# Chapter 38
# Second-Order Inference for Functional Data with Application to DNA Minicircles

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

**Abstract** The problem of comparison of second-order (covariance) properties of two samples of random curves is considered. The work is motivated by the study of the mechanical properties of short strands of DNA. Our test is based on the common empirical Karhunen–Loève expansion and truncated approximation of the Hilbert–Schmidt distance of the empirical covariance operators.

## 38.1 Introduction

The development of the statistical methods described here was motivated by a dataset consisting of reconstructed three-dimensional electron microscope images of loops (called minicircles) obtained from short strands of DNA (Amzallag, Vaillant, Jacob, Unser, Bednar, Kahn, Dubochet, Stasiak and Maddocks, 2006). There are two types (called TATA and CAP) of DNA minicircles with identical base-pair sequences, except for short susubsequence where they differ. The main question is whether this difference affects the geometry of the minicircle.

Mathematically, DNA minicircles are closed curves in $\mathbb{R}^3$. Figure 38.1 shows projections of these curves on the planes given by the axes of the coordinate system. In Figure 38.2 coordinates on the axes are plotted against the arc length of the curve. This plot suggests that the data could be analysed by means of functional data analysis.

Victor M. Panaretos
Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, e-mail: victor.panaretos@epfl.ch

David Kraus
Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, e-mail: david.kraus@epfl.ch

John H. Maddocks
Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, e-mail: john.maddocks@epfl.ch

F. Ferraty (ed.), *Recent Advances in Functional Data Analysis and Related Topics*,
DOI 10.1007/978-3-7908-2736-1_38, 

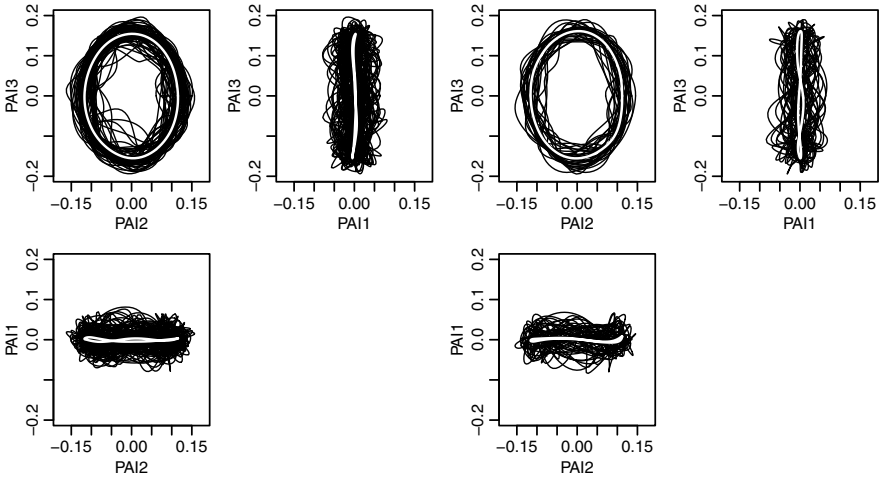


Fig. 38.1: Projections of DNA minicircles on the planes given by the principal axes of inertia (three panels on the left side: TATA curves, right: CAP curves). Mean curves are plotted in white.

