Testing the Equality of Covariance Operators in Functional Samples

STEFAN FREMDT and JOSEF G. STEINEBACH
Mathematical Institute, University of Cologne

LAJOS HORVÁTH
Department of Mathematics, University of Utah

PIOTR KOKOSZKA
Department of Statistics, Colorado State University

ABSTRACT. We propose a non-parametric test for the equality of the covariance structures in two functional samples. The test statistic has a chi-square asymptotic distribution with a known number of degrees of freedom, which depends on the level of dimension reduction needed to represent the data. Detailed analysis of the asymptotic properties is developed. Finite sample performance is examined by a simulation study and an application to egg-laying curves of fruit flies.

Key words: asymptotic distribution, covariance operator, functional data, quadratic forms, two sample problem

1. Introduction

The last decade has seen increasing interest in methods of functional data analysis which offer novel and effective tools for dealing with problems where curves can naturally be viewed as data objects. The books by Ramsay & Silverman (2005) and Ramsay et al. (2009) offer comprehensive introductions to the subject, the collection by Ferraty & Romain (2011) reviews some recent developments focusing on advances in the relevant theory, while the monographs of Bosq (2000), Ferraty & Vieu (2006) and Horváth & Kokoszka (2012) develop the field in several important directions. Despite the emergence of many alternative ways of looking at functional data, and many dimension reduction approaches, the functional principal components (FPCs) still remain the most important starting point for many functional data analysis procedures, and Reiss & Ogden (2007), Gervini (2008), Yao & Müller (2010), Gabrys et al. (2010) are just a handful of illustrative references. The FPCs are the eigenfunctions of the covariance operator. This paper focuses on testing if the covariance operators of two functional samples are equal. By the Karhunen–Loève expansion, this is equivalent to testing if both samples have the same set of FPCs. Benko et al. (2009) developed bootstrap procedures for testing the equality of specific FPCs. Panaretos et al. (2010) proposed a test of the type we consider, but assuming that the curves have a Gaussian distribution. The main result of Panaretos et al. (2010) follows as a corollary of our more general approach (theorem 2). A generalization to non-Gaussian data was discussed in Panaretos et al. (2010, 2011). For some recent work confer also Boente et al. (2011) who studied a related approach together with a corresponding bootstrap procedure.

Despite their importance, two sample problems for functional data received relatively little attention. In addition to the work of Benko et al. (2009) and Panaretos et al. (2010), the relevant references are Horváth et al. (2009) and Horváth et al. (2012) who focus, respectively, on the regression kernels in functional linear models and the mean of functional data exhibiting temporal dependence. For a recent contribution, see also Gaines et al. (2011), who
use a likelihood ratio-type approach for testing the equality of two covariance operators. Clearly, if some population parameters of two functional samples are different, estimating them using the pooled sample may lead to spurious conclusions. Due to the importance of the FPCs, a relatively simple and non-parametric procedure for testing the equality of the covariance operators is called for.

The remainder of this paper is organized as follows. Section 2 sets out the notation and definitions. The construction of the test statistic and its asymptotic properties are developed in section 3. Section 4 reports the results of a simulation study and illustrates the procedure by application to egg-laying curves of Mediterranean fruit flies. The proofs of the asymptotic results of section 3 are given in section 5.

### 2. Preliminaries

Let \(X_1, X_2, \ldots, X_N\) be independent, identically distributed random variables with values in \(L_2[0, 1]\), the Hilbert space of square-integrable \(\mathbb{R}\)-valued functions on \([0, 1]\), and set \(EX_i(t) = \mu(t)\) and \(\text{cov}(X_i(t), X_i(s)) = C(t,s)\). We assume that another sample \(X_1^*, X_2^*, \ldots X_M^*\) is also available and let \(\bar{X}_i = EX_i(t)\) and \(C^*(t,s) = \text{cov}(X_i^*(t), X_i^*(s))\) for \(t, s \in [0, 1]\). We wish to test the null hypothesis

\[H_0 : C = C^*\]

against the alternative \(H_A\) that \(H_0\) does not hold.

A crucial assumption considering the asymptotics of our test procedure will be that

\[N/M + N \to \Theta \in (0, 1) \quad \text{as} \quad N, M \to \infty. \tag{1}\]

For the construction of our test procedure, we will use an estimate of the asymptotic pooled covariance operator \(\mathfrak{R}\) of the two given samples [cf. (4)] which is defined by the kernel

\[R(t,s) = \Theta C(t,s) + (1-\Theta)C^*(t,s).\]

In the case of samples \(\{X_i\}\) and \(\{X_i^*\}\) of Gaussian random functions, the latter approach has successfully been applied by Panaretos et al. (2010) to construct an asymptotic test for checking the equality of two covariance operators (see also Panaretos et al., 2011).

Denote by \((\lambda_1, \varphi_1), (\lambda_2, \varphi_2), \ldots\) the eigenvalue/eigenfunction pairs of \(\mathfrak{R}\), which are defined by

\[\lambda_k \varphi_k(t) = \Re \varphi_k(t) = \int_0^1 R(t,s) \varphi_k(s) ds, \quad t \in [0, 1], \quad 1 \leq k < \infty. \tag{2}\]

Throughout this paper, we assume

\[\lambda_1 > \lambda_2 > \cdots > \lambda_p > \lambda_{p+1}, \tag{3}\]

i.e. there exist at least \(p\) distinct (positive) eigenvalues. Under assumption (3), we can uniquely (up to signs) choose \(\varphi_1, \ldots, \varphi_p\) satisfying (2), if we require \(\|\varphi_i\| = 1\), where \(\|\cdot\|\) always denotes the \(L_2\)-norm, e.g. for \(x \in L_2([0,1])\),

\[\|x\| = \left( \int_0^1 x^2(t) dt \right)^{1/2}.\]

Thus, under (3), \(\{\varphi_i, 1 \leq i \leq p\}\) is an orthonormal system that can be extended to an orthonormal basis \(\{\varphi_i, 1 \leq i < \infty\}\).
If $H_0$ holds, then $(\hat{\lambda}_i, \phi_i), 1 \leq i < \infty,$ are also the eigenvalues/eigenfunctions of the covariance operators $\mathcal{C}$ of the first and $\mathcal{C}^*$ of the second sample. To construct a test statistic which converges under $H_0,$ we can therefore pool the two samples, as explained in section 3.

