Second-Order Inference for Gaussian Random Curves
With Application to DNA Minicircles

Victor Panaretos  David Kraus  John Maddocks

Ecole Polytechnique Fédérale de Lausanne
Are they different?
1. DNA minicircles

2. Functional Data Analysis background

3. Testing procedures

4. Analysis of DNA minicircles

5. Summary
Geometry of DNA molecules, base-pairs sequences

- Goal: study of mechanical properties of DNA molecules (via their geometry)
- Question: influence of base pair sequences

\texttt{c11T15} (158 bp, J=95 nM):

\begin{verbatim}
GATGAATTCACGGATCCCGTTTTTTGGCCCGTTTTTTTCGCCCTTTTTTGC
GACCTAGGGTCTAATGAGTGAGCTAACTCACAATTAATGGTGCCTGCGCC
ATGGAAATC
\end{verbatim}

\texttt{t11T15} (158 bp, J\sim3500 nM):

\begin{verbatim}
GATGAATTCACGGATCCCGTTTTTTGGCCCGTTTTTTTCGCCCTTTTTTGC
GACCTAGGGTGCTAATGAGTGCCCTTTTATAGCTTAAACGCCTGTGCGCC
ATGGAAATC
\end{verbatim}

158 base-pairs, 18 bp fragment different (TATA vs CAP)
Electron microscope images

50 nm layer of ice at $-170^\circ$C, images tilted $\pm 15^\circ$

Minicircles diam $\sim 17$ nm
- Each curve 200 $xyz$-coordinates
- Curves not directly comparable
- Adjustment
  - Centering (center of mass $= 0$)
  - Scaling (length $= 1$)
- Not sufficient, further alignment necessary
Registration of functional data

- Standard alignment methods
  1. Landmark alignment
  2. Warping
- Standard methods cannot be used
  - Landmark alignment not applicable: no landmarks available
  - Warping inappropriate: don’t want to change the shape, need a rigid method
  - Curves have no beginning/end, no orientation
- Rotate each curve to make them as close as possible
  - Global optimization over $n = 99$ orthogonal transformations would be difficult
- Instead, rotate each curve separately
Moments of inertia tensor

- Consider an object in $\mathbb{R}^3$ with distribution of mass $\mu$
- For DNA minicircles, $\mu$ is the uniform measure supported on the curve
- Consider an axis given by a unit vector $u \in \mathbb{R}^3$ ($\|u\| = 1$)
- Moment of inertia tensor defined as

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

(integrated squared distance from the axis given by $u$)

- Interpretation: $J(u)$ measures how difficult it is to rotate the object around the axis $u$
- In matrix form

$$J(u) = u^T Ju, \quad \text{where } J = \int_{\mathbb{R}^3} (x^T x I - xx^T) \mu(dx)$$
Principal axes of inertia

- $\mathcal{J}(u) = u^T Ju$ is positive semidefinite, hence it possesses nonnegative eigenvalues and orthonormal eigenvectors.

- The first eigenvector $w_1$ determines the axis around which the curve is most difficult to rotate ($\mathcal{J}(u) = u^T Ju$ is maximized at $u = w_1$).

- The projection of the curve on the plane orthogonal to $w_1$ is most spread.

- The second eigenvector $w_2$ determines the axis within the first principal plane around which the projected curve is most difficult to rotate.

- Within the first principal plane, the projection on the line orthogonal to $w_2$ is most spread.

- The axes given by $w_1, w_2, w_3$ are called principal axes of inertia (PAI1, PAI2, PAI3).

- PAI3 carries the most spatial information, PAI1 contains the smallest amount of information.
Moments of inertia alignment

- Each curve aligned separately (no averaging over the sample)
- For each curve, the principal axes of inertia are determined and the curve is rotated so that the PAI’s agree with the axes of the coordinate system (i.e., the curves are projected on PAI’s)
- The procedure is similar to the balancing of a tyre (adjusting the distribution of mass of a wheel such that its PAI is aligned with the axle)
Aligned DNA minicircles

TATA

CAP

Panaretos, Kraus, Maddocks (EPFL)
From curves to functions

- Curves have no starting point
  - The intersection of the projection on the first principal plane and the horizontal (PAI2) positive semi-axis

- Curves have no orientation
  - Counterclockwise in the first principal plane

- No correspondence between points on the curves
  - Parametrization by arc length
PAI coordinates of aligned DNA minicircles

