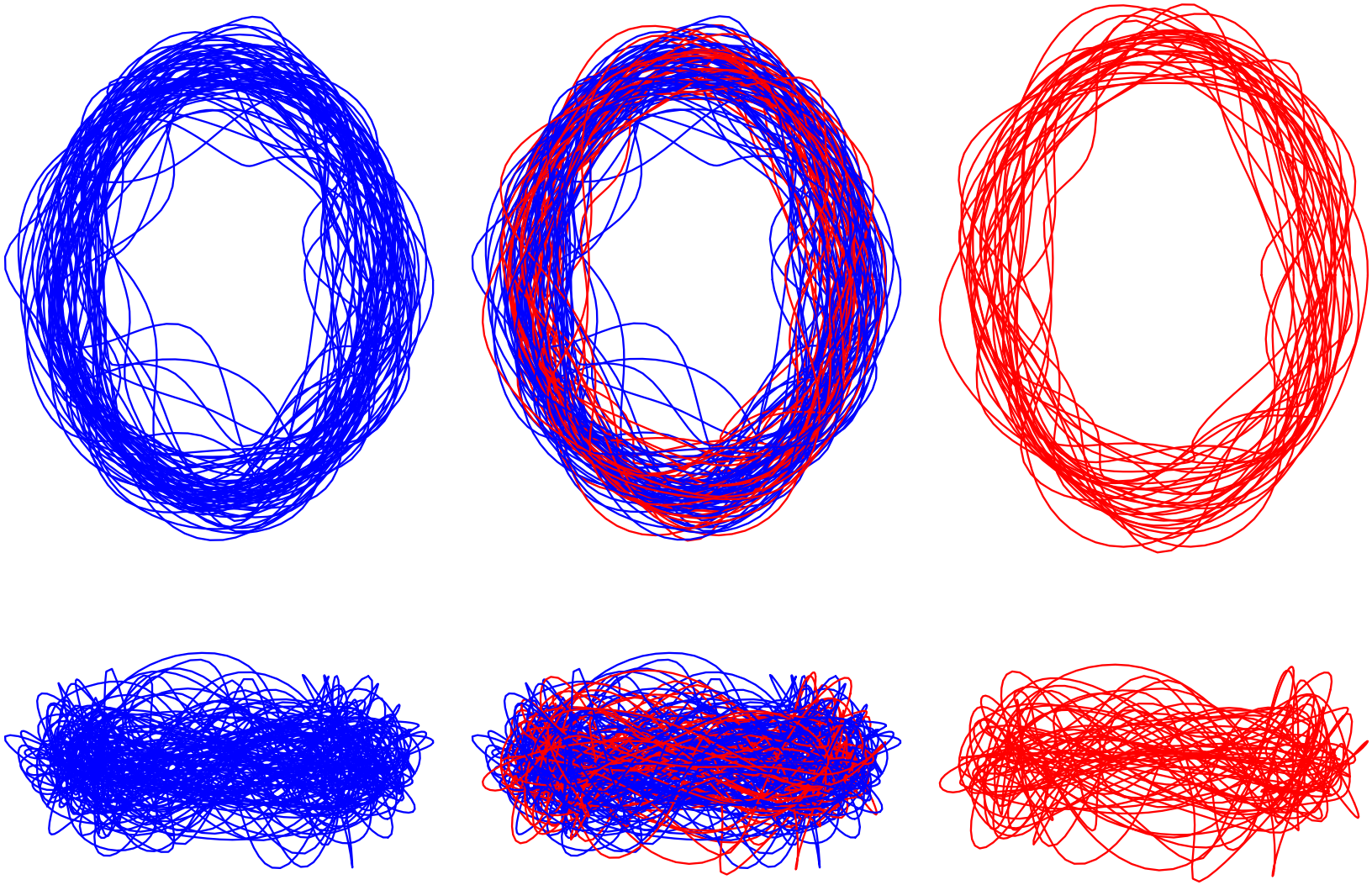
Second-Order Inference for Gaussian Random Curves

With Application to DNA Minicircles

Victor Panaretos David Kraus John Maddocks

Ecole Polytechnique Fédérale de Lausanne

Are they different?



Outline

Outline

Geometry of DNA molecules, base-pairs sequences

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

c11T15 (158 bp, J=95 nM):

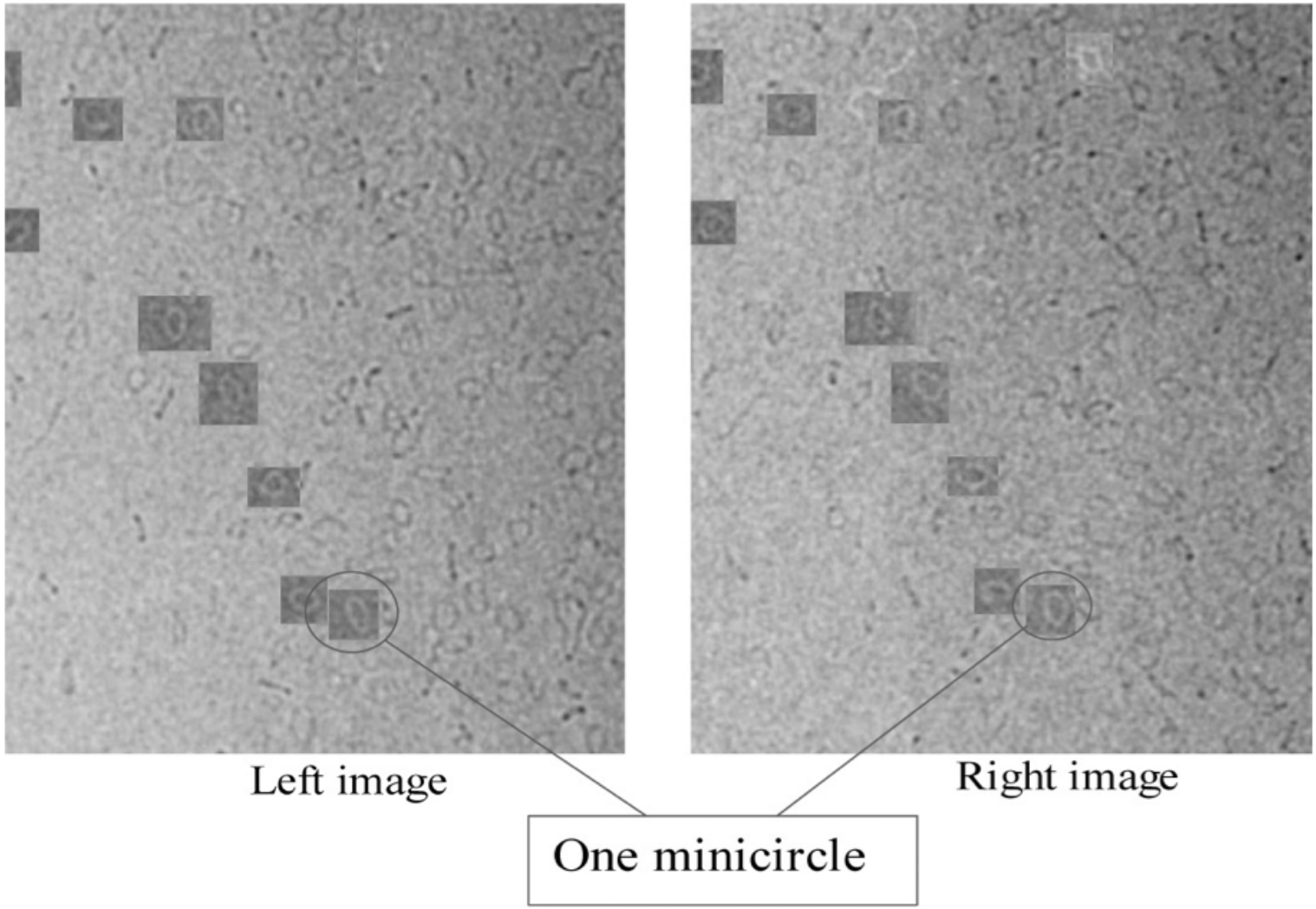
GATGAATTCACGGATCCGGTTTTTTGCCCGTTTTTTGCCGTTTTTTGCCC
GTTTTTTGCCGTTTTTTGCCCGTTTTTTCCGGATCCGTACAGGAATTCTA
GACCTAGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCC
ATGGAATC