### 3. The test and the asymptotic results

Along the lines of Panaretos et al. (2010), our procedure is also based on projecting the observations onto a suitably chosen finite-dimensional space. To define this space, introduce the empirical pooled covariance operator $\hat{R}_{N,M}$ defined by the kernel

$$
\hat{R}_{N,M}(t,s) = \frac{1}{N+M} \left\{ \sum_{k=1}^{N} (X_k(t) - \bar{X}_N(t))(X_k(s) - \bar{X}_N(s)) \\
+ \sum_{k=1}^{M} (X^*_k(t) - \bar{X}^*_M(t))(X^*_k(s) - \bar{X}^*_M(s)) \right\},
$$

(4)

where

$$
\bar{X}_N(t) = \frac{1}{N} \sum_{k=1}^{N} X_k(t) \quad \text{and} \quad \bar{X}^*_M(t) = \frac{1}{M} \sum_{k=1}^{M} X^*_k(t)
$$

are the sample mean functions. Let $(\hat{\lambda}_i, \phi_i)$ denote the eigenvalues/eigenfunctions of $\hat{R}_{N,M},$ i.e.

$$
\hat{\lambda}_i \phi_i(t) = \hat{R}_{N,M}(t,s) \phi_i(s) \, ds, \quad t \in [0, 1], \quad 1 \leq i \leq N + M,
$$

with $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots.$ We can and will assume that the $\phi_i$ form an orthonormal system. We consider the projections

$$
\hat{a}_k(i) = \langle X_k - \bar{X}, \phi_i \rangle = \int_0^1 (X_k(t) - \bar{X}_N(t)) \phi_i(t) \, dt
$$

(5)

and

$$
\hat{a}_k^*(j) = \langle X^*_k - \bar{X}^*_M, \phi_j \rangle = \int_0^1 (X^*_k(t) - \bar{X}^*_M(t)) \phi_j(t) \, dt,
$$

(6)

where $\langle \cdot, \cdot \rangle$ denotes the inner product of two elements of the Hilbert space $L_2[0, 1].$ To test $H_0,$ we compare the matrices $\hat{\Lambda}_N$ and $\hat{\Lambda}^*_M$ with entries

$$
\hat{\Lambda}_N(i,j) = \frac{1}{N} \sum_{k=1}^{N} \hat{a}_k(i) \hat{a}_k(j), \quad 1 \leq i,j \leq p,
$$

and

$$
\hat{\Lambda}^*_M(i,j) = \frac{1}{M} \sum_{k=1}^{M} \hat{a}_k^*(i) \hat{a}_k^*(j), \quad 1 \leq i,j \leq p.
$$

We note that $\hat{\Lambda}_N(i,j) - \hat{\Lambda}^*_M(i,j)$ is the projection of $\hat{C}_N(t,s) - \hat{C}^*_M(t,s)$ in the direction of $\phi_i(t)\phi_j(s),$ where

$$
\hat{C}_N(t,s) = \frac{1}{N} \sum_{k=1}^{N} (X_k(t) - \bar{X}_N(t))(X_k(s) - \bar{X}_N(s))
$$
and
\[ C_M^*(t, s) = \frac{1}{M} \sum_{k=1}^{M} (X_k^*(t) - \tilde{X}_M(t))(X_k^*(s) - \tilde{X}_M(s)) \]
are the empirical covariances of the two samples.

We create the vector \( \hat{\xi}_{N,M} \) from the columns below the diagonal of \( \hat{\Lambda}_N - \hat{\Lambda}_M^* \) as follows:
\[
\hat{\xi}_{N,M} = \text{vech}\left( \hat{\Lambda}_N - \hat{\Lambda}_M^* \right) = \begin{pmatrix}
\hat{\Lambda}_N(1,1) - \hat{\Lambda}_M^*(1,1) \\
\hat{\Lambda}_N(2,1) - \hat{\Lambda}_M^*(2,1) \\
\vdots \\
\hat{\Lambda}_N(p,p) - \hat{\Lambda}_M^*(p,p)
\end{pmatrix}.
\]

(7)

For the properties of the vech operator, we refer to Abadir & Magnus (2005).

Next, we estimate the asymptotic covariance matrix of \( (M^N)(N+M)^{1/2}\hat{\xi}_{N,M} \). Note that, in general, this estimate differs from the one which was used in the Gaussian case (cf. Panaretos et al., 2010, and theorem 2). Let
\[
\hat{L}_{N,M}(k,k') = (1 - \Theta_{N,M}) \left\{ \frac{1}{N} \sum_{\ell=1}^{N} \hat{a}_\ell(i)\hat{a}_\ell(j)\hat{a}_\ell(i')\hat{a}_\ell(j') - \langle \hat{C}_N \hat{\phi}_1, \hat{\phi}_1 \rangle \langle \hat{C}_N \hat{\phi}_1, \hat{\phi}_1 \rangle \right\} \\
+ \Theta_{N,M} \left\{ \frac{1}{M} \sum_{\ell=1}^{M} \hat{a}_\ell(i)\hat{a}_\ell(j)\hat{a}_\ell(i')\hat{a}_\ell(j') - \langle \hat{C}_M \hat{\phi}_1, \hat{\phi}_1 \rangle \langle \hat{C}_M \hat{\phi}_1, \hat{\phi}_1 \rangle \right\},
\]
where \( i, j, i', j' \) depend on \( k, k' \) (see below), and \( \hat{C}_N (\hat{C}_M^*) \) is interpreted as an operator with \( \hat{C}_N \) defined as
\[
\hat{C}_N \hat{\phi}_1 = \int_0^1 \hat{C}_N(t,s)\hat{\phi}_1(s)\,ds.
\]