Panaretos, Kraus, Maddocks (EPFL)
1. DNA minicircles

2. Functional Data Analysis background

3. Testing procedures

4. Analysis of DNA minicircles

5. Summary
Each minicircle curve is modelled as the realisation of a stochastic process indexed by $[0, 1]$,

$$X = \{X(t), t \in [0, 1]\}$$

taking values in $\mathbb{R}^3$

$X$ is seen as a random element of the Hilbert space $L^2[0, 1]$ of coordinate-wise square-integrable $\mathbb{R}^3$-valued functions with the inner product

$$\langle f, g \rangle = \int_0^1 \langle f(t), g(t) \rangle dt$$

Wlog assume that the mean function

$$\mu(t) = \mathbb{E}X(t)$$

is zero
Covariance operator

• Denote the covariance function (kernel)

\[ R(s, t) = \text{cov}(X(s), X(t)) = \mathbb{E}(X(s)X(t)^T) \]

• The covariance operator is defined as

\[ \mathcal{R} : L^2[0, 1] \to L^2[0, 1] \]

\[ \mathcal{R}(f) = \text{cov}(\langle X, f \rangle X) = \int_0^1 R(\cdot, t)f(t)dt, \]

• Equivalently

\[ \mathcal{R} = \mathbb{E}(X \otimes X) \]

• The tensor product of \( a, b \in L^2[0, 1] \) is defined as the operator

\[ (a \otimes b) : L^2[0, 1] \to L^2[0, 1], \quad (a \otimes b)(f) = \int_0^1 a(\cdot)\langle b(t), f(t) \rangle dt \]

• Multivariate analog: \( a \otimes b = ab^T \) for \( a, b \in \mathbb{R}^p \)
The covariance kernel admits the representation

\[ R(s, t) = \sum_{k=1}^{\infty} \lambda_k \varphi_k(s) \varphi_k(t)^T, \]

where \( \lambda_k \geq 0 \) are nonincreasing eigenvalues and \( \varphi_k \) orthonormal eigenfunctions of \( R \), i.e., \( R(\varphi_k) = \lambda_k \varphi_k \).

The process \( X \) can be represented as

\[ X(t) = \sum_{k=1}^{\infty} \langle X, \varphi_k \rangle \varphi_k(t) = \sum_{k=1}^{\infty} \lambda_k^{1/2} \xi_k \varphi_k(t) \]

where the Fourier coefficients \( \xi_k = \lambda_k^{-1/2} \langle X, \varphi_k \rangle \) are uncorrelated random variables with zero mean and unit variance.

If the process is Gaussian, the scores \( \xi_k, k \geq 1 \) are iid standard Gaussian.
The first $K$ eigenelements $\lambda_k, \varphi_k, k = 1, \ldots, K$ provide the optimal rank $K$ approximation in the sense

$$\min_{\varphi_1, \ldots, \varphi_K \text{ orthonormal}} \mathbb{E}\left\| X - \sum_{k=1}^{K} \langle X, \varphi_k \rangle \varphi_k \right\|_2^2$$

$$\min_{\varphi_1, \ldots, \varphi_K \text{ orthonormal, } \lambda_1, \ldots, \lambda_K \geq 0} \mathbb{E}\left\| X \otimes X - \sum_{k=1}^{K} \lambda_k (\varphi_k \otimes \varphi_k) \right\|_{HS}^2$$

$$\max_{\varphi_1, \ldots, \varphi_K \text{ orthonormal}} \sum_{k=1}^{K} \text{var}(\langle X, \varphi_k \rangle)$$
Functional Principal Component Analysis is the empirical version of the Karhunen–Loève decomposition.