t11T15 (158 bp, J~3500 nM):

GATGAATTCACGGATCCGGTTTTTTGCCCGTTTTTTGCCGTTTTTTGCCC
GTTTTTTGCCGTTTTTTGCCCGTTTTTTCCGGATCCGTACAGGAATTCTA
GACCTAGGGTGCCTAATGAGTGCCCTTTTATAGCTTAAACGCGTTGCGCC
ATGGAATC

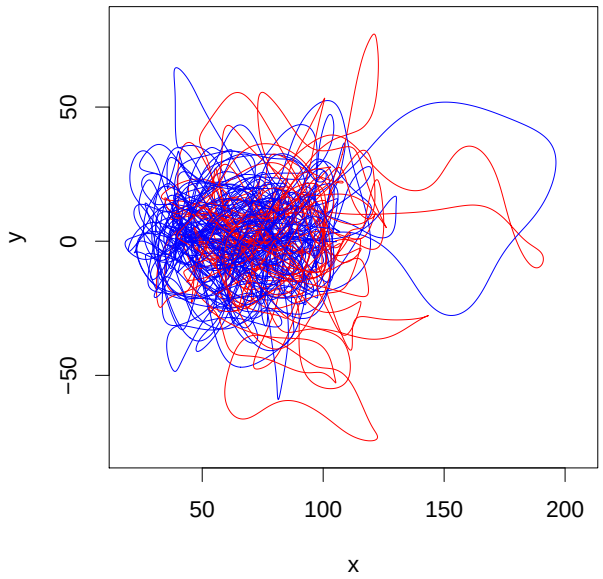
158 base-pairs, 18 bp fragment different (TATA vs CAP)

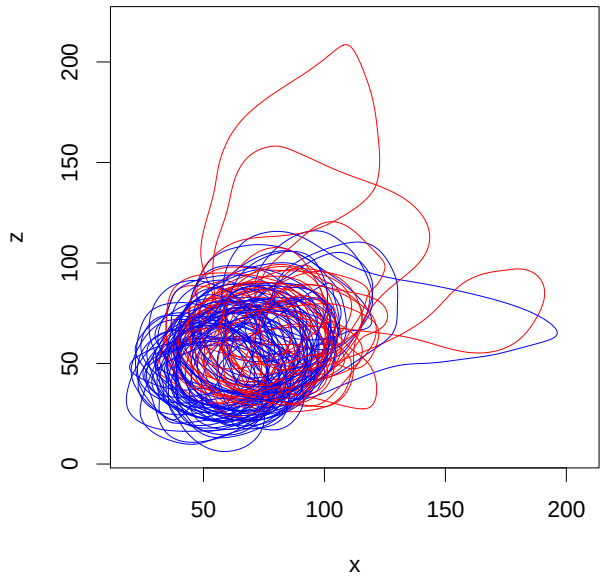
Electron microscope images

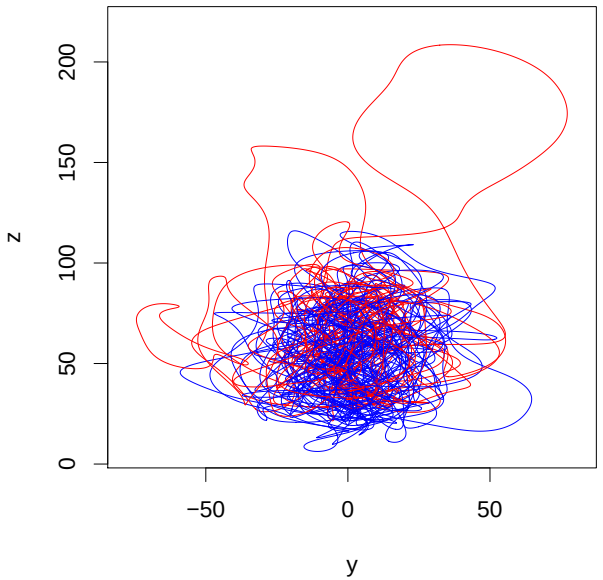


50 nm layer of ice at $-170\,^{\circ}$C, images tilted $\pm 15\,^{\circ}$
Minicircles diam ~ 17 nm

Original data







- 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 μ
- For DNA minicircles, μ 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

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

(integrated squared distance from the axis given by u)

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

$$\mathscr{J}(u) = u^\mathsf{T} J u, \quad \text{where } J = \int_{\mathbb{R}^3} (x^\mathsf{T} x I - x x^\mathsf{T})\mu(dx)$$