(An analogous definition holds for \( \hat{C}_M^* \).) From this definition it follows that
\[
\langle \hat{C}_N \hat{\phi}_1, \hat{\phi}_1 \rangle = \frac{1}{N} \sum_{\ell=1}^{N} \hat{a}_\ell(i)\hat{a}_\ell(j).
\]

There are other ways to estimate the asymptotic covariance matrix. We note that one can use \( \hat{L}_{N,M}(k,k') \) instead of \( \hat{L}_{N,M}(k,k') \), where \( \hat{L}_{N,M}(k,k') \) is defined like \( \hat{L}_{N,M}(k,k') \), but \( \langle \hat{C}_N \hat{\phi}_1, \hat{\phi}_1 \rangle \) and \( \langle \hat{C}_M \hat{\phi}_1, \hat{\phi}_1 \rangle \) are replaced with 0 if \( i \neq j \) and \( \hat{\lambda}_j \) if \( i = j \). In the same spirit, \( \langle \hat{C}_N \hat{\phi}_1, \hat{\phi}_1 \rangle \) and \( \langle \hat{C}_M \hat{\phi}_1, \hat{\phi}_1 \rangle \) are replaced with 0 for \( i' \neq j' \) and \( \hat{\lambda}_j \) if \( i' = j' \).

The index \((i,j)\) is computed from \( k \) in the following way: Let
\[
k' = \frac{p(p+1)}{2} - k + 1, \quad i' = p - i + 1 \quad \text{and} \quad j' = p - j + 1.
\]

(8)

We look at an upper triangle matrix \( (a_{i',j'}) \). Then, for column \( j' \), we have that \((j' - 1)/2 < k \leq j'(j' + 1)/2\). Thus,
\[
\hat{a} = \left\lfloor \sqrt{2k' + \frac{1}{4}} - \frac{1}{2} \right\rfloor \quad \text{and} \quad i' = k' - (j' - 1)/2,
\]
where \([r] = \min\{k \in \mathbb{Z} : k \geq r\}\) for \( r \in \mathbb{R} \). Consequently, the index \((i,j)\) can be computed from \( k \) via
\[ j = p - \left[ \sqrt{p(p+1) - 2k + \frac{9}{4} - \frac{1}{2}} \right] + 1 \text{ and } i = k + p - pj + \frac{j(j-1)}{2}. \] (9)

With the above notation, we can formulate the main result of this paper in the non-Gaussian case. The latter case has briefly been mentioned (without any mathematical details) in the concluding remarks of Panaretos et al. (2010) (see also Panaretos et al., 2011).

**Theorem 1.** We assume that \( H_0 \), (1) and (3) hold, and
\[
\int_0^1 E(X_i(t))^4 dt < \infty, \quad \int_0^1 E(X_i^*(t))^4 dt < \infty.
\] (10)

Then,
\[
\frac{N M \hat{\xi}^T_{N,M} \hat{L}_{N,M}^{-1} \hat{\xi}_{N,M}}{N + M} \xrightarrow{D} \chi^2_{p(p+1)/2}, \quad \text{as } N, M \to \infty,
\]
where \( \chi^2_{p(p+1)/2} \) stands for a \( \chi^2 \) random variable with \( p(p+1)/2 \) degrees of freedom.

Theorem 1 implies that the null hypothesis is rejected if the test statistic
\[
\hat{T}_1 = \frac{N M}{N + M} \hat{\xi}^T_{N,M} \hat{L}_{N,M}^{-1} \hat{\xi}_{N,M}
\]
exceeds a critical quantile of the chi-square distribution with \( p(p+1)/2 \) degrees of freedom. If both samples are Gaussian random processes, the quadratic form \( \hat{\xi}^T_{N,M} \hat{L}_{N,M}^{-1} \hat{\xi}_{N,M} \) can be replaced with the normalized sum of the squares of \( \hat{\Delta}_{N,M}(i,j) - \hat{\Delta}_{N,M}^*(i,j) \), as stated in theorem 2 (cf. Panaretos et al., 2010).

**Theorem 2.** If \( X_1, X_1^* \) are Gaussian processes and the conditions of theorem 1 are satisfied, then, as \( N, M \to \infty \),
\[
\hat{T}_2 = \frac{N M}{N + M} \sum_{1 \leq i, j \leq p} \frac{1}{2} (\hat{\Delta}_{N}(i,j) - \hat{\Delta}_{N}^*(i,j))^2 \frac{1}{\hat{\lambda}_i \hat{\lambda}_j} \xrightarrow{D} \chi^2_{p(p+1)/2}.
\]

Observe that the statistic \( \hat{T}_2 \) can be written as
\[
\hat{T}_2 = \frac{N M}{N + M} \left\{ \sum_{1 \leq i < j \leq p} \frac{(\hat{\Delta}_{N}(i,j) - \hat{\Delta}_{N}^*(i,j))^2}{\hat{\lambda}_i \hat{\lambda}_j} + \sum_{i=1}^{p} \frac{(\hat{\Delta}_{N}(i,i) - \hat{\Delta}_{N}^*(i,i))^2}{2 \hat{\lambda}_i^2} \right\}.
\]

Next, we discuss the asymptotic consistency of the testing procedure based on theorem 1. Analogously to the definition of \( \hat{\xi}_{N,M} \) we define the vector \( \xi = (\xi(1), \ldots, \xi(p(p+1)/2)) \) using the columns of the matrix
\[
D = \left( \int_0^1 \int_0^1 (C(t,s) - C^*(t,s)) \phi_j(t) \phi_j(s) \, dt \, ds \right)_{i,j=1, \ldots, p}
\] (11)
instead of \( \hat{\Delta}_{N} - \hat{\Delta}_{N}^* \), i.e.
\[
\xi = \text{vech}(D).
\]
**Theorem 3.** We assume that $H_A$, (1), (3) and (10) hold. Then, there exist random variables $\hat{h}_1 = \hat{h}_1(N, M), \ldots, \hat{h}_{p(p+1)/2} = \hat{h}_{p(p+1)/2}(N, M)$, taking values in $\{-1, 1\}$ such that, as $N, M \to \infty$,

$$\max_{1 \leq i \leq p(p+1)/2} |\hat{\xi}_{N, M}^i(i) - \hat{h}_i(i)| = o_p(1)$$

and therefore

$$|\hat{\xi}_{N, M}| \xrightarrow{P} |\xi|,$$

where $|\cdot|$ denotes the Euclidean norm. If $\xi \neq 0$ and the $p$ largest eigenvalues of $C$ and $C^*$ are positive, we also have

$$\hat{T}_1 \xrightarrow{P} \infty, \text{ as } N, M \to \infty.$$

The assumption that the $p$ largest eigenvalues of $C$ and $C^*$ are positive implies that the random functions $X_i, i = 1, \ldots, N$, and $X^*_j, j = 1, \ldots, M$, are not included in a $(p-1)$-dimensional subspace.