The empirical covariance operator is given by:

$$\hat{\mathcal{R}} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}) \otimes (X_i - \bar{X})$$

The functional eigenproblem is:

$$\hat{\mathcal{R}} \hat{\phi}_k = \hat{\lambda}_k \hat{\phi}_k$$

Usually, observations are represented in a basis,

$$X_i(t) = \sum_{j=1}^{\infty} c_{ij} \psi_j(t)$$

If the basis \{\psi_j\} is orthonormal (such as the Fourier basis for periodic data like minicircles), then functional PCA is usual PCA of the coefficient matrix \(C = (c_{ij})\).
Outline

1. DNA minicircles
2. Functional Data Analysis background
3. Testing procedures
4. Analysis of DNA minicircles
5. Summary
Means of PAI coordinates

Panaretos, Kraus, Maddocks (EPFL)
Means on the principal plane

Proj. on Prin. Plane 1, TATA

Proj. on Prin. Plane 1, all

Proj. on Prin. Plane 1, CAP

Panaretos, Kraus, Maddocks (EPFL)
Situation: $X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$ independent samples of Gaussian stochastic processes with means $\mu_X, \mu_Y$ and covariance operators $\mathcal{R}_X, \mathcal{R}_Y$

Mean functions appear to be equal
\[ \rightarrow \] Focus on covariance operators

Hypothesis testing problem

\[ H_0 : \mathcal{R}_X = \mathcal{R}_Y \quad \text{vs} \quad H_1 : \mathcal{R}_X \neq \mathcal{R}_Y \]

Use of a statistic like $\hat{\mathcal{R}}_X^{-1} \hat{\mathcal{R}}_Y$ impossible (noninvertibility)

Instead, use the difference

\[ \hat{\mathcal{R}}_X - \hat{\mathcal{R}}_Y \]

which should be close to the zero operator under the null
Hilbert–Schmidt operator norm

- Need to measure the distance of $\hat{R}_X - \hat{R}_Y$ from a zero operator
  $\hookrightarrow$ Need an operator norm

- The Hilbert–Schmidt operator norm is defined as

\[ \|A\|_{HS}^2 = \sum_i \|A e_i\|_2^2 = \sum_{i,j} \langle e_i, A e_j \rangle^2 \]

(multivariate analog: $\|A\|_F^2 = \sum_{i,j} a_{ij}^2$ for a matrix $A$)

- The distribution of $\|\hat{R}_X - \hat{R}_Y\|_{HS}^2$ intractable

$\hookrightarrow$ Perform dimension reduction, focus on the projected operators
Projection, truncation of the HS norm

- Let $f_1, \ldots, f_K$ be some orthonormal $L^2$ functions
- Let
  
  $$
  \pi_K = \sum_{k=1}^{K} f_k \otimes f_k
  $$

  be the projection operator onto the span of $f_1, \ldots, f_K$
- The test will be based on
  
  $$
  \left\| \hat{R}^K_X - \hat{R}^K_Y \right\|_{\text{HS}}^2
  $$

  where $\hat{R}^K_X = \pi_K \hat{R}_X \pi_K$, $\hat{R}^K_Y = \pi_K \hat{R}_Y \pi_K$
- Need a common basis (the same $\pi_K$ for $X, Y$)
  (a common reference coordinate system)
- We use $\pi_K = \hat{\pi}_K$ projecting on
  
  $\hat{\phi}_{XY}^1, \ldots, \hat{\phi}_{XY}^K$

  (eigenfunctions of the pooled-sample estimator
  
  $\hat{R} = \frac{n_1}{n} \hat{R}_X + \frac{n_2}{n} \hat{R}_Y$)
The statistic is

\[ \| \hat{\mathcal{R}}_X^K - \hat{\mathcal{R}}_Y^K \|_{HS}^2 = \sum_{k=1}^K \sum_{j=1}^K \langle \hat{\phi}^{XY}_k, (\hat{\mathcal{R}}_X - \hat{\mathcal{R}}_Y) \hat{\phi}^{XY}_j \rangle^2 \]

The terms

\[
\hat{\lambda}^{X,XY}_{kj} = \langle \hat{\phi}^{XY}_k, \hat{\mathcal{R}}_X \hat{\phi}^{XY}_j \rangle, \quad \hat{\lambda}^{Y,XY}_{kj} = \langle \hat{\phi}^{XY}_k, \hat{\mathcal{R}}_Y \hat{\phi}^{XY}_j \rangle
\]

are empirical var/cov of the X and Y scores w.r.t. the common basis

Their asymptotic var/cov under \( H_0 \) is \( 2 \delta_{kj} \lambda_k \lambda_j \)