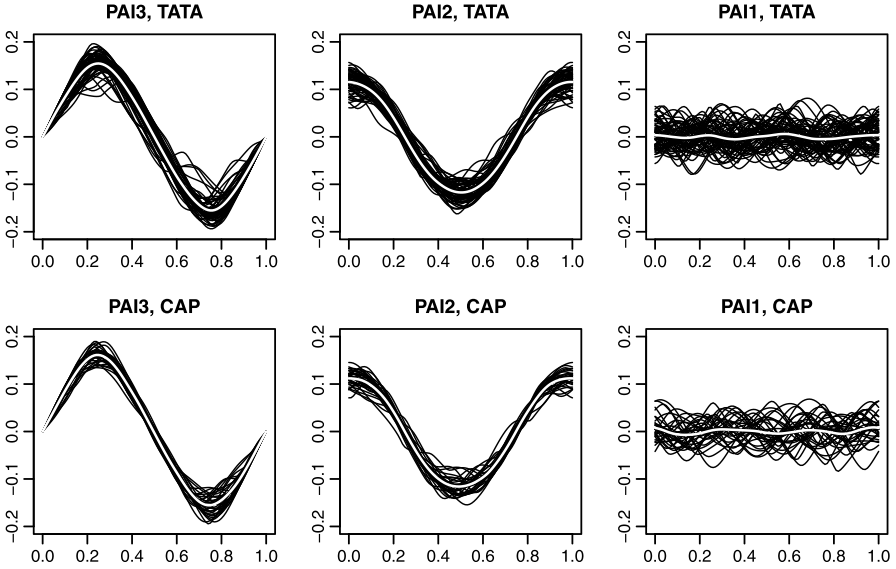


Fig. 38.2: Coordinates of DNA minicircles on the principal axes of inertia. Mean curves are plotted in white.

Plots of estimated mean functions do not suggest any difference between the two types of curves. We tested the hypothesis of equal mean functions and the results were insignificant. Therefore we focused on second-order properties and developed methods for comparing covariance operators.

In this extended abstract we sketch the main idea of the testing procedure (Section 2) and summarise results of the analysis of DNA minicircles (Section 3). Details of the statistical methods and data application mentioned here can be found in Panaretos, Kraus and Maddocks (2010).

## 38.2 Test

Let $X_1,\dots,X_{n_1}$ and $Y_1,\dots,Y_{n_2}$ be two independent samples of stochastic processes with paths in $L^2[0,1]$ with mean functions $\mu_X,\mu_Y$ and covariance operators $\mathscr{R}_X,\mathscr{R}_Y$. The aim is to test the null hypothesis $\mathscr{R}_X=\mathscr{R}_Y$ against the general alternative $\mathscr{R}_X\neq\mathscr{R}_Y$.

The problem of comparing covariance operators of functional data has received relatively little attention in the literature. Related but different second-order problems were studied by Benko, Härdle and Kneip (2009) and Horváth, Hušková and Kokoszka (2010).

Our test is based on the comparison of the empirical covariance operators

$$\hat{\mathscr{R}}_X=\frac{n_1}{n_1+n_2}\sum_{i=1}^{n_1}(X_i-\bar{X})\otimes(X_i-\bar{X}),\qquad \hat{\mathscr{R}}_Y=\frac{n_1}{n_1+n_2}\sum_{i=1}^{n_2}(Y_i-\bar{Y})\otimes(Y_i-\bar{Y}).$$

The test will reject the null hypothesis when the operator $\mathscr{D}=\hat{\mathscr{R}}_X-\hat{\mathscr{R}}_Y$ is significantly far from the zero operator.

The distance of $\mathscr{D}$ from zero can be measured by the squared Hilbert–Schmidt norm

$$\|\mathscr{D}\|^2=\sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\langle\varphi_j,\mathscr{D}\varphi_k\rangle^2$$

where $\{\varphi_j, j=1,2,\dots\}$ is any orthonormal basis of the sample Hilbert space $L^2[0,1]$. This random variable does not have a tractable asymptotic distribution. Therefore we perform dimension reduction and study the infinite-dimensional object $\mathscr{D}$ on a finite-dimensional subspace. Let $\Phi$ be the $K$-dimensional linear subspace generated by an orthonormal basis $\{\varphi_1,\dots,\varphi_K\}$ (where $K$ is a finite number small than or equal to the rank of the covariance operator). Instead of measuring the difference of $\mathscr{D}$ from zero on the whole Hilbert space $L^2[0,1]$, we restrict our attention to $\Phi$. More precisely, instead of $\mathscr{D}$ we use the operator $\pi_\Phi\mathscr{D}\pi_\Phi$ where $\pi_\Phi=\sum_{k=1}^{K}\varphi_k\otimes\varphi_k$ is the projection operator on $\Phi$. The square of its Hilbert–Schmidt norm equals

$$\|\pi_\Phi \mathscr{D} \pi_\Phi\|^2 = \sum_{j=1}^{K}\sum_{k=1}^{K} \langle \varphi_j, \mathscr{D}\varphi_k\rangle^2.$$