**Remark 1.** The application of the test requires the selection of the number $p$ of the empirical FPCs to be used. A rule of thumb is to choose $p$ so that the first $p$ empirical FPCs in each sample (i.e. those calculated as the eigenfunctions of $\hat{C}_N$ and $\hat{C}_M^*$) explain about 85–90 per cent of the variance in each sample. Choosing $p$ too large generally negatively affects the finite sample performance of tests of this type, and for this reason, we do not study asymptotics as $p$ tends to infinity. It is often illustrative to apply the test for a range of the values of $p$; each $p$ specifies a level of relevance of differences in the curves or kernels. A good practical approach is to look at the Karhunen–Loève approximations of the curves in both samples, and choose $p$ which gives approximation errors that can be considered unimportant. Cross validation has also been suggested in the literature without investigating its properties in detail. For a more formal discussion of this selection, confer also section 3.3 in Panaretos et al. (2010).

4. A simulation study and an application

We first describe the results of a simulations study designed to evaluate finite sample properties of the tests based on the statistics $\hat{T}_1$ and $\hat{T}_2$. The emphasis is on verifying the advantage of a non-parametric procedure, i.e. to see the ‘robustness’ to the violation of the assumption of normality. We simulated Gaussian curves as Brownian motions and Brownian bridges, and non-Gaussian curves via

$$X(t) = A \sin(\pi t) + B \sin(2\pi t) + C \sin(4\pi t),$$

where $A = 5Y_1$, $B = 3Y_2$, $C = Y_3$ and $Y_1, Y_2, Y_3$ are independent $t_5$-distributed random variables (similarly $X^*(t)$ for the second sample). All curves were simulated at 1000 equidistant points in the interval $[0, 1]$, and transformed into functional data objects using the Fourier basis with 49 basis functions. For each data generating process, we used one thousand replications.

Table 1 shows the empirical sizes for non-Gaussian data. The test based on $\hat{T}_2$ has severely inflated size, due to the violation of the assumption of normality. As documented in Panaretos et al. (2010), and confirmed by our own simulations, this test has very good empirical size when the data are Gaussian. The test based on $\hat{T}_1$ is conservative, especially for smaller sample sizes. This is true for both Gaussian and non-Gaussian data; there is not much difference in the empirical size of this test for different data-generating processes.
Table 1. Empirical sizes of the tests based on statistics $\hat{T}_1$ and $\hat{T}_2$ for non-Gaussian data. The curves in each sample were generated according to (15)

| Sample sizes | $\hat{T}_1$ | $\hat{T}_2$ |
|--------------|-------------|-------------|
|              | 1%          | 5%          | 10%         | 1%          | 5%          | 10%         |
| $p = 2$      |             |             |             |             |             |             |
| $N = M = 100$| 0.005       | 0.028       | 0.061       | 0.152       | 0.275       | 0.380       |
| $N = M = 200$| 0.003       | 0.021       | 0.058       | 0.163       | 0.314       | 0.402       |
| $N = M = 1000$| 0.002      | 0.021       | 0.056       | 0.190       | 0.313       | 0.426       |
| $p = 3$      |             |             |             |             |             |             |
| $N = M = 100$| 0.004       | 0.028       | 0.065       | 0.167       | 0.332       | 0.434       |
| $N = M = 200$| 0.004       | 0.024       | 0.064       | 0.194       | 0.338       | 0.423       |
| $N = M = 1000$| 0.004     | 0.028       | 0.070       | 0.240       | 0.384       | 0.484       |

Table 2 gives an example of the empirical power of the test based on statistic $\hat{T}_1$. The test was carried out for two equally sized samples of 200, 500 and 1000 realizations, respectively, of (15) for the first sample and scaled versions of (15), i.e. $X^*(t) = cX(t)$, for the second sample. The results are displayed for a selection of values for the scaling parameter $c$. It can be seen that in all cases the power increases with the sample size. As can be expected, the convergence of the power towards 1 improves for larger deviations ($c \neq 1$) from the null hypothesis. Since, due to the inflated size of the test based on $\hat{T}_2$ in the non-Gaussian case (cf. Table 1), its power is (misleadingly) higher than that of the test based on $\hat{T}_1$ and thus will not be displayed here. We also studied a Monte Carlo version of the test based on the statistic $\hat{T}_3 = NM(N + M)^{-\frac{1}{2}} \hat{\theta}_{N,M}^T \hat{\theta}_{N,M}$ and found that its finite sample properties were similar to those of the test based on $\hat{T}_1$.

