theorem 1.

Proof of Theorem 4. The proof uses perturbation theory in which $\hat{\mathscr{R}}$ is regarded as a perturbed version of $\mathscr{R}$, i.e., $\hat{\mathscr{R}} = \mathscr{R} + (\hat{\mathscr{R}} - \mathscr{R})$. Recall that the perturbation satisfies $\mathrm{E}\,\|\hat{\mathscr{R}} - \mathscr{R}\|_2^2 = O(n^{-1})$ [35, Proposition 1], and, therefore, $\|\hat{\mathscr{R}} - \mathscr{R}\|_\infty = O_P(n^{-1/2})$.

Similarly to the proof of Theorem 3.1 in [10], we rewrite $n^{1/2}(\hat\lambda_m - \lambda_m) = n^{1/2}(\hat\lambda_m - \lambda_m)1_{\Omega_n} + n^{1/2}(\hat\lambda_m - \lambda_m)1_{\Omega_n^c}$, where $\Omega_n = \{\omega : \|\hat{\mathscr{R}} - \mathscr{R}\|_\infty < \varepsilon_n\}$ for a numerical sequence ε_n satisfying $n^{-1/2} \ll \varepsilon_n \ll n^{-1/4}$. Since $\Pr(\Omega_n) \to 1$ as $n \to \infty$, the term $n^{1/2}(\hat\lambda_m - \lambda_m)1_{\Omega_n^c}$ converges to 0 in probability. For $\|\hat{\mathscr{R}} - \mathscr{R}\|_\infty$ sufficiently small, i.e., on Ω_n for n large enough, we have by Corollary 3.4 of [22] that $n^{1/2}(\hat\lambda_m - \lambda_m)1_{\Omega_n} = n^{1/2}\langle(\hat{\mathscr{R}} - \mathscr{R})\varphi_m, \varphi_m\rangle 1_{\Omega_n} + n^{1/2}O(\|\hat{\mathscr{R}} - \mathscr{R}\|_\infty^2)1_{\Omega_n}$. Here the last term converges to 0 in probability because $\varepsilon_n \ll n^{-1/4}$ and the first term on the right side converges in distribution to the limit given in part (i) of the theorem. Hence the result follows from Slutsky's theorem. The expression for the limiting variance is obtained by rewriting $\mathrm{var}\langle\mathscr{H}'^\infty\varphi_m, \varphi_m\rangle = \mathrm{var}\langle\mathscr{H}'^\infty, \varphi_m \otimes \varphi_m\rangle = \langle\varphi_m \otimes \varphi_m, \mathfrak{H}'(\varphi_m \otimes \varphi_m)\rangle$.

Next, we can write $n^{1/2}(\hat s_m\hat\varphi_m - \varphi_m) = n^{1/2}(\hat s_m\hat\varphi_m - \varphi_m)1_{\Omega_n} + n^{1/2}(\hat s_m\hat\varphi_m - \varphi_m)1_{\Omega_n^c}$. For n sufficiently large, Corollary 3.3 of [22] gives $n^{1/2}(\hat s_m\hat\varphi_m - \varphi_m)1_{\Omega_n} = n^{1/2}\mathscr{Q}_m(\hat{\mathscr{R}} - \mathscr{R})\varphi_m 1_{\Omega_n} + n^{1/2}O(\|\hat{\mathscr{R}} - \mathscr{R}\|_\infty^2)1_{\Omega_n}$. The first term on the right converges in distribution to the limiting distribution as claimed in part (ii) and the other terms converge in probability to 0. The limiting covariance operator is obtained by inspecting the cross-covariance operator for each pair of summands in the series $\mathscr{Q}_m\mathscr{H}'^\infty\varphi_m$. The cross-covariance between $(\varphi_k\otimes\varphi_k)\mathscr{H}'^\infty\varphi_m = \langle\varphi_k, \mathscr{H}'^\infty\varphi_m\rangle\varphi_k$ and $(\varphi_l\otimes\varphi_l)\mathscr{H}'^\infty\varphi_m = \langle\varphi_l, \mathscr{H}'^\infty\varphi_m\rangle\varphi_l$ is

$$\begin{aligned}\mathrm{cov}(\langle\varphi_k, \mathscr{H}'^\infty\varphi_m\rangle, \langle\varphi_l, \mathscr{H}'^\infty\varphi_m\rangle)(\varphi_k\otimes\varphi_l) &= \mathrm{cov}\{\langle(\varphi_m\otimes\varphi_k), \mathscr{H}'^\infty\rangle, \langle(\varphi_m\otimes\varphi_l), \mathscr{H}'^\infty\rangle\}(\varphi_k\otimes\varphi_l)\\ &= \langle(\varphi_m\otimes\varphi_k), \mathfrak{H}'(\varphi_m\otimes\varphi_l)\rangle(\varphi_k\otimes\varphi_l).\end{aligned}$$

The inner product in the last expression above equals the integral in part (ii) of the theorem.