The test statistic based on standardized components is

\[
T = \frac{n_1 n_2}{2n} \sum_{k=1}^K \sum_{j=1}^K \frac{(\hat{\lambda}^{X,XY}_{kj} - \hat{\lambda}^{Y,XY}_{kj})^2}{\hat{\lambda}_k \hat{\lambda}_j}
\]
Under $H_0$ and Gaussian assumption,

$$T \xrightarrow{D} \chi^2_{K(K+1)/2}$$

Sketch of the proof

- Using consistency of $\hat{\varphi}^{XY}_k$, replace $\hat{\pi}_K$ by $\pi_K$
- By the Hilbert space CLT,

$$\hat{R}_X = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathcal{X}_i \quad \text{with} \quad \mathcal{X}_i = X_i \otimes X_i$$

is asymptotically a Gaussian random operator (random element in the space of operators with Gaussian fdd’s)

Investigation of the covariance operator of the limit (an operator on operators on $L^2$) yields that the components of $T$ are asymptotically independent Gaussian
Modifications of the test statistic

- **Diagonal statistic**

  \[ T_1 = \frac{n_1 n_2}{2n} \sum_{k=1}^{K} \left( \frac{\hat{\lambda}_{kk}^{X,Y} - \hat{\lambda}_{kk}^{Y,X}}{\hat{\lambda}_{k}^2} \right)^2 \]

  (compares only eigenvalues, might be good when eigenfunctions equal)

- **Variance-stabilizing transformations**
  - log of the diagonal (variance) terms
  - Fisher’s z-transformation of the off-diagonal (covariance) terms
Selection of $K$

1. Scree plots, cumulative explained proportion of variance, . . .
2. Minimization of the penalized fit criterion

$$PFC(K) = GOF_X(K) + GOF_Y(K) + \frac{n_1}{n} \text{PEN}_X(K) + \frac{n_2}{n} \text{PEN}_Y(K)$$

(no formal result on the post-selection test)
Simulations: design

- Simulated processes are combinations of Fourier basis functions with independent Gaussian coefficients.
- Mimicking the ‘elbow effect’: 3 or 4 dominating components and several components with smaller variance.
- $n_1 = n_2 = 50$
- Nominal level 5%
Scenario A

- Equal covariance operators

<table>
<thead>
<tr>
<th>Test</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>$K^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>off-diag</td>
<td>0.051</td>
<td>0.056</td>
<td>0.057</td>
<td>0.056</td>
<td>0.059</td>
</tr>
<tr>
<td>diag</td>
<td>0.051</td>
<td>0.054</td>
<td>0.056</td>
<td>0.061</td>
<td>0.061</td>
</tr>
</tbody>
</table>

- The tests maintain the nominal level when $K \leq \text{rank}$
- This is true also for $K^*$. The selection criterion aims at estimating the effective complexity of the distributions, it does not optimize the power, does not reflect validity or invalidity of $H_0$. 

Panaretos, Kraus, Maddocks (EPFL)

Second-Order Inference for Gaussian Random Curves
Scenario B

- The same sequence of eigenfunctions (in the same order)
- The first eigenvalues differ

<table>
<thead>
<tr>
<th>Test</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>$K^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>off-diag</td>
<td>0.443</td>
<td>0.315</td>
<td>0.223</td>
<td>0.174</td>
<td>0.175</td>
</tr>
<tr>
<td>diag</td>
<td>0.443</td>
<td>0.350</td>
<td>0.306</td>
<td>0.267</td>
<td>0.267</td>
</tr>
</tbody>
</table>

- The power decreases (difference only in the first component)
- Diagonal better (the same eigenfunctions)
Scenario E

- The same set of eigenfunctions in a different order (permuted)
- The first eigenfunctions equal
- The same eigenvalues

<table>
<thead>
<tr>
<th>Test</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>$K^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>off-diag</td>
<td>0.055</td>
<td>0.267</td>
<td>0.686</td>
<td>0.976</td>
<td>0.975</td>
</tr>
<tr>
<td>diag</td>
<td>0.055</td>
<td>0.250</td>
<td>0.509</td>
<td>0.620</td>
<td>0.617</td>
</tr>
</tbody>
</table>

- No power with $K = 1$ (equal first eigenelements)
- Lower power for the diagonal test (the same eigenvalues)
Scenario F

- Completely different eigenfunctions (sines vs time-shifted sines)
- Equal eigenvalues

<table>
<thead>
<tr>
<th>Test</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>$K^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>off-diag</td>
<td>0.273</td>
<td>0.706</td>
<td>0.916</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>diag</td>
<td>0.273</td>
<td>0.496</td>
<td>0.544</td>
<td>0.594</td>
<td>0.655</td>
</tr>
</tbody>
</table>

- Lower power for the diagonal test
Outliers?