We now describe the results of the application of both tests to an interesting data set consisting of egg-laying trajectories of Mediterranean fruit flies (medflies). The data were kindly made available to us by Hans–Georg Müller. This data set has been extensively studied in biological and statistical literature; see Müller & Stadtmüller (2005) and references therein. We consider 534 egg-laying curves of medflies who lived at least 34 days, but we only consider the egg-laying activities on the first 30 days. We examined two versions of these egg-laying curves. The curves are scaled such that the functions in either version are defined on

Table 2. Power of the test based on statistic $\hat{T}_1$ for non-Gaussian data. The curves in the equally sized samples were generated according to (15) in the first sample and as a scaled version of (15) in the second sample, i.e. $X^*(t) = cX(t)$

| $c = 0.8$ | $c = 0.9$ |
|-----------|-----------|
| $p = 2$   | $p = 3$   | $p = 2$   | $p = 3$   |
| $N, M$    | 1%        | 5%        | 10%       | 1%        | 5%        | 10%       | 1%        | 5%        | 10%       |
| 200       | 0.061     | 0.289     | 0.491     | 0.033     | 0.181     | 0.351     | 0.007     | 0.041     | 0.106     |
| 500       | 0.532     | 0.811     | 0.917     | 0.485     | 0.804     | 0.896     | 0.032     | 0.159     | 0.274     |
| 1000      | 0.947     | 0.986     | 0.993     | 0.965     | 0.997     | 0.998     | 0.113     | 0.360     | 0.519     |
| $c = 1.2$ | $c = 1.1$ |
| $p = 2$   | $p = 3$   | $p = 2$   | $p = 3$   |
| $N, M$    | 1%        | 5%        | 10%       | 1%        | 5%        | 10%       | 1%        | 5%        | 10%       |
| 200       | 0.017     | 0.160     | 0.295     | 0.015     | 0.124     | 0.231     | 0.001     | 0.041     | 0.099     |
| 500       | 0.244     | 0.570     | 0.718     | 0.186     | 0.524     | 0.689     | 0.009     | 0.096     | 0.222     |
| 1000      | 0.762     | 0.913     | 0.952     | 0.760     | 0.956     | 0.980     | 0.091     | 0.295     | 0.461     |

© 2012 Board of the Foundation of the Scandinavian Journal of Statistics.
the interval $[0, 1]$. Version 1 curves (denoted $X_i(t)$) are the absolute counts of eggs laid by fly $i$ on day $\lfloor 30t \rfloor$. Version 2 curves (denoted $Y_i(t)$) are the counts of eggs laid by fly $i$ on day $\lfloor 30t \rfloor$ relative to the total number of eggs laid in the lifetime of fly $i$. The 534 flies are classified into long-lived, i.e. those who lived 44 days or longer, and short-lived, i.e. those who died before the end of the 43rd day after birth. In the data set, there are 256 short-lived and 278 long-lived flies. This classification naturally defines two samples: Sample 1: the egg-laying curves $\{X_i(t) | 0 \leq t \leq 1, i = 1, 2, \ldots, 256\}$ of the short-lived flies. Sample 2: the egg-laying curves $\{X_j^*(t) | 0 < t \leq 30, j = 1, 2, \ldots, 278\}$ of the long-lived flies. The egg-laying curves are very irregular; Fig. 1 shows ten (smoothed) curves of short- and long-lived flies for version 1, and Fig. 2 shows ten (smoothed) curves for version 2 (both using a B-spline basis for the representation).

Table 3 shows the $p$-values for the absolute egg-laying counts (version 1). For the statistic $\hat{T}_1$, the null hypothesis cannot be rejected irrespective of the choice of $p$. For the statistic $\hat{T}_2$, the result of the test varies depending on the choice of $p$. As explained in section 3, the usual recommendation is to use the values of $p$ which explain 85 to 90 per cent of the variance.

Fig. 1. Ten randomly selected smoothed egg-laying curves of short-lived medflies (left panel) and ten such curves for long-lived medflies (right panel).

Fig. 2. Ten randomly selected smoothed egg-laying curves of short-lived medflies (left panel) and ten such curves for long-lived medflies (right panel), relative to the number of eggs laid in the fly’s lifetime.
Table 3. The p-values (in per cent) of the test based on statistics $\hat{T}_1$ and $\hat{T}_2$ applied to absolute medfly data. Here $f_p$ denotes the fraction of the sample variance explained by the first $p$ FPCs, i.e. $f_p = \frac{\sum_{k=1}^{p} \hat{\lambda}_k}{\sum_{k=1}^{N+M} \hat{\lambda}_k}$.

| $p$ | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| $\hat{T}_1$ | 82.70 | 36.22 | 30.59 | 63.84 | 37.71 | 39.03 | 33.77 | 34.77 |
| $\hat{T}_2$ | 0.54 | 0.13 | 0.11 | 0.12 | 0.02 | 0.00 | 0.00 | 0.00 |
| $f_p$ | 72.93 | 78.36 | 81.87 | 83.94 | 85.62 | 87.08 | 88.49 | 89.72 |

For such values of $p$, $\hat{T}_2$ leads to a clear rejection. Since this test has however overinflated size, we conclude that there is little evidence that the covariance structures of version 1 curves for long- and short-lived flies are different. For the version 2 curves, the statistic $\hat{T}_2$ yields $p$-values equal to zero (in machine precision), potentially indicating that the covariance structures for the short- and long-lived flies are different. The assumption of a normal distribution is however questionable, as the QQ-plots in Fig. 3 show. These QQ-plots are constructed for the inner products $\langle Y_i, e_k \rangle$ and $\langle Y^*_i, e_k \rangle$, where the $Y_i$ are the curves from one of the samples (we cannot pool the data to construct QQ-plots because we test if the stochastic structures are different), and $e_k$ is the $k$th element of the Fourier basis. The normality of a functional sample implies the normality of all projections onto a complete orthonormal system.

For $\langle X_i, e_k \rangle$, the QQ-plots show a strong deviation from a straight line for some projections. Almost all projections $\langle Y_i, e_k \rangle$ have QQ-plots indicating a strong deviation from normality. It is therefore important to apply the non-parametric test based on the statistic $\hat{T}_1$. The corresponding $p$-values for version 2 are displayed in Table 4. For most values of $p$, these $p$-values

![Fig 3. Normal QQ-plots for the scores of the version 2 medfly data with respect to the first two Fourier basis functions. Left – sample 1, Right – sample 2.](image)
indicate the rejection of \( H_0 \). Many of them hover around the 5 per cent level, but since the test is conservative, we can with confidence view them as favouring \( H_A \).

The above application confirms the properties of the statistics established through the simulation study. It shows that while there is little evidence that the covariance structures for the absolute counts are different, there is strong evidence that they are different for relative counts.