Proof of Theorem 5. Let $\hat{\mathfrak{D}}$ be the linear operator on the product space $\mathrm{HS}(L^2([0,1]))^K$ that maps $\mathscr{F} = (\mathscr{F}_1, \ldots, \mathscr{F}_K)^\top$, where $\mathscr{F}_j$ are Hilbert–Schmidt operators on $L^2([0,1])$ with kernels $f_j(s,t)$, to $\mathscr{G} = (\mathscr{G}_1, \ldots, \mathscr{G}_K)^\top$ where $\mathscr{G}_j$ has kernel $g_j(s,t) = \sum_{l=1}^K\{\delta_{jl} - M_j(s,t)^{1/2}\hat w_l(s,t)I_l(s,t)M_l(s,t)^{-1/2}\}f_l(s,t)$. The mapping $\hat{\mathfrak{D}}$ is a random linear operator on $\mathrm{HS}(L^2([0,1]))^K$ that acts by pointwise multiplication and linear combination of integral kernels; $\hat{\mathfrak{D}}$ itself is not an integral operator but it is bounded because the functions in the braces above are bounded. It converges in probability to the non-random bounded linear operator $\mathfrak{D}$ that maps $\mathscr{F}$ to $\mathscr{G}$ with $\mathscr{G}_j$ with kernel $\sum_{l=1}^K\{\delta_{jl} - \nu_j(s,t)^{1/2}a_j^{1/2}w_l(s,t)\nu_l(s,t)^{-1/2}a_l^{-1/2}\}$ $f_l(s,t)$. The convergence is in the sense of the operator norm on linear operators on $\mathrm{HS}(L^2([0,1]))^K$, that is, $\|\hat{\mathfrak{D}} - \mathfrak{D}\|_\infty \xrightarrow{P} 0$, where $\|\mathfrak{D}\|_\infty = \sup\{\|\mathfrak{D}\mathscr{F}\|_2/\|\mathscr{F}\|_2 : \mathscr{F} \in \mathrm{HS}(L^2([0,1]))^K\}$ with $\|\cdot\|_2$ being the Hilbert–Schmidt norm on $\mathrm{HS}(L^2([0,1]))^K$.

Now consider the standardized contrasts $\mathscr{Z} = (\mathscr{Z}_1, \ldots, \mathscr{Z}_K)^\top$ with kernels $z_j(s,t) = M_j(s,t)^{1/2}\{\hat\rho_j(s,t) - \hat\rho(s,t)\}$. They are obtained as $\mathscr{Z} = \hat{\mathfrak{D}}\mathscr{H}$, where $\mathscr{H} = (\mathscr{H}_1, \ldots, \mathscr{H}_K)^\top$ with $\mathscr{H}_j$ with kernel $h_j(s,t) = M_j(s,t)^{1/2}\{\hat\rho(s,t) - \rho(s,t)\}$. Under the null hypothesis Theorem 3 yields that $\mathscr{H}$ converges in distribution to $\mathscr{H}^\infty$, a vector of K independent mean zero Gaussian random operators with covariance operators $\mathfrak{H}_j$. Therefore, $\mathscr{Z} = \hat{\mathfrak{D}}\mathscr{H}$ converges in distribution to $\mathscr{Z}^\infty = \mathfrak{D}\mathscr{H}^\infty$ by Slutsky's and continuous mapping theorem.

The covariance operator $\mathfrak{B}$ of $\mathscr{Z}^\infty$ is given by the cross-covariance operators $\mathfrak{B}_{jk}$ between the components $\mathscr{Z}_j$ and $\mathscr{Z}_k$ whose estimator $\hat{\mathfrak{B}}_{jk}$ has kernel

$$\hat\beta_{jk}(s,t,u,v) = \sum_{l=1}^K\{\delta_{jl} - M_j(s,t)^{1/2}\hat w_l(s,t)M_l(s,t)^{-1/2}\}\hat\eta_l(s,t,u,v)\{\delta_{kl} - M_k(u,v)^{1/2}\hat w_l(u,v)M_l(u,v)^{-1/2}\}.$$

The test statistic $S_{\mathrm{HS}} = \|\mathscr{Z}\|_2^2$ is asymptotically distributed as $\|\mathscr{Z}^\infty\|_2^2$. The random variable $\mathscr{Z}^\infty$ is a Gaussian element of the separable Hilbert space $\mathrm{HS}(L^2([0,1]))^K$, therefore it can be expanded in a Karhunen–Loève series with independent Gaussian coefficients. Therefore, its squared norm is distributed as the series of independent chi-square variables weighted by the eigenvalues of the covariance operator and part (i) of the theorem follows.

The components of the score vector satisfy $R_{jlm} = \langle\hat\nu_j(\cdot,\cdot)^{1/2}z_j(\cdot,\cdot), \hat{\mathscr{U}}_{lm}\rangle$. Due to the consistency of the estimated eigenfunctions [35, Proposition 2], the operator $\hat{\mathscr{U}}_{lm}$ (up to the sign ambiguity for $l \neq m$) converges to $\mathscr{U}_{lm}$ defined by the true eigenfunctions, with kernel $u_{lm}(s,t)$. Therefore, the score vector weakly converges to the mean zero Gaussian vector with components $R^\infty_{jlm} = \langle\nu_j(\cdot,\cdot)^{1/2}z_j^\infty(\cdot,\cdot), \mathscr{U}_{lm}\rangle = \langle z_j^\infty(\cdot,\cdot), \nu_j(\cdot,\cdot)^{1/2}u_{lm}(\cdot,\cdot)\rangle$ whose covariance matrix has entries $W_{jlm,kpq} = \langle\nu_j(\cdot,\cdot)^{1/2}u_{lm}(\cdot,\cdot), \mathfrak{B}_{jk}\{\nu_k(\cdot,\cdot)^{1/2}u_{pq}(\cdot,\cdot)\}\rangle$, $j,k \in \{1,\ldots,K\}$, $1 \le l \le m \le d$, $1 \le p \le q \le d$. The vector of operators with kernels $\nu_j(s,t)^{1/2}z_j^\infty(s,t)$ lies in a hyperplane in $\mathrm{HS}(L^2([0,1]))^K$, thus the matrix W has rank $(K-1)d(d+1)/2$. The consistency of $\hat W$ follows from the convergence of all quantities involved. Hence the limiting distribution is the chi-square distribution as claimed in part (ii).