In light of the Karhunen–Loève expansion and Mercer's theorem, it is natural to choose the functions $\varphi_k$ as the first $K$ eigenfunctions $\hat{\varphi}_k$ of the pooled sample covariance estimator $\hat{\mathscr{R}} = \frac{n_1}{n_1+n_2}\hat{\mathscr{R}}_X + \frac{n_2}{n_1+n_2}\hat{\mathscr{R}}_Y$. (Note that one cannot perform eigendecomposition of each covariance operator separately because a common basis is needed.)

The terms $S_{jk} = \langle \varphi_j, \mathscr{D}\varphi_k\rangle$ can be seen as differences of the empirical covariances of the Fourier coefficients of the observations with respect to $\varphi_1, \ldots, \varphi_K$. That is, for $\beta_{ik}^X = \langle X_i, \varphi_k\rangle$, $\beta_{ik}^Y = \langle Y_i, \varphi_k\rangle$ one can see that $S_{jk} = \hat{\lambda}_{jk}^X - \hat{\lambda}_{jk}^Y$ where

$$\hat{\lambda}_{jk}^X = \frac{1}{n_1}\sum_{i=1}^{n_1}(\beta_{ij}^X - \bar{\beta}_j^X)(\beta_{ik}^X - \bar{\beta}_k^X), \qquad \hat{\lambda}_{jk}^Y = \frac{1}{n_2}\sum_{i=1}^{n_2}(\beta_{ij}^Y - \bar{\beta}_j^Y)(\beta_{ik}^Y - \bar{\beta}_k^Y).$$

The variable $\|\pi_\Phi \mathscr{D} \pi_\Phi\|^2$ thus equals the squared Frobenius norm of the difference of the empirical covariance matrices of the Fourier scores.

Instead of simply summing the squares of $S_{jk}$, one combines the $K(K+1)/2$ different terms $S_{jk}$, $1 \le j \le k \le K$ in a quadratic form reflecting their covariance structure as follows. Under certain assumptions it can be shown using the Hilbert space Central Limit Theorem that under the null hypothesis the test operator

$$\frac{n_1^{1/2} n_2^{1/2}}{(n_1+n_2)^{1/2}}\mathscr{D}$$

is asymptotically distributed as a zero-mean Gaussian random linear operator on $L^2[0,1]$. Consequently, in view of the consistency of empirical eigenfunctions the vector with components $S_{jk}$, $1 \le j \le k \le K$ converges to a mean zero Gaussian vector whose covariance matrix can be consistently estimated by the empirical covariance matrix, say $W$, of the summands in $S_{jk}$. Then the quadratic test statistic follows the form

$$\frac{n_1 n_2}{n_1+n_2} S^{\mathsf{T}} W S.$$

Its asymptotic distribution under the null is chi-square with $K(K+1)/2$ degrees of freedom. The test rejects $H_0$ when the value of the statistic is significantly large.