5. Proofs of the results of section 3

The proof of theorem 1 follows from several lemmas, which we establish first. We can and will assume without loss of generality that \( \mu(t) = \mu'(t) = 0 \) for all \( t \in [0, 1] \).

We will use the identity
\[
\frac{1}{N^{1/2}} \sum_{k=1}^{N} (X_k(t) - \bar{X}_N(t)) (X_k(s) - \bar{X}_N(s)) = \frac{1}{N^{1/2}} \sum_{k=1}^{N} X_k(t) X_k(s) - N^{1/2} \bar{X}_N(t) \bar{X}_N(s),
\]
and an analogous identity for the second sample.

Our first lemma establishes bounds in probability which will often be used in the proofs.

**Lemma 1.** Under the assumptions of theorem 1, as \( N, M \to \infty \),
\[
\left\| N^{-1/2} \sum_{k=1}^{N} \{X_k(t)X_k(s) - C(t, s)\} \right\| = O_P(1),
\]
\[
\left\| N^{1/2} \bar{X}_N(t) \right\| = O_P(1),
\]
and
\[
\left\| M^{-1/2} \sum_{k=1}^{M} \{X_k^*(t)X_k^*(s) - C^*(t, s)\} \right\| = O_P(1),
\]
\[
\left\| M^{1/2} \bar{X}_M^*(t) \right\| = O_P(1),
\]
where here and in the sequel the notation \( \| \cdot \| \) is also used for the corresponding norm in \( L_2([0, 1]^2) \).
Lemma 3. If the conditions of theorem 1 are satisfied, then, as $N,M \to \infty$, 
\[
\max_{1 \leq i \leq p} |\hat{\lambda}_i - \lambda_i| = O_P \left( (N + M)^{-1/2} \right)
\]
and 
\[
\max_{1 \leq i \leq p} \|\hat{\phi}_i - \phi_i\| = O_P \left( (N + M)^{-1/2} \right),
\]
where 
\[
\hat{\phi}_i = \hat{\phi}_i(N, M) = \text{sign}(\langle \hat{\phi}_i, \phi_i \rangle).
\]
Proof. These estimates are also well-known (cf., e.g. Bosq, 2000, lemma 4.3 and assertion (4.43), or Horváth & Kokoszka, 2012, lemmas 2.2–2.3). Note that the first rate above is independent of \( p \), whereas the second one may actually depend on the projection dimension \( p \).

Lemma 3 now allows us to replace the estimated eigenfunctions by their population counterparts. The random signs \( \hat{c}_i \) must appear in the formulation of lemma 4, but they cancel in the subsequent results.

**Lemma 4.** If the conditions of theorem 1 are satisfied, then, for all \( 1 \leq i, j \leq p \), as \( N, M \to \infty \),

\[
\left( \frac{NM}{N + M} \right)^{1/2} (\hat{\Delta}_N(i,j) - \hat{\Delta}_M(i,j))
= \left( \frac{NM}{N + M} \right)^{1/2} \left\{ \frac{1}{N} \sum_{k=1}^{N} \langle X_k, \hat{\phi}_i \rangle \langle X_k, \hat{\phi}_j \rangle - \frac{1}{M} \sum_{k=1}^{M} \langle X_k^*, \hat{\phi}_i \rangle \langle X_k^*, \hat{\phi}_j \rangle \right\} + o_p(1).
\]

**Proof.** We write

\[
\frac{1}{N} \sum_{k=1}^{N} \langle X_k, \hat{\phi}_i \rangle \langle X_k, \hat{\phi}_j \rangle - \int_0^1 \int_0^1 C(t,s)\hat{\phi}_i(t)\hat{\phi}_j(s) \, dt \, ds
= N^{1/2} \int_0^1 \int_0^1 \left\{ \frac{1}{N^{1/2}} \sum_{k=1}^{N} (X_k(t)X_k(s) - C(t,s)) \right\} \hat{\phi}_i(t)\hat{\phi}_j(s) \, dt \, ds.
\]

Using lemmas 1–3 we get

\[
\left| \int_0^1 \int_0^1 \left\{ \frac{1}{N^{1/2}} \sum_{k=1}^{N} (X_k(t)X_k(s) - C(t,s)) \right\} (\hat{\phi}_i(t)\hat{\phi}_j(s) - \hat{c}_i \phi_i(t)\hat{c}_j \phi_j(s)) \, dt \, ds \right|
\leq \left( \int_0^1 \int_0^1 \left\{ \frac{1}{N^{1/2}} \sum_{k=1}^{N} (X_k(t)X_k(s) - C(t,s)) \right\}^2 \, dt \, ds \right)^{1/2}
\times \left( \int_0^1 \int_0^1 (\hat{\phi}_i(t) - \hat{c}_i \phi_i(t))^2 \hat{\phi}_j^2(s) \, dt \, ds \right)^{1/2}
+ \left( \int_0^1 \int_0^1 \left\{ \frac{1}{N^{1/2}} \sum_{k=1}^{N} (X_k(t)X_k(s) - C(t,s)) \right\}^2 \, dt \, ds \right)^{1/2}
\times \left( \int_0^1 \int_0^1 \hat{\phi}_i^2(t)(\hat{\phi}_i(s) - \hat{c}_i \phi_i(s))^2 \, dt \, ds \right)^{1/2}

= \frac{1}{N^{1/2}} \sum_{k=1}^{N} (X_k(t)X_k(s) - C(t,s)) \left\| \hat{\phi}_i - \hat{c}_i \phi_i \right\| + \left\| \hat{\phi}_j - \hat{c}_j \phi_j \right\|
= o_p(1).
\]
Similar arguments give that

\[ \left| \int_0^1 \int_0^1 \left\{ \frac{1}{M^{1/2}} \sum_{k=1}^M (X^*_k(t)X^*_k(s) - C^*(t,s)) \right\} \left\{ \hat{\phi}_j(t)\hat{\phi}_j(s) - \hat{c}_j(t)\hat{c}_j(s) \right\} \, dt \, ds \right| = o_p(1). \]

Since \( C = C^* \), the lemma is proven.