Proof of Theorem 6. First, we prove the convergence in distribution of one-dimensional projections using Lindeberg's central limit theorem. It follows from assumption (i) that for $f \in \mathcal{H}$ such that $\mathscr{G}f \neq 0$, $\mathrm{var}\langle S_n, f\rangle = \langle f, \mathscr{G}_n f\rangle \to \langle f, \mathscr{G}f\rangle$ as

$n \to \infty$. To verify Lindeberg's condition, we compute

$$n^{-1}\sum_{i=1}^{n} \mathrm{E}(\langle Y_{ni}, f\rangle^2 1_{[|\langle Y_{ni},f\rangle|>n^{1/2}\langle f,\mathscr{G}_n f\rangle^{1/2}\varepsilon]}) \le n^{-1}\sum_{i=1}^{n} \mathrm{E}(\|Y_{ni}\|^2\|f\|^2 1_{[\|Y_{ni}\|>n^{1/2}\langle f,\mathscr{G}_n f\rangle^{1/2}\|f\|^{-1}\varepsilon]}).$$

Now in light of assumption (i), there is a positive constant c such that for sufficiently large n, $\langle f, \mathscr{G}_n f\rangle^{1/2}/\|\mathscr{G}_n\|_\infty > c$, and the above expression is further dominated by $n^{-1}\sum_{i=1}^{n} \mathrm{E}(\|Y_{ni}\|^2\|f\|^2 1_{[\|Y_{ni}\|>n^{1/2}\|\mathscr{G}_n\|_\infty c\|f\|^{-1}\varepsilon]})$, which converges to 0 by assumption (ii). Hence one-dimensional projections converge, and due to Theorem 2.3 of Bosq [5], all finite-dimensional projections converge.

To complete the proof, let us prove the tightness of the sequence S_n, $n = 1, 2, \ldots$ The idea of the proof is similar to that of Bosq [5, Theorem 2.7] but in the present situation the variables $Y_{n1}, \ldots, Y_{nn}$ are possibly non-identically distributed. Let v_j and δ_j, $j = 1, 2, \ldots$ be the eigenfunctions and eigenvalues of the limiting operator $\mathscr{G}$. Consider a sequence l_k, $k = 1, 2, \ldots$ such that $l_k \to \infty$ for $k \to \infty$. For $\varepsilon > 0$, let N_k, $k = 1, 2, \ldots$ be an increasing sequence of integers such that $\sum_{k=1}^{\infty} l_k r_{N_k}^2 < \varepsilon$, where $r_N^2 = \sum_{j=N}^{\infty}\delta_j$. Define $B_k = \{x \in \mathcal{H} : \sum_{j=N_k}^{\infty}\langle x, v_j\rangle^2 \le l_k^{-1}\}$. It follows from assumptions (i) and (iii) that

$$\begin{aligned}\Pr(S_n \in B_k^{\mathsf{C}}) &= P\left(\sum_{j=N_k}^{\infty}\langle S_n, v_j\rangle^2 > l_k^{-1}\right) \le l_k\,\mathrm{E}\left(\sum_{j=N_k}^{\infty}\langle S_n, v_j\rangle^2\right) = l_k\,\mathrm{E}\left(\|S_n\|^2 - \sum_{j=1}^{N_k-1}\langle S_n, v_j\rangle^2\right)\\ &= l_k\left(\operatorname{tr}\mathscr{G}_n - \sum_{j=1}^{N_k-1}\langle v_j, \mathscr{G}_n v_j\rangle\right) \to l_k\left(\operatorname{tr}\mathscr{G} - \sum_{j=1}^{N_k-1}\langle v_j, \mathscr{G} v_j\rangle\right) = l_k\sum_{j=N_k}^{\infty}\langle v_j, \mathscr{G} v_j\rangle = l_k r_{N_k}^2.\end{aligned}$$

Consider the compact set $K_\varepsilon = \cap_{k=1}^{\infty} B_k$ and compute

$$\limsup_{n\to\infty}\Pr(S_n \in K_\varepsilon^{\mathsf{C}}) \le \limsup_{n\to\infty}\sum_{k=1}^{\infty}\Pr(S_n \in B_k^{\mathsf{C}}) \le \sum_{k=1}^{\infty}\limsup_{n\to\infty}\Pr(S_n \in B_k^{\mathsf{C}}) \le \sum_{k=1}^{\infty} l_k r_{N_k}^2 < \varepsilon,$$

where the second inequality is due to Fatou's lemma. This proves the tightness.

Appendix C. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2019.05.002. The supplementary document available online contains further simulation results and additional graphs for the data application. R code is available online.