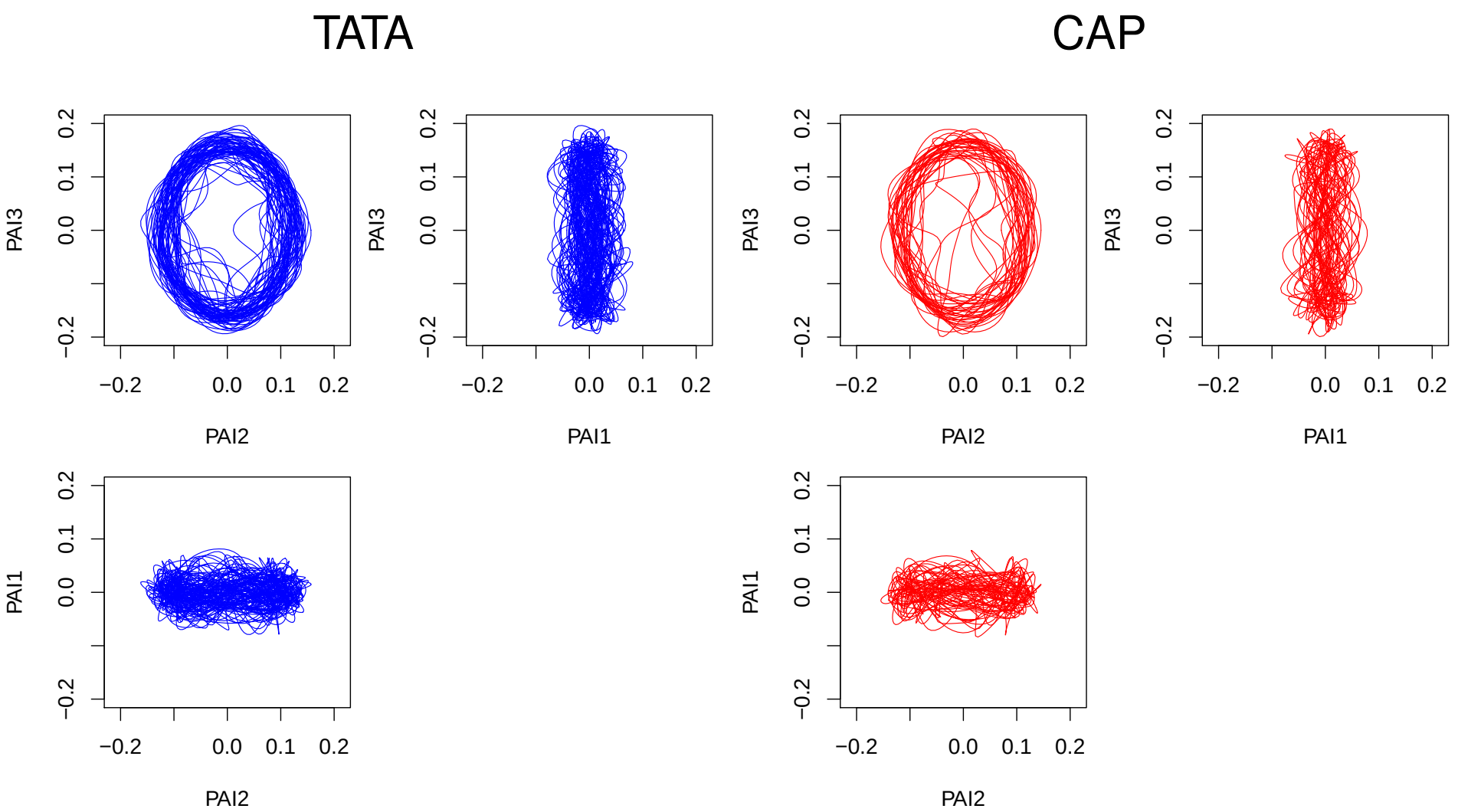
Principal axes of inertia

- $\mathscr{J}(u) = u^{\mathsf{T}} J u$ 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
  ($\mathscr{J}(u) = u^{\mathsf{T}} J u$ is maximized at $u = w_1$)
  $\hookrightarrow$ 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
  $\hookrightarrow$ 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

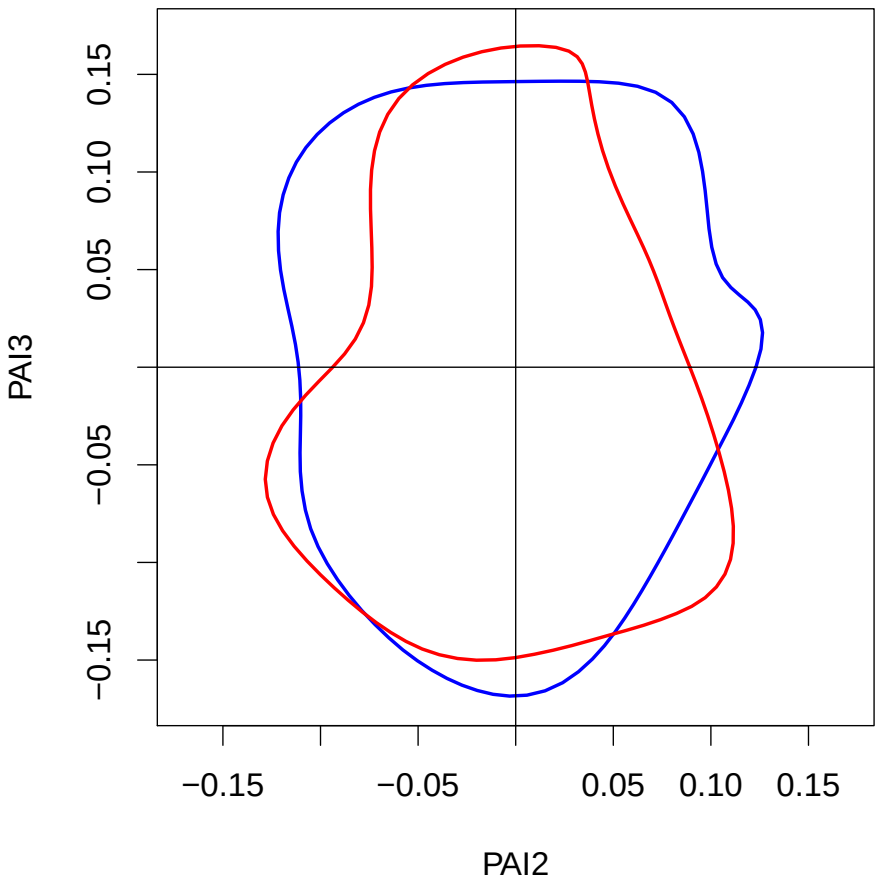
- Each curve aligned separately (no averaging over the sample)
- For each curve, the principal axes of inertia are determined and the curve is rotated so that the PAI's agree with the axes of the coordinate system (i.e., the curves are projected on PAI's)
- The procedure is similar to the balancing of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