The previous lemmas isolated the main terms in the differences \( \hat{\Delta}_N(i,j) - \hat{\Delta}_M^*(i,j) \). The following lemma describes the limits of these main terms (without the random signs).

**Lemma 5.** If the conditions of theorem 1 are satisfied, then, as \( N, M \to \infty \),

\[ \{ \Delta_{N,M}(i,j), 1 \leq i, j \leq p \} \overset{D}{\to} \{ \Delta(i,j), 1 \leq i, j \leq p \}, \]

where

\[ \Delta_{N,M}(i,j) = \left( \frac{NM}{N+M} \right)^{1/2} \left\{ \frac{1}{N} \sum_{k=1}^N \langle X_k, \phi_i \rangle \langle X_k, \phi_j \rangle - \frac{1}{M} \sum_{k=1}^M \langle X^*_k, \phi_i \rangle \langle X^*_k, \phi_j \rangle \right\}, \]

and \( \{ \Delta(i,j), 1 \leq i, j \leq p \} \) is a Gaussian matrix with \( E\Delta(i,j) = 0 \) and

\[ E\Delta(i,j)\Delta(i',j') = (1 - \Theta) (E\langle X_1, \phi_i \rangle \langle X_1, \phi_j \rangle \langle X_1, \phi_{i'} \rangle \langle X_1, \phi_{j'} \rangle) \]

\[ - E\langle X_1, \phi_i \rangle \langle X_1, \phi_{i'} \rangle E\langle X_1, \phi_j \rangle \langle X_1, \phi_{j'} \rangle) \]

\[ + \Theta (E\langle X^*_1, \phi_i \rangle \langle X^*_1, \phi_j \rangle \langle X^*_1, \phi_{i'} \rangle \langle X^*_1, \phi_{j'} \rangle) \]

\[ - E\langle X^*_1, \phi_i \rangle \langle X^*_1, \phi_{i'} \rangle E\langle X^*_1, \phi_j \rangle \langle X^*_1, \phi_{j'} \rangle) \].

**Proof.** First we note that

\[ E\langle X_1, \phi_i \rangle \langle X_1, \phi_j \rangle = E\langle X^*_1, \phi_{i'} \rangle \langle X^*_1, \phi_{j'} \rangle = \begin{cases} 0 & \text{if } i \neq j, \\ \lambda_i & \text{if } i = j. \end{cases} \]

Since \( E\langle X_1, \phi_i \rangle \langle X_1, \phi_{i'} \rangle \rangle^2 < \infty \) and \( E\langle X^*_1, \phi_{i'} \rangle \rangle^2 < \infty \), the multivariate central limit theorem implies the result.

Finally, we need an asymptotic approximation to the covariances \( \hat{\Delta}_{N,M}(k,k') \). Let

\[ \hat{L}_{N,M}(k,k') = (1 - \Theta_{N,M}) \left\{ \frac{1}{N} \sum_{\ell=1}^N a_\ell(i)a_\ell(j)a_\ell(i')a_\ell(j') - \langle \hat{E}_N \phi_i, \phi_j \rangle \langle \hat{E}_N \phi_{i'}, \phi_{j'} \rangle \right\} \]

\[ + \Theta_{N,M} \left\{ \frac{1}{M} \sum_{\ell=1}^M a^*_\ell(i)a^*_\ell(j)a^*_\ell(i')a^*_\ell(j') - \langle \hat{E}_M \phi_i, \phi_j \rangle \langle \hat{E}_M \phi_{i'}, \phi_{j'} \rangle \right\}, \]

where

\[ a_\ell(i) = \langle X_\ell, \phi_i \rangle \text{ and } a^*_\ell(i) = \langle X^*_\ell, \phi_i \rangle, \]

and \( i, j, i', j' \) are determined from \( k \) and \( k' \) as in (8) and (9).

**Lemma 6.** If the conditions of theorem 1 are satisfied, then for all \( 1 \leq k, k' \leq p(p+1)/2 \),

\[ \hat{L}_{N,M}(k,k') - \hat{c}_\ell \hat{c}_\ell \hat{c}_\ell \hat{c}_\ell L_{N,M}(k,k') = o_p(1) \text{ as } N, M \to \infty, \]

where \( (i, j) \) and \( (i', j') \) are determined from \( k \) and \( k' \) as in (8) and (9).

**Proof.** The result follows from lemma 3 along the lines of the proof of lemma 4.
Proof of theorem 1. According to lemma 2 and lemmas 4–6, the asymptotic distribution of $\hat{\Sigma}_{N,M}^T L_{N,M} \hat{\Sigma}_{N,M}$ does not depend on the signs $\hat{c}_1, \ldots, \hat{c}_p$, so it is sufficient to prove the result for $\hat{c}_1 = \cdots = \hat{c}_p = 1$. The law of large numbers yields that

$$L_{N,M}(k,k') \overset{p}{\to} L(k,k'),$$

where

$$L(k,k') = (1 - \Theta) \{ E(a_1(i)a_1(j)a_1(i')a_1(j')) - E(a_1(i)a_1(j)a_1(i')a_1(j')) \} + \Theta \{ E(\hat{a}_1(i)\hat{a}_1(j)\hat{a}_1(i')\hat{a}_1(j')) - E(a_1(i)a_1(j)a_1(i')a_1(j')) \}.$$  

(22)

The result then follows from lemmas 1–5. For more details, we refer to Panaretos et al. (2010).

Proof of theorem 2. In the case of Gaussian observations, $\Delta(i,j)$, $1 \leq i \leq j \leq p$, are independent normal random variables with mean 0 and

$$E\Delta^2(i,j) = \begin{cases} \hat{a}_i \hat{a}_j & \text{if } i \neq j, \\ 2\hat{a}_i^2 & \text{if } i = j. \end{cases}$$

Now the result follows from lemmas 1–5. For more details, we refer to Panaretos et al. (2010).