Principal Axis 3, TATA

Principal Axis 2, TATA

Principal Axis 1, TATA

Prin. Plane, TATA

Principal Axis 3, CAP

Principal Axis 2, CAP

Principal Axis 1, CAP

Prin. Plane, CAP

Panaretos, Kraus, Maddocks (EPFL)
The functional spatial median is defined as the solution to

$$\min_{m \in L^2} \sum_{i=1}^{n} \|X_i - m\|_2 \quad \text{or} \quad \sum_{i=1}^{n} \frac{m - X_i}{\|m - X_i\|_2} = 0$$

The solution $\hat{m}$ can be written as the weighted sum

$$\hat{m} = \sum_{i=1}^{n} w_i X_i, \quad w_i \geq 0, \quad \sum_{i=1}^{n} w_i = 1$$

Outliers have small weights $w_i$, large values of $1/w_i$ indicate outliers
Inverse spatial median weights

Inverse median weights, PAI3

Inverse median weights, PAI2

Inverse median weights, PAI1

Inverse median weights, PAI1,2,3

Panaretos, Kraus, Maddocks (EPFL)
Outliers

Panaretos, Kraus, Maddocks (EPFL)
Joint PCA of PAI2,3: coordinates of eigenfunctions

Panaretos, Kraus, Maddocks (EPFL)
Joint PCA of PAI2,3: eigencircles

PCA function 1 (Percentage of variability 26.1 )

PCA function 2 (Percentage of variability 22.4 )

PCA function 3 (Percentage of variability 12.4 )

PCA function 4 (Percentage of variability 11 )

PCA function 1 (Percentage of variability 28.1 )

PCA function 2 (Percentage of variability 20.1 )

PCA function 3 (Percentage of variability 12.5 )

PCA function 4 (Percentage of variability 11.6 )
Joint PCA of PAI2,3: eigenvalues

Lambda for PAI2,3

Var prop for PAI2,3

Cum var prop for PAI2,3

Panaretos, Kraus, Maddocks (EPFL)
Selection of $K$ (joint analysis of PAI2,3)

Scree plot w.r.t. the common (pooled sample) eigenbasis

Plots suggest $K = 6$ or 7

Automatic choice $K = 7$
Test statistic: off-diagonal, transformed, $\chi^2$ approximation

<table>
<thead>
<tr>
<th>$K$</th>
<th>PAI3</th>
<th>PAI2</th>
<th>PAI1</th>
<th>PAI2,3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.252</td>
<td>0.313</td>
<td>0.976</td>
<td>0.167</td>
</tr>
<tr>
<td>2</td>
<td>0.001</td>
<td>0.118</td>
<td>0.823</td>
<td>0.005</td>
</tr>
<tr>
<td>3</td>
<td>0.000</td>
<td>0.087</td>
<td>0.782</td>
<td>0.025</td>
</tr>
<tr>
<td>4</td>
<td>0.001$^S$</td>
<td>0.022</td>
<td>0.886</td>
<td>0.051</td>
</tr>
<tr>
<td>5</td>
<td>0.001$^A$</td>
<td>0.053$^S$</td>
<td>0.555</td>
<td>0.009</td>
</tr>
<tr>
<td>6</td>
<td>0.010</td>
<td>0.087</td>
<td>0.327</td>
<td>0.005$^S$</td>
</tr>
<tr>
<td>7</td>
<td>0.019</td>
<td>0.098$^A$</td>
<td>0.360</td>
<td>0.023$^A$</td>
</tr>
<tr>
<td>8</td>
<td>0.046</td>
<td>0.173</td>
<td>0.148</td>
<td>0.094</td>
</tr>
</tbody>
</table>

(S = Selection based on scree plots, A = Automatic selection)
1. DNA minicircles
2. Functional Data Analysis background
3. Testing procedures
4. Analysis of DNA minicircles
5. Summary
Summary

DNA minicircle data, alignment, . . .
Functional data approach
Tests based on an approximation of the Hilbert–Schmidt distance between empirical covariance operators, asymptotics, simulations, . . .
Minicircle data analysis (outlier detection, order selection, testing, . . .)

Outlook

Robustification
Normality testing
. . .