Aligned DNA minicircles

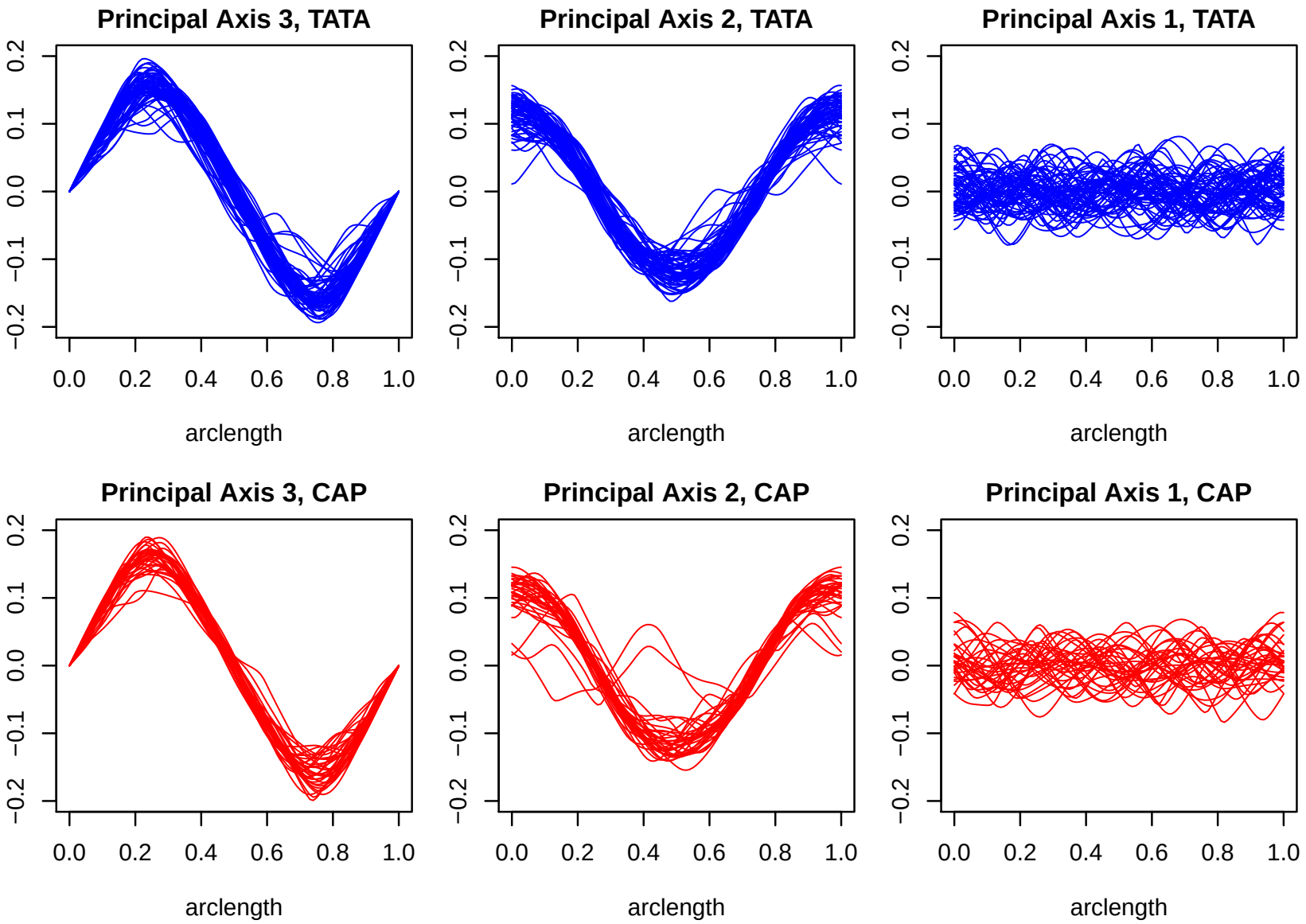


From curves to functions

- 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



Outline

Functional data

- 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

$$\mu(t) = \mathrm{E}\, X(t)$$

  is zero

Covariance operator

- Denote the covariance function (kernel)

$$R(s,t) = \mathrm{cov}(X(s), X(t)) = \mathrm{E}(X(s)X(t)^{\mathsf{T}})$$

- The covariance operator is defined as

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

$$\mathscr{R}(f) = \mathrm{cov}(\langle X, f\rangle X) = \int_0^1 R(\cdot, t) f(t) dt,$$

- Equivalently

$$\mathscr{R} = \mathrm{E}(X \otimes X)$$

- 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: $a \otimes b = ab^{\mathsf{T}}$ for $a, b \in \mathbb{R}^p$

Karhunen–Loève decomposition

- The covariance kernel admits the representation

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

  where $\lambda_k \geq 0$ are nonincreasing eigenvalues and φ_k orthonormal eigenfunctions of $\mathscr{R}$, i.e., $\mathscr{R}(\varphi_k) = \lambda_k \varphi_k$
- 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 ξ_k, $k \geq 1$ are iid standard Gaussian

Truncated KL decomposition, dimension reduction

- The first K eigenelements $\lambda_k, \varphi_k, k = 1, \ldots, K$ provide the optimal rank K approximation in the sense

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

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

$$\max_{\substack{\varphi_1,\ldots,\varphi_K \\ \text{orthonormal}}} \sum_{k=1}^{K} \mathrm{var}(\langle X, \varphi_k \rangle)$$

Functional Principal Component Analysis

- Functional Principal Component Analysis is the empirical version of the Karhunen–Loève decomposition
- Empirical covariance operator

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

- Functional eigenproblem

$$\hat{\mathscr{R}}\hat{\varphi}_k = \hat{\lambda}_k\hat{\varphi}_k$$

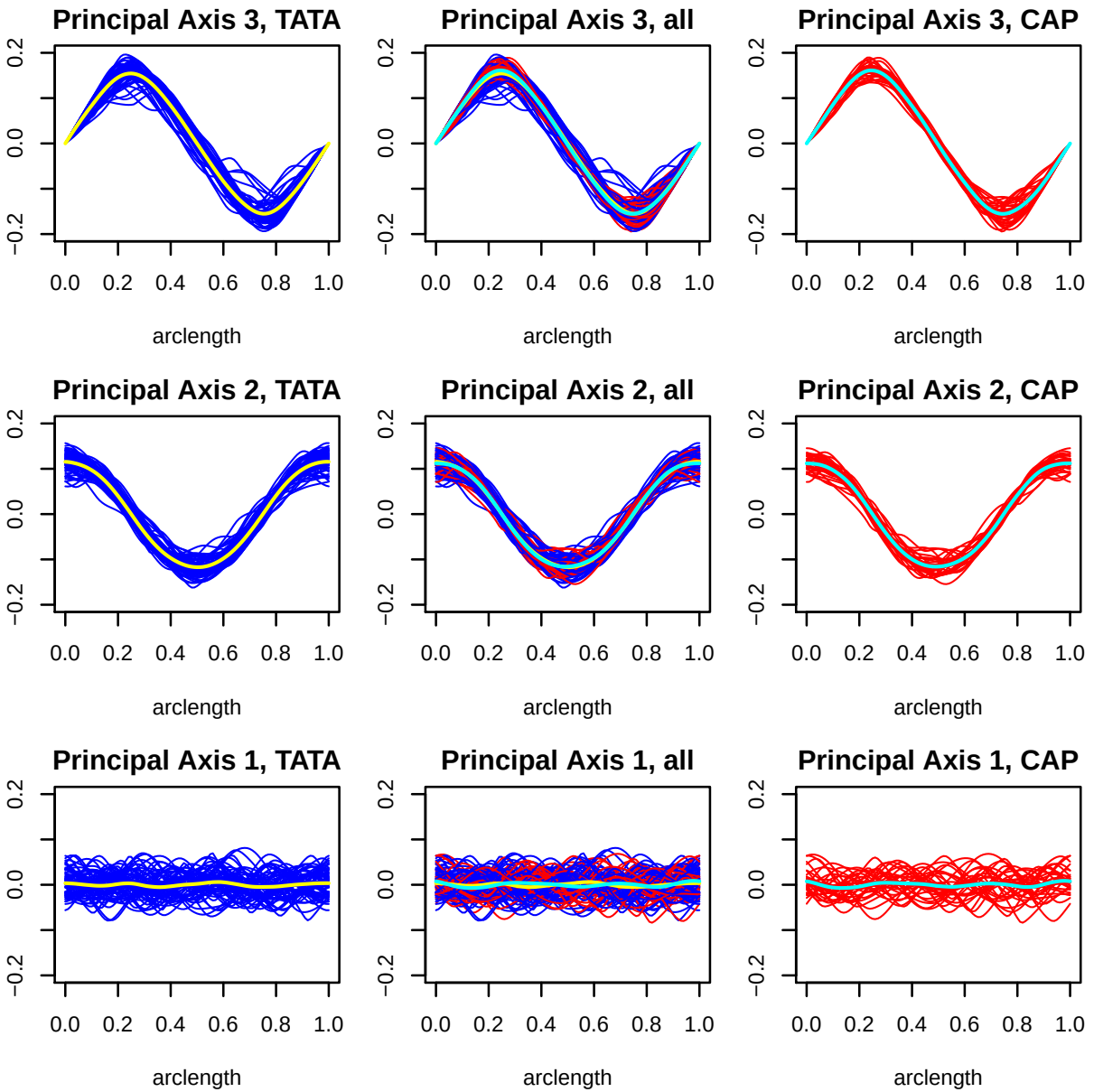
- Usually, observations are represented in a basis,