In the case of Gaussian data the limiting covariance structure of $S$ simplifies. It turns out that the components $S_{jk}$ are asymptotically independent and their limiting variances can be expressed in terms of the eigenvalues of $\mathscr{R}_1 = \mathscr{R}_2$. This leads to the statistic

$$T = \frac{n_1 n_2}{n_1+n_2}\sum_{j=1}^{K}\sum_{k=1}^{K}\frac{(\hat{\lambda}_{jk}^X - \hat{\lambda}_{jk}^Y)^2}{2(\frac{n_1}{n}\hat{\lambda}_{jj}^X + \frac{n_2}{n}\hat{\lambda}_{jj}^Y)(\frac{n_1}{n}\hat{\lambda}_{kk}^X + \frac{n_2}{n}\hat{\lambda}_{kk}^Y)}$$

with asymptotic $\chi^2$ distribution with $K(K+1)/2$ degrees of freedom. When one a priori expects the eigenfunctions in the two samples to be equal, the test can be based

only on the diagonal ($j = k$) terms in the sum above (comparing only variances, not covariances of the scores). Such a statistic is asymptotically $\chi^2_K$-distributed. Modifications of the test statistics can be obtained by variance stabilising transformations of the summands.

The truncation level $K$ can be selected with the help of scree plots and cumulative variance plots. We have also proposed an automatic procedure based on a penalised fit criterion.

## 38.3 Application to DNA minicircles

The original original $(x, y, z)$-coordinates of the curves were obtained from electron microscope images of a frozen liquid containing the minicircles. Therefore the original curves are randomly rotated and shifted, thus not directly comparable. So it is necessary to align them. We cannot apply landmark alignment methods because there are no landmarks (the sequence of DNA base-pairs is not observed). Warping methods are not appropriate as they could modify the second-order properties. Instead, after centering (setting the center of mass to 0) and scaling to unit length, we align each curve separately by rotating it in a way given by the moment of inertia tensor.

The moment of inertia tensor is defined as

$$J(u) = \int_{\mathbb{R}^3} \|(I - uu^\mathsf{T})x\|^2 \mu(dx)$$

where $u$ is a unit vector in $\mathbb{R}^3$ and $\mu$ is the uniform distribution of mass on the curve. By integrating the squared distance of the points on the curve from the axis, the tensor measures how difficult it is to rotate the curve around the axis given by $u$. The first eigenvector (corresponding to the largest eigenvalue) determines the first principal axis of inertia (PAI1) around which the curve is most difficult to rotate. The projection on the plane orthogonal to PAI1 is most spread. Then PAI2 given by the second eigenvector is the axis orthogonal to PAI1 around which the projection of the curve on the first principal plane is most difficult to rotate. Within this plane, the projection on the axis PAI3 orthogonal to PAI2 is most spread.

For each curve we computed the principal axes of inertia and rotated the curve so that its principal axes agree with the $(x, y, z)$-axes. This procedure is similar to the balancing of a tyre. Figure 38.1 shows the rotated minicircles. These closed curves have no starting point and no orientation. As the starting point of each curve we chose the point where the projection of the curve on the first principal plane intersects the positive horizontal semi-axis; we chose the counter-clockwise orientation. As the 'time' argument of each functional observation we use the arc length of the curve from the starting point. The resulting functional data set is plotted in Figure 38.2.

The test comparing the covariance operators suggests significant differences between the samples. For example, when applied to the projections on the first princi-

pal plane (PAI2,3) with $K = 7$ (selected by the automatic procedure), the $p$-value is 0.023.

## References

1. Amzallag, A., Vaillant, C., Jacob, M., Unser, M., Bednar, J., Kahn, J. D., Dubochet, J., Stasiak, A., Maddocks, J. H.: 3D reconstruction and comparison of shapes of DNA minicircles observed by cryo-electron microscopy. Nucleic Acids Research **34**, e125 (2006)
2. Benko, M., Härdle W., Kneip, A.: Common functional principal components. Ann. Stat. **37**, 1–34 (2009)
3. Horváth, L., Hušková, M., Kokoszka, P.: J. Multivariate Anal. **101**, 352–367.
4. Panaretos, V. M., Kraus, D., Maddocks, J. H.: Second-order comparison of Gaussian random functions and the geometry of DNA minicircles. J. Am. Stat. Assoc. **105**, 670–682 (2010)