Proof of theorem 3. First, we observe that by the law of large numbers we have

$$\int_0^1 \int_0^1 (\hat{\Phi}_{N,M}(t,s) - R(t,s))^2 \, dt \, ds = o_P(1).$$

Hence, using the result in section VI.1. of Gohberg et al. (1990), we get that

$$\max_{1 \leq i \leq p} |\hat{\lambda}_i - \lambda_i| = o_P(1)$$

(23)

and

$$\max_{1 \leq i \leq p} \|\hat{\varphi}_i - \varphi_i\| = o_P(1),$$

(24)

where $\hat{c}_i = \hat{c}_i(N,M) = \text{sign}(\hat{\varphi}_i, \varphi_i)$. Relations (23) and (24) show that lemma 3 remains true. It follows from the law of large numbers and (24) that for all $1 \leq i, j \leq p$

$$\left| \Delta_N(i,j) - \Delta^*_M(i,j) - \hat{c}_i \hat{c}_j \int_0^1 \int_0^1 (C(t,s) - C^*(t,s)) \varphi_j(t) \varphi_j(s) \, dt \, ds \right|$$

$$= \left| \int_0^1 \int_0^1 \left( \hat{C}_N(t,s) - \hat{C}_M(t,s) \right) \hat{\varphi}_j(t) \hat{\varphi}_j(s) \, dt \, ds - \hat{c}_i \hat{c}_j \int_0^1 \int_0^1 (C(t,s) - C^*(t,s)) \varphi_j(t) \varphi_j(s) \, dt \, ds \right|$$

$$\leq \left| \int_0^1 \int_0^1 \left( \hat{C}_N(t,s) - C(t,s) - \left( \hat{C}_M(t,s) - C^*(t,s) \right) \right) \hat{\varphi}_j(t) \hat{\varphi}_j(s) \, dt \, ds \right|$$

$$+ \left| \int_0^1 \int_0^1 (C(t,s) - C^*(t,s)) (\varphi_j(t) \varphi_j(s) - \hat{c}_i \hat{c}_j \varphi_j(t) \varphi_j(s)) \, dt \, ds \right|$$

$$\leq \| \hat{C}_N - C \| + \| \hat{C}_M - C^* \| + \| C - C^* \| \| \hat{\varphi}_j \| \| \hat{\varphi}_j - \varphi_j \| \| \hat{c}_i \| \| \hat{c}_j \|$$

$$= o_P(1),$$

where the fact that $\| \varphi_j \| = 1 = \| \hat{\varphi}_j \|$ was used. Hence, the proof of (12) is complete. It is also clear that (12) implies (13).
Next, we observe that lemma 6 and (21) remain true under the alternative. Now, by some
lengthy calculations, it can be verified that \( L \) given in (22) is positive definite so that (14)
follows from (13).

Acknowledgements

Research partially supported by NSF grants DMS 0905400 at the University of Utah, DMS-
0804165 and DMS-0931948 at Colorado State University and DFG grant STE 306/22-1 at
the University of Cologne.

References

Abadir, K. M. & Magnus, J. R. (2005). Matrix algebra. Cambridge University Press, New York.
Benko, M., Härdle, W. & Kneip, A. (2009). Common functional principal components. Ann. Statist. 37,
1–34.
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), 49–53. Physica-Verlag,
Heidelberg.
Bosq, D. (2000). Linear processes in function spaces. Springer, New York.
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. Multivariate Anal. 12, 136–154.
Ferraty, F. & Romain, Y. eds (2011). The Oxford handbook of functional data analysis. Oxford University
Press, Oxford.
Ferraty, F. & Vieu, P. (2006). Nonparametric functional data analysis: theory and practice. Springer, New
York.
Gabrys, R., Horváth, L. & Kokoszka, P. (2010). Tests for error correlation in the functional linear model.
J. Amer. Statist. Assoc. 105, 1113–1125.
Gaines, G., Kaphle, K. & Ruymgaart, F. (2011). Application of a delta-method for random operators
to testing equality of two covariance operators. Math. Meth. Statist. 20, 232–245.
Gervini, D. (2008). Robust functional estimation using the spatial median and spherical principal com-
ponents. Biometrika 95, 587–600.
Gohberg, I., Goldberg, S. & Kaashoek, M. A. (1990). Classes of linear operators. Operator theory:
Advances and applications, 49. Birkhäuser, Basel.
Horváth, L. & Kokoszka, P. (2012). Inference for functional data with applications. Springer Series in
Statistics. Springer, New York (in press).
Horváth, L., Kokoszka, P. & Reeder, R. (2012). Estimation of the mean of functional time series and a
two sample problem. J. Roy. Statist. Soc., Ser. B (in press).
Horváth, L., Kokoszka, P. & Reimherr, M. (2009). Two sample inference in functional linear models.
Canad. J. Statist. 37, 571–591.
Müller, H. G. & Stadtmüller, U. (2005). Generalized functional linear models. Ann. Statist. 33, 774–805.
Panaretos, V. M., Kraus, D. & Maddocks, J. H. (2010). Second-order comparison of Gaussian random
functions and the geometry of DNA minicircles. J. Amer. Statist. Assoc. 105, 670–682.
Panaretos, V. M., Kraus, D. & Maddocks, J. H. (2011). Second-order inference for functional data with
application to DNA minicircles. In Recent advances in functional data analysis and related topics (ed.
F. Ferraty), 245–250. Physica-Verlag, Heidelberg.
Ramsay, J. O. & Silverman, B. W. (2005). Functional data analysis. Springer, New York.
Ramsay, J., Hooker, G. & Graves, S. (2009). Functional data analysis with R and MATLAB. Springer,
New York.
Reiss, P. T. & Ogden, R. T. (2007). Functional principal component regression and functional partial
least squares. J. Amer. Statist. Assoc. 102, 984–996.
Yao, F. & Müller, H. G. (2010). Functional quadratic regression. Biometrika 97, 49–64.

Received February 2011, in final form February 2012

Josef G. Steinebach, Mathematical Institute, University of Cologne, Weyertal 86-90, 50931 Köln,
Germany.
E-mail: jost@math.uni-koeln.de

© 2012 Board of the Foundation of the Scandinavian Journal of Statistics.