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

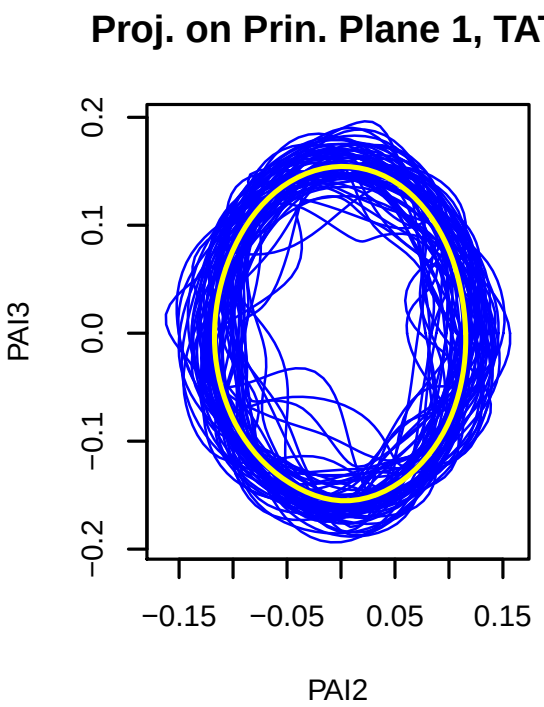
- If the basis $\{\psi_j\}$ is orthonormal (such as the Fourier basis for periodic data like minicircles), then functional PCA is usual PCA of the coefficient matrix $C = (c_{ij})$