References

[1] J.A.D. Aston, D. Pigoli, S. Tavakoli, Tests for separability in nonparametric covariance operators of random surfaces, Ann. Statist. 45 (4) (2017) 1431–1461.
[2] A. Aue, R. Gabrys, L. Horváth, P. Kokoszka, Estimation of a change-point in the mean function of functional data, J. Multivariate Anal. 100 (10) (2009) 2254–2269.
[3] M. Benko, W. Härdle, A. Kneip, Common functional principal components, Ann. Statist. 37 (1) (2009) 1–34.
[4] G. Boente, D. Rodriguez, M. Sued, Testing equality between several populations covariance operators, Ann. Inst. Statist. Math. (2017) 1–32.
[5] D. Bosq, Linear Processes in Function Spaces, Springer, New York, 2000.
[6] F.A. Bugni, Specification test for missing functional data, Econom. Theory 28 (5) (2012) 959–1002.
[7] A. Cabassi, D. Pigoli, P. Secchi, P.A. Carter, Permutation tests for the equality of covariance operators of functional data with applications to evolutionary biology, Electron. J. Stat. 11 (2) (2017) 3815–3840.
[8] G. Cao, L. Yang, D. Todem, Simultaneous inference for the mean function based on dense functional data, J. Nonparametr. Stat. 24 (2) (2012) 359–377.
[9] A. Cuevas, M. Febrero, R. Fraiman, An anova test for functional data, Comput. Statist. Data Anal. 47 (1) (2004) 111–122.
[10] J. Cupidon, D. Gilliam, R. Eubank, F. Ruymgaart, The delta method for analytic functions of random operators with application to functional data, Bernoulli 13 (4) (2007) 1179–1194.
[11] J. Dauxois, A. Pousse, Y. Romain, Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference, J. Multivariate Anal. 12 (1) (1982) 136–154.
[12] A.C. Davison, D.V. Hinkley, Bootstrap methods and their application, Cambridge University Press, Cambridge, 1997, p. x+582.
[13] M. Dawson, H.-G. Müller, Dynamic modeling of conditional quantile trajectories, with application to longitudinal snippet data, J. Amer. Statist. Assoc. 113 (524) (2018) 1612–1624.
[14] A. Delaigle, P. Hall, Classification using censored functional data, J. Amer. Statist. Assoc. 108 (504) (2013) 1269–1283.
[15] A. Delaigle, P. Hall, Approximating fragmented functional data by segments of Markov chains, Biometrika 103 (4) (2016) 779–799.
[16] M.-H. Descary, V.M. Panaretos, Recovering covariance from functional fragments, Biometrika 106 (1) (2019) 145–160.
[17] F. Ferraty, Y. Romain (Eds.), The Oxford Handbook of Functional Data Analysis, Oxford University Press, Oxford, 2011, p. xviii+494.
[18] C.B. Fogarty, D.S. Small, Equivalence testing for functional data with an application to comparing pulmonary function devices, Ann. Appl. Stat. 8 (4) (2014) 2002–2026.
[19] S. Fremdt, L. Horváth, P. Kokoszka, J.G. Steinebach, Functional data analysis with increasing number of projections, J. Multivariate Anal. 124 (2014) 313–332.