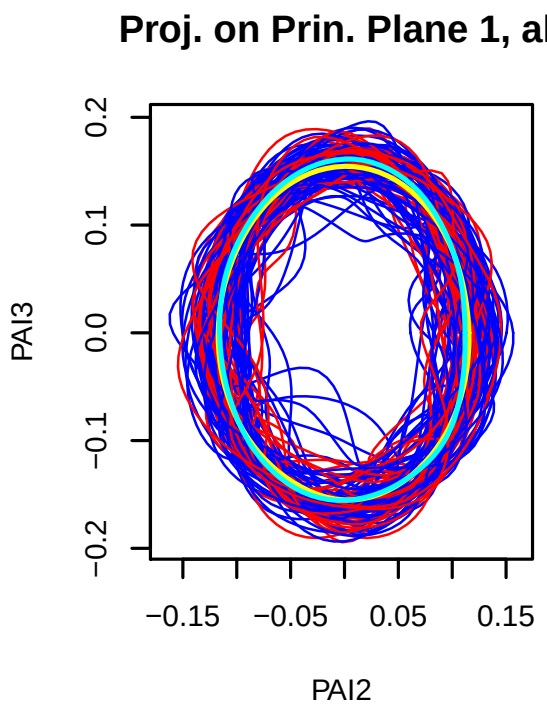
Outline

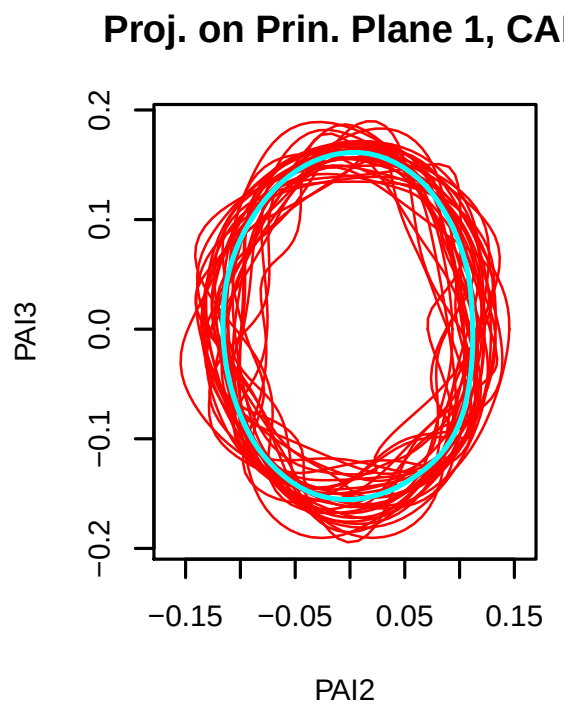
Means of PAI coordinates



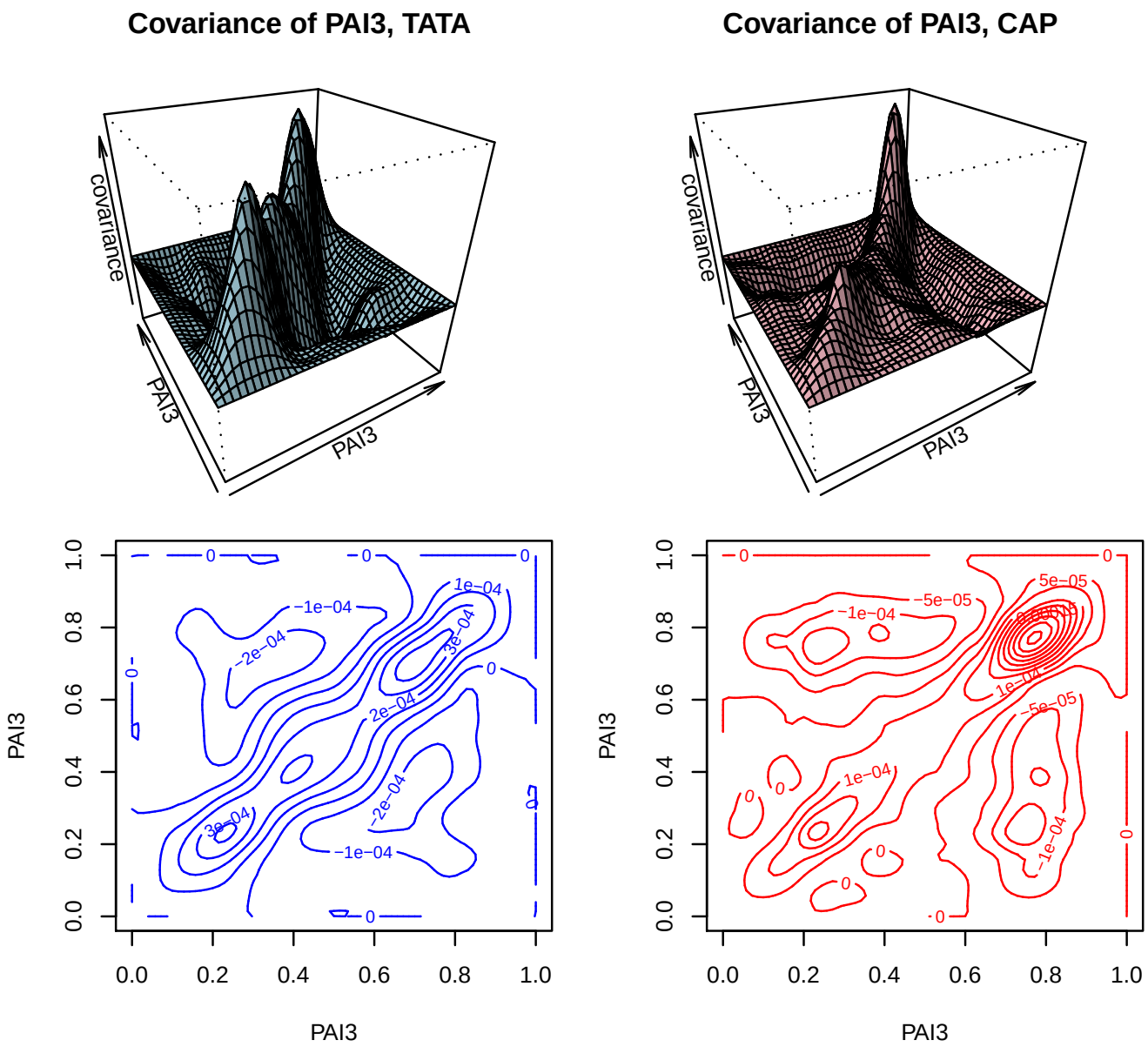
Means on the principal plane







Covariance functions



Testing problem

- Situation: $X_1, \ldots, X_{n_1}$, $Y_1, \ldots, Y_{n_2}$ independent samples of Gaussian stochastic processes with means μ_X, μ_Y and covariance operators $\mathscr{R}_X, \mathscr{R}_Y$
- Mean functions appear to be equal
  $\hookrightarrow$ Focus on covariance operators
- Hypothesis testing problem

$$H_0 : \mathscr{R}_X = \mathscr{R}_Y \qquad \text{vs} \qquad H_1 : \mathscr{R}_X \neq \mathscr{R}_Y$$

- Use of a statistic like $\hat{\mathscr{R}}_X^{-1} \hat{\mathscr{R}}_Y$ impossible (noninvertibility)
- Instead, use the difference

$$\hat{\mathscr{R}}_X - \hat{\mathscr{R}}_Y$$

which should be close to the zero operator under the null

Hilbert–Schmidt operator norm

- Need to measure the distance of $\hat{\mathscr{R}}_X - \hat{\mathscr{R}}_Y$ from a zero operator
  $\hookrightarrow$ Need an operator norm
- The Hilbert–Schmidt operator norm is defined as

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

  (multivariate analog: $\|A\|_{\mathrm{F}}^2 = \sum_{i,j} a_{ij}^2$ for a matrix A)
- The distribution of

$$\|\hat{\mathscr{R}}_X - \hat{\mathscr{R}}_Y\|_{\mathrm{HS}}^2$$

  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 based on
$$\|\hat{\mathscr{R}}_X^K - \hat{\mathscr{R}}_Y^K\|_{\mathrm{HS}}^2$$
where $\hat{\mathscr{R}}_X^K = \pi_K \hat{\mathscr{R}}_X \pi_K$, $\hat{\mathscr{R}}_Y^K = \pi_K \hat{\mathscr{R}}_Y \pi_K$
- Need a common basis (the same π_K for X, Y)
(a common reference coordinate system)
- We use $\pi_K = \hat{\pi}_K$ projecting on
$$\hat{\varphi}_1^{XY}, \ldots, \hat{\varphi}_K^{XY}$$
(eigenfunctions of the pooled-sample estimator $\hat{\mathscr{R}} = \frac{n_1}{n}\hat{\mathscr{R}}_X + \frac{n_2}{n}\hat{\mathscr{R}}_Y$)

Test statistic

- The statistic is

$$\|\hat{\mathscr{R}}_X^K - \hat{\mathscr{R}}_Y^K\|_{\mathrm{HS}}^2 = \sum_{k=1}^{K}\sum_{j=1}^{K}\langle \hat{\varphi}_k^{XY}, (\hat{\mathscr{R}}_X - \hat{\mathscr{R}}_Y)\hat{\varphi}_j^{XY}\rangle^2$$

- The terms

$$\hat{\lambda}_{kj}^{X,XY} = \langle \hat{\varphi}_k^{XY}, \hat{\mathscr{R}}_X \hat{\varphi}_j^{XY}\rangle, \qquad \hat{\lambda}_{kj}^{Y,XY} = \langle \hat{\varphi}_k^{XY}, \hat{\mathscr{R}}_Y \hat{\varphi}_j^{XY}\rangle$$

  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

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

Asymptotics

- Under H_0 and Gaussian assumption,

$$T \xrightarrow[n\to\infty]{\mathcal{D}} \chi^2_{K(K+1)/2}$$

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

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

    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

$$T_1 = \frac{n_1 n_2}{2n} \sum_{k=1}^{K} \frac{(\hat{\lambda}_{kk}^{X,XY} - \hat{\lambda}_{kk}^{Y,XY})^2}{\hat{\lambda}_k^2}$$

  (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

$$\mathrm{PFC}(K) = \mathrm{GOF}_X(K) + \mathrm{GOF}_Y(K) + \frac{n_1}{n}\,\mathrm{PEN}_X(K) + \frac{n_2}{n}\,\mathrm{PEN}_Y(K)$$

(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 %

Simulations: level

Scenario A

- Equal covariance operators

	K				
Test	1	2	3	4	K^*
off-diag	0.051	0.056	0.057	0.056	0.059
diag	0.051	0.054	0.056	0.061	0.061

- The tests maintain the nominal level when $K \leq$ 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.