[20] S. Fremdt, J.G. Steinebach, L. Horváth, P. Kokoszka, Testing the equality of covariance operators in functional samples, Scand. J. Stat. 40 (1) (2013) 138–152.
[21] J.E. Gellar, E. Colantuoni, D.M. Needham, C.M. Crainiceanu, Variable-domain functional regression for modeling ICU data, J. Amer. Statist. Assoc. 109 (508) (2014) 1425–1439.
[22] D.S. Gilliam, T. Hohage, X. Ji, F. Ruymgaart, The Fréchet derivative of an analytic function of a bounded operator with some applications, Int. J. Math. Math. Sci. 2009 (2009).
[23] Y. Goldberg, Y. Ritov, A. Mandelbaum, Predicting the continuation of a function with applications to call center data, J. Statist. Plann. Inference 147 (2014) 53–65.
[24] O. Gromenko, P. Kokoszka, J. Sojka, Evaluation of the cooling trend in the ionosphere using functional regression with incomplete curves, Ann. Appl. Stat. 11 (2) (2017) 898–918.
[25] J. Guo, B. Zhou, J.-T. Zhang, New tests for equality of several covariance functions for functional data, J. Amer. Statist. Assoc. (2018) To appear.
[26] J. Guo, B. Zhou, J.-T. Zhang, Testing the equality of several covariance functions for functional data: a supremum-norm based test, Comput. Statist. Data Anal. 124 (2018) 15–26.
[27] L. Horváth, M. Hušková, P. Kokoszka, Testing the stability of the functional autoregressive process, J. Multivariate Anal. 101 (2) (2010) 352–367.
[28] L. Horváth, P. Kokoszka, Inference for functional data with applications, Springer, New York, 2012, p. xiv+422.
[29] L. Horváth, P. Kokoszka, R. Reeder, Estimation of the mean of functional time series and a two-sample problem, J. R. Stat. Soc. Ser. B Stat. Methodol. 75 (1) (2013) 103–122.
[30] D. Jarušková, Testing for a change in covariance operator, J. Statist. Plann. Inference 143 (9) (2013) 1500–1511.
[31] A. Kashlak, J. Aston, R. Nickl, Inference on covariance operators via concentration inequalities: k-sample tests, classification, and clustering via rademacher complexities, Sankhya A (2018).
[32] A. Kneip, D. Liebl, On the Optimal Reconstruction of Partially Observed Functional Data, Ann. of Statist. to appear, 2019.
[33] P. Kokoszka, M. Reimherr, Asymptotic normality of the principal components of functional time series, Stochastic Process. Appl. 123 (5) (2013) 1546–1562.
[34] P. Kokoszka, M. Reimherr, Introduction to Functional Data Analysis, CRC Press, 2017.
[35] D. Kraus, Components and completion of partially observed functional data, J. R. Stat. Soc. Ser. B Stat. Methodol. 77 (4) (2015) 777–801.
[36] D. Kraus, V.M. Panaretos, Dispersion operators and resistant second-order functional data analysis, Biometrika 99 (4) (2012) 813–832.
[37] D. Kraus, M. Stefanucci, Classification of functional fragments by regularized linear classifiers with domain selection, Biometrika 106 (1) (2019) 161–180.
[38] D. Liebl, Modeling and forecasting electricity spot prices: a functional data perspective, Ann. Appl. Stat. 7 (3) (2013) 1562–1592.
[39] D. Liebl, Nonparametric testing for differences in electricity prices: The case of the fukushima nuclear accident, Ann. Appl. Stat. (2019) To appear.
[40] D. Liebl, S. Rameseder, Partially observed functional data: The case of systematically missing parts, Comput. Statist. Data Anal. 131 (2019) 104–115.
[41] R.Y. Liu, Bootstrap procedures under some non-i.i.d. models, Ann. Statist. 16 (4) (1988) 1696–1708.
[42] C.J. Lloyd, Estimating test power adjusted for size, J. Stat. Comput. Simul. 75 (11) (2005) 921–933.
[43] A. Mas, Testing for the mean of random curves: a penalization approach, Stat. Inference Stoch. Process. 10 (2) (2007) 147–163.
[44] V. Masarotto, Procrustes Metric and Optimal Transport for Covariance Operators, Ph. D. thesis, Ecole Polytechnique Fédérale de Lausanne, 2019.
[45] M. Mojirsheibani, C. Shaw, Classification with incomplete functional covariates, Statist. Probab. Lett. 139 (2018) 40–46.
[46] V.M. Panaretos, D. Kraus, J.H. Maddocks, Second-order comparison of Gaussian random functions and the geometry of DNA minicircles, J. Amer. Statist. Assoc. 105 (490) (2010) 670–682.
[47] V.M. Panaretos, D. Kraus, J.H. Maddocks, Second-order inference for functional data with application to dna minicircles, in: Recent Advances in Functional Data Analysis and Related Topics, Springer, 2011, pp. 245–250.
[48] E. Paparoditis, T. Sapatinas, Bootstrap-Based K-Sample Testing For Functional Data, arXiv:1409.4317v4, 2016.
[49] E. Paparoditis, T. Sapatinas, Bootstrap-based testing of equality of mean functions or equality of covariance operators for functional data, Biometrika 103 (3) (2016) 727–733.
[50] D. Pigoli, J.A. Aston, I.L. Dryden, P. Secchi, Distances and inference for covariance operators, Biometrika 101 (2) (2014) 409–422.
[51] A. Pini, L. Spreafico, S. Vantini, A. Vietti, Multi-aspect local inference for functional data: Analysis of ultrasound tongue profiles, J. Multivariate Anal. 170 (2019) 162–185.
[52] A. Pini, A. Stamm, S. Vantini, Hotelling's T^2 in separable Hilbert spaces, J. Multivariate Anal. 167 (2018) 284–305.
[53] A. Pini, S. Vantini, The interval testing procedure: a general framework for inference in functional data analysis, Biometrics 72 (3) (2016) 835–845.
[54] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, Springer, New York, 2005.
[55] M. Stefanucci, L.M. Sangalli, P. Brutti, PCA-Based discrimination of partially observed functional data, with an application to AneuRisk65 data set, Stat. Neerl. 72 (3) (2018) 246–264.
[56] O. Vsevolozhskaya, M. Greenwood, D. Holodov, Pairwise comparison of treatment levels in functional analysis of variance with application to erythrocyte hemolysis, Ann. Appl. Stat. 8 (2) (2014) 905–925.
[57] J.-T. Zhang, Analysis of Variance for Functional Data, Chapman and Hall/CRC, 2013.
[58] J.-T. Zhang, X. Liang, One-way ANOVA for functional data via globalizing the pointwise F-test, Scand. J. Stat. 41 (1) (2014) 51–71.
[59] C. Zhang, H. Peng, J.-T. Zhang, Two samples tests for functional data, Commun. Statist. – Theory Methods 39 (4) (2010) 559–578.

Supplementary material for "Inferential procedures for partially observed functional data"

David Kraus*

Abstract: This supplementary document contains additional simulation results and further results of the data analysis.

Key words and phrases: Bootstrap; covariance operator; functional data; K-sample test; partial observation; principal components.

S1 Extended simulation results

Table S1 is an extended version of Table 1 presented in the main body of the paper. It includes additional simulation results for tests of equal means for non-Gaussian distributed curves and for groups with unequal covariance operators. The same model as in the paper is used except that for the non-Gaussian case independent t_5 distributed coefficients are generated and for the case of unequal covariance operators we set $\lambda_{3,0} = 0.2$. Since the empirical size deviates from the nominal level in some cases, Table S2 additionally reports size-adjusted powers for the same settings using the method described by Lloyd (2005, Subsection 3.2).

Table S3 reports results for tests of equal covariance operators. In addition to the results presented in Table 2 in the main body of the paper it contains results for t_5 distributed coefficients in the model for random curves. Table S4 reports size-adjusted powers for the same settings.

S2 Additional results for the data analysis

Fig. S1 contains additional plots of the covariance function estimates of the heart rate data shown in the main body of the paper. Fig. S2 shows the null estimates of the covariance functions and their leading eigenfunctions that the projection covariance test uses. Components of the score vector standardized by their estimated standard deviation are plotted in Fig. S3

Acknowledgements

We are grateful to all reviewers for their valuable comments and suggestions. This work was supported by the Czech Science Foundation under Grant GJ17-22950Y. Access to computing and storage facilities owned by parties and projects contributing to the MetaCentrum

*Department of Mathematics and Statistics, Masaryk University, Kotlářská 2, 611 37 Brno, Czech Republic; *david.kraus@mail.muni.cz.*

Table S1
Empirical rejection probability (in %) of the L^2 test, T_{L^2}, and projection test, T_d, of equal means. A dash indicates the same value as on the preceding row. The observation patterns (1)–(9) and mean configurations A–D are described in Section 5 of the paper.

Distrib.	Covar. oper.	Observ. pattern	Mean configuration							
			A		B		C		D	
			T_{L^2}	T_d	T_{L^2}	T_d	T_{L^2}	T_d	T_{L^2}	T_d
Gaussian	Equal	(1)	5.6	6.2	69	60	49	56	52	63
		(2)	5.4	6.7	59	52	28	29	38	50
		(3)	—	—	—	—	50	56	44	62
		(4)	4.4	6.5	66	58	51	57	51	62
		(5)	—	—	—	—	44	49	50	58
		(6)	5.4	7.1	58	51	50	55	42	49
		(7)	—	—	—	—	28	34	37	42
		(8)	5.4	5.8	55	47	34	37	42	48
		(9)	5.4	7.8	37	40	20	23	26	34
Gaussian	Unequal	(1)	4.2	5.2	79	75	58	63	57	67
		(2)	4.0	5.6	66	62	28	32	37	52
		(3)	—	—	—	—	56	62	47	66
		(4)	4.0	5.7	77	72	58	62	55	64
		(5)	—	—	—	—	50	55	53	63
		(6)	3.9	4.9	64	60	55	57	43	52
		(7)	—	—	—	—	29	36	38	46
		(8)	4.5	7.0	64	62	39	42	47	54
		(9)	4.0	6.5	42	48	23	25	27	38
t_5	Equal	(1)	5.4	7.3	72	61	51	58	54	63
		(2)	4.7	7.6	58	53	27	30	38	52
		(3)	—	—	—	—	50	60	44	63
		(4)	5.1	6.4	70	60	52	57	51	60
		(5)	—	—	—	—	46	52	50	60
		(6)	3.7	6.1	56	50	50	54	41	50
		(7)	—	—	—	—	27	32	37	43
		(8)	5.1	7.1	58	52	33	36	44	51
		(9)	5.4	6.6	38	42	21	24	26	34
t_5	Unequal	(1)	5.8	7.4	82	77	59	65	60	68
		(2)	4.7	6.9	68	64	32	35	44	57
		(3)	—	—	—	—	60	66	50	68
		(4)	5.2	6.7	80	76	62	65	59	66
		(5)	—	—	—	—	53	60	56	65
		(6)	3.9	6.1	65	63	57	61	47	57
		(7)	—	—	—	—	32	37	42	50
		(8)	4.8	7.5	65	64	39	42	50	56
		(9)	5.5	6.2	44	50	24	28	30	40

Table S2
Size-adjusted empirical power (in %) for the same settings as in Table S1.

Distrib.	Covar. oper.	Observ. pattern	Mean configuration					
			B		C		D	
			T_{L^2}	T_d	T_{L^2}	T_d	T_{L^2}	T_d
Gaussian	Equal	(1)	66	56	47	52	49	59
		(2)	56	43	25	23	34	41
		(3)	—	—	47	48	40	54
		(4)	68	52	52	48	52	54
		(5)	—	—	45	43	51	51
		(6)	58	46	50	49	42	45
		(7)	—	—	28	29	37	37
		(8)	54	45	34	34	41	45
		(9)	36	33	20	17	26	27
Gaussian	Unequal	(1)	83	73	63	62	62	66
		(2)	72	59	35	29	44	49
		(3)	—	—	62	59	56	64
		(4)	81	72	63	62	61	63
		(5)	—	—	56	55	59	62
		(6)	68	60	60	57	47	53
		(7)	—	—	34	36	43	46
		(8)	67	54	42	36	49	48
		(9)	45	45	25	23	31	35
t_5	Equal	(1)	71	55	50	51	52	57
		(2)	60	44	28	23	39	42
		(3)	—	—	51	47	46	53
		(4)	69	53	51	53	50	56
		(5)	—	—	44	45	49	55
		(6)	60	48	53	52	45	48
		(7)	—	—	31	30	40	40
		(8)	57	44	32	30	43	44
		(9)	38	38	21	20	26	31
t_5	Unequal	(1)	80	71	58	59	58	62
		(2)	68	56	32	27	44	48
		(3)	—	—	61	56	50	60
		(4)	80	71	62	60	59	62
		(5)	—	—	53	54	56	61
		(6)	70	61	61	57	51	54
		(7)	—	—	37	35	46	47
		(8)	66	56	40	35	51	49
		(9)	43	45	23	24	28	36