Simulations: power

Scenario B

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

	K				
Test	1	2	3	4	K^*
off-diag	0.443	0.315	0.223	0.174	0.175
diag	0.443	0.350	0.306	0.267	0.267

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

Simulations: power

Scenario E

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

	K				
Test	1	2	3	4	K^*
off-diag	0.055	0.267	0.686	0.976	0.975
diag	0.055	0.250	0.509	0.620	0.617

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

Simulations: power

Scenario F

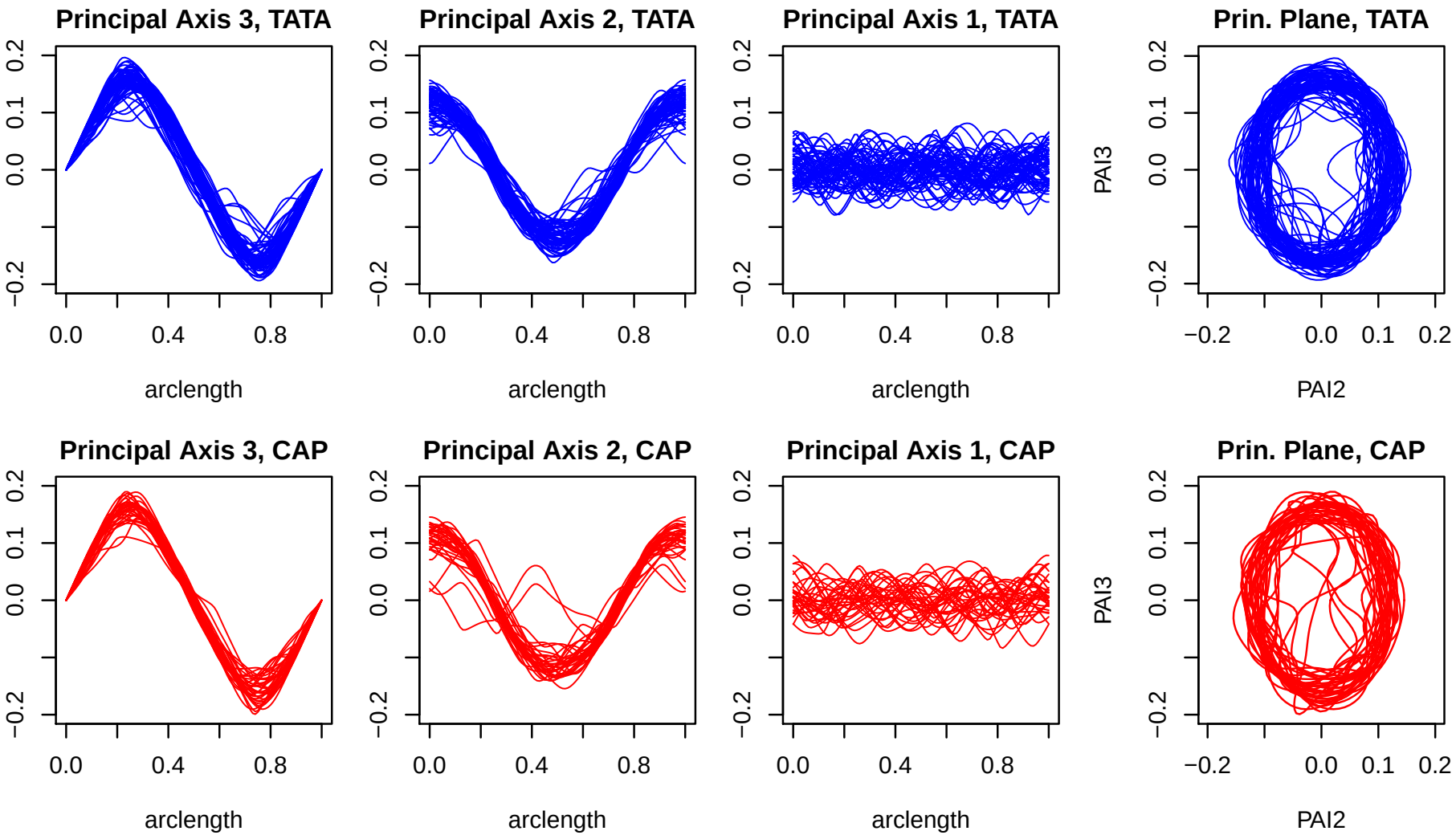
- Completely different eigenfunctions (sines vs time-shifted sines)
- Equal eigenvalues

	K				
Test	1	2	3	4	K^*
off-diag	0.273	0.706	0.916	1.000	1.000
diag	0.273	0.496	0.544	0.594	0.655

- Lower power for the diagonal test

Outline

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
arclength
PAI2
PAI3
−0.2
0.0
0.1
0.2
0.4
0.8

Spatial median, outlier detection

- The functional spatial median is defined as the solution to

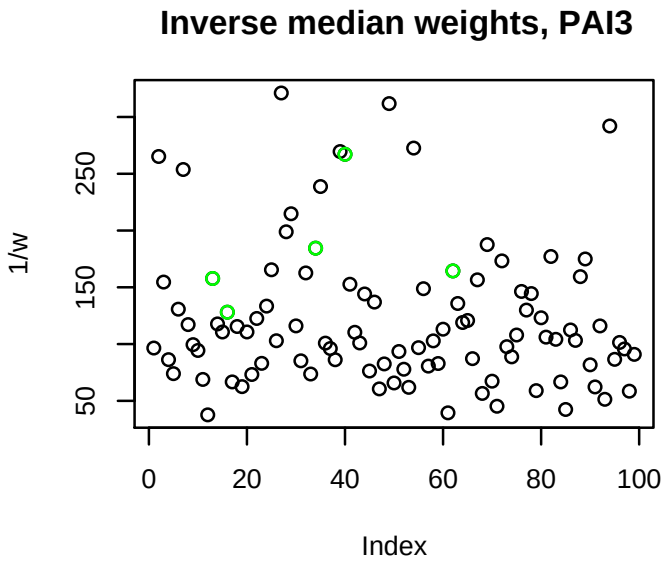
$$\min_{m \in L^2} \sum_{i=1}^{n} \|X_i - m\|_2 \qquad \text{or} \qquad \sum_{i=1}^{n} \frac{m - X_i}{\|m - X_i\|_2} = 0$$

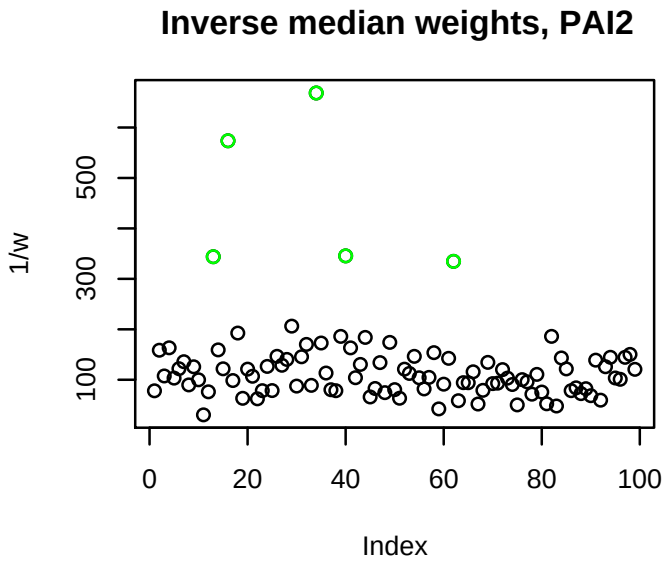
- The solution $\hat{m}$ can be written as the weighted sum

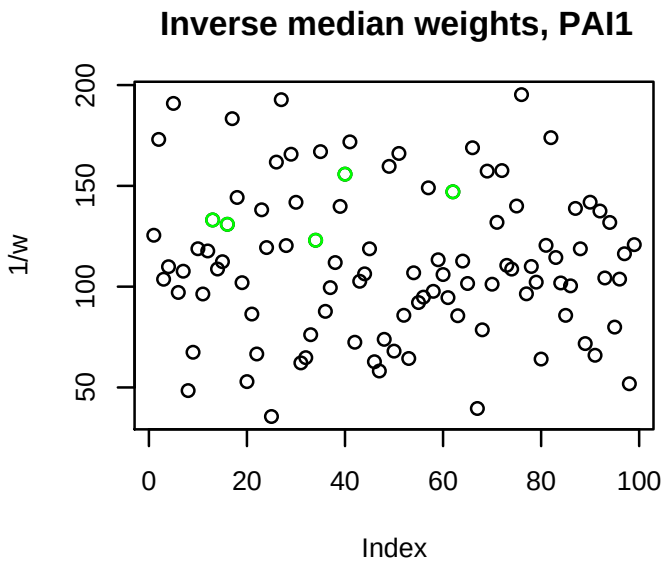
$$\hat{m} = \sum_{i=1}^{n} w_i X_i, \quad w_i \geq 0, \quad \sum_{i=1}^{n} w_i = 1$$

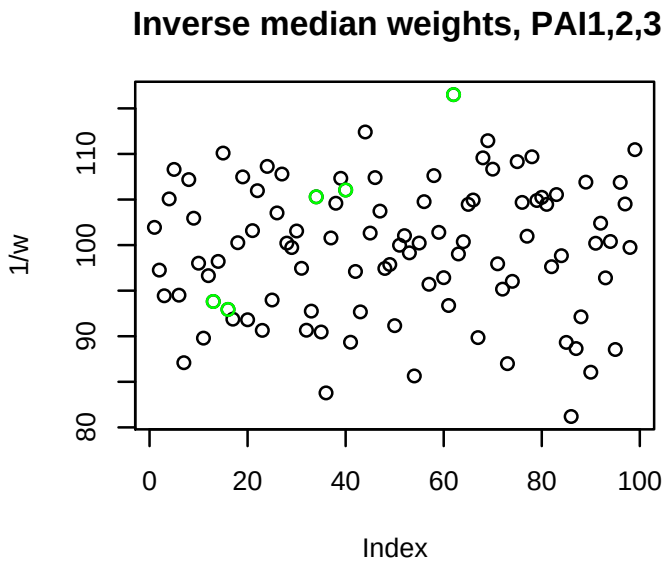
- Outliers have small weights w_i, large values of $1/w_i$ indicate outliers

Inverse spatial median weights

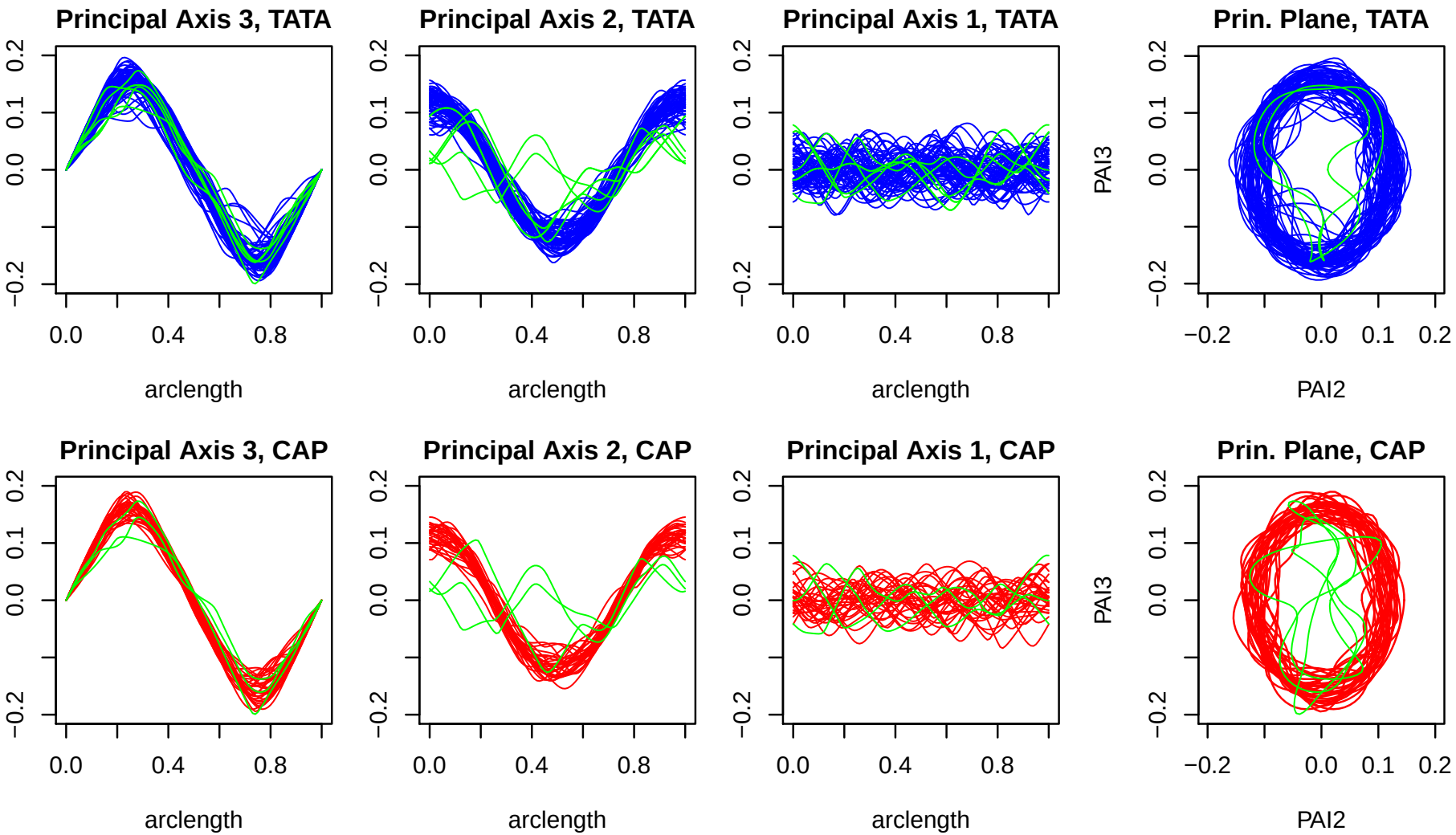