Table S3
Empirical rejection probability (in %) of the Hilbert–Schmidt norm test, S_{HS}, projection test, S_d, and square root covariance test, S_{sqrt}, of equal covariance operators. A dash indicates the same value as on the preceding row. The observation patterns (1)–(5) and covariance configurations A–D are described in Section 5 of the paper.

Distrib.	Observ. pattern	Covariance configuration											
		A			B			C			D		
		S_{HS}	S_d	S_{sqrt}	S_{HS}	S_d	S_{sqrt}	S_{HS}	S_d	S_{sqrt}	S_{HS}	S_d	S_{sqrt}
Gaussian	(1)	5.4	5.8	4.8	69	82	80	69	58	69	78	62	81
	(2)	4.6	6.4	4.9	54	63	41	37	32	38	76	64	54
	(3)	—	—	—	—	—	—	—	—	—	46	30	48
	(4)	5.0	5.1	5.8	64	74	72	61	53	62	72	56	73
	(5)	—	—	—	—	—	—	—	—	—	77	60	77
t_5	(1)	3.6	5.7	4.2	26	32	35	30	26	35	38	41	44
	(2)	3.3	6.5	3.4	22	31	18	14	17	16	38	41	23
	(3)	—	—	—	—	—	—	—	—	—	16	16	20
	(4)	4.0	6.4	4.8	23	32	30	25	25	31	30	33	34
	(5)	—	—	—	—	—	—	—	—	—	36	38	40

Table S4
Size-adjusted empirical power (in %)) for the same settings as in Table S3.

Distrib.	Observ. pattern	Covariance configuration								
		B			C			D		
		S_{HS}	S_d	S_{sqrt}	S_{HS}	S_d	S_{sqrt}	S_{HS}	S_d	S_{sqrt}
Gaussian	(1)	66	79	81	67	56	69	78	60	81
	(2)	54	59	42	38	29	38	78	60	55
	(3)	—	—	—	—	—	—	47	26	49
	(4)	64	73	69	61	52	59	72	56	71
	(5)	—	—	—	—	—	—	77	59	74
t_5	(1)	32	29	39	36	23	38	44	38	48
	(2)	23	24	20	18	14	18	40	33	26
	(3)	—	—	—	—	—	—	20	12	23
	(4)	29	26	31	31	19	32	37	27	35
	(5)	—	—	—	—	—	—	43	32	41

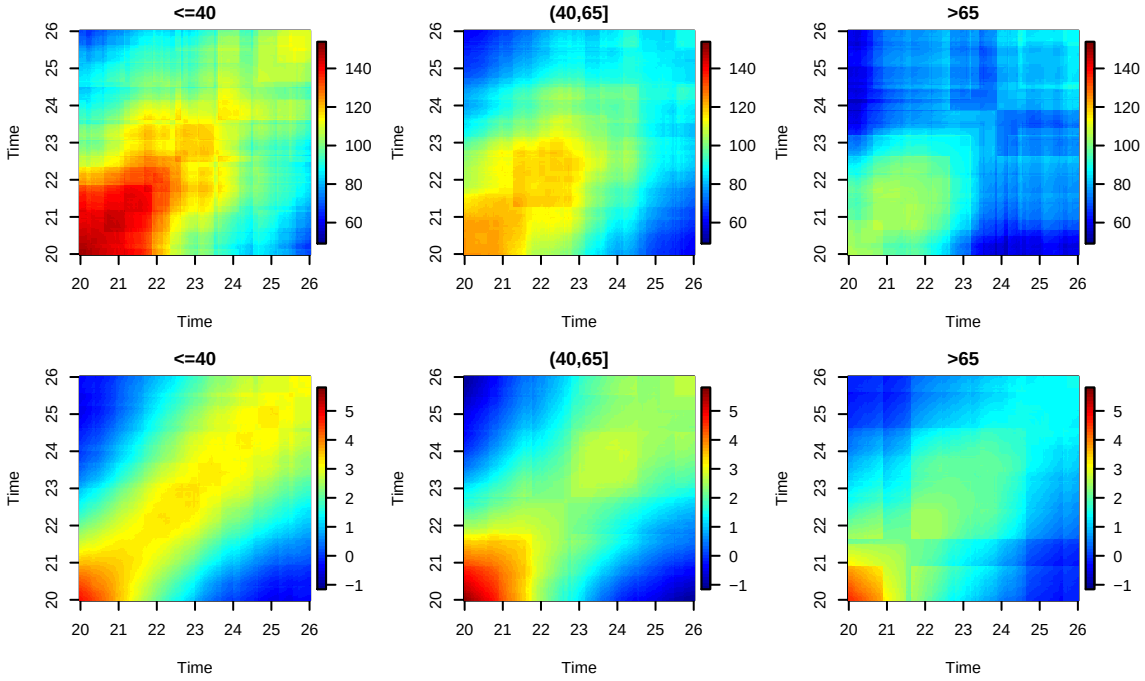

Fig. S1. Estimated covariance functions of heart rate profiles (top row) and of their derivatives (bottom row) in age groups.

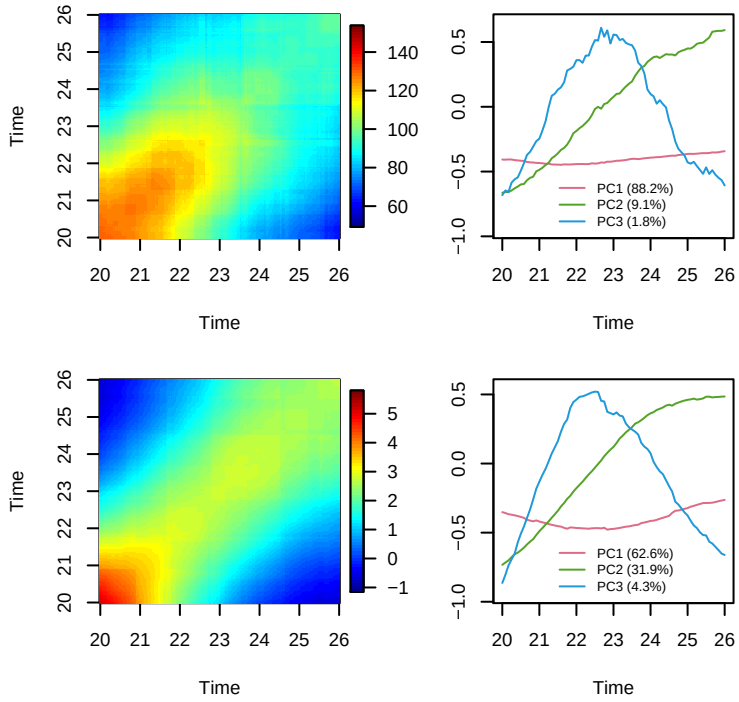

Fig. S2. The null estimate of the covariance function (left column) and its three leading principal components (right column) for heart rate profiles (top row) and for their first derivative (bottom row).

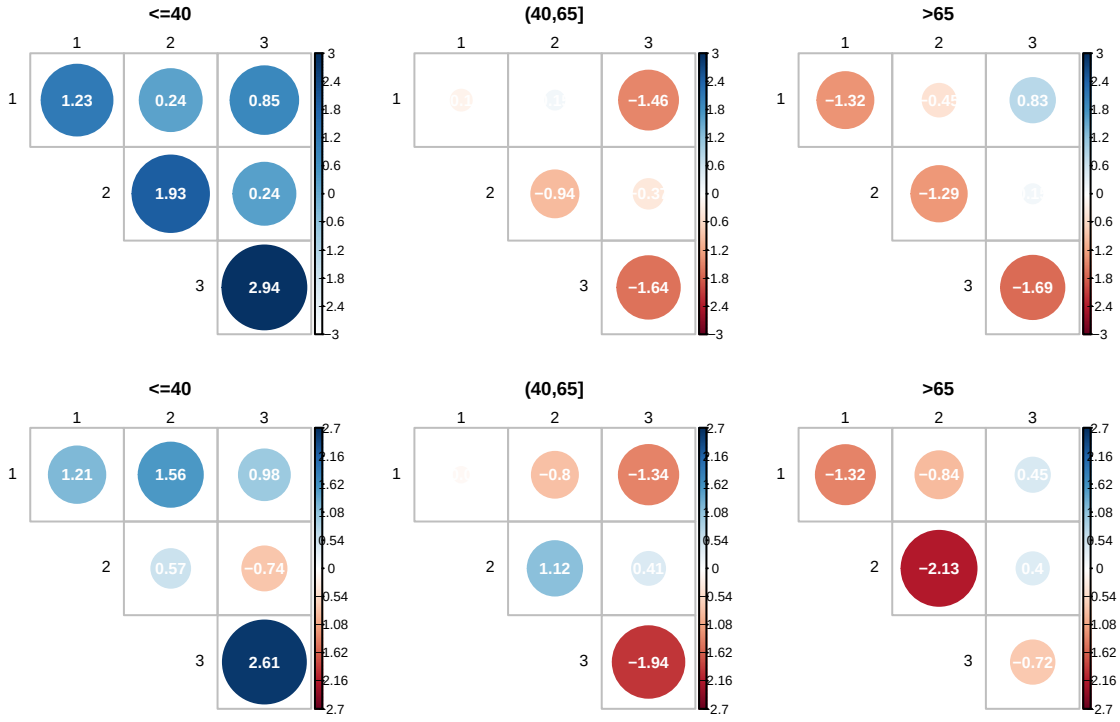

Fig. S3. Standardized components of the score vector for testing equal covariances contrasting age groups against the null for heart rate profiles (top row) and for their derivatives (bottom row).

National Grid Infrastructure provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042) is greatly appreciated.

References

Lloyd, C. J. (2005). Estimating test power adjusted for size. *Journal of Statistical Computation and Simulation*, 75(11):921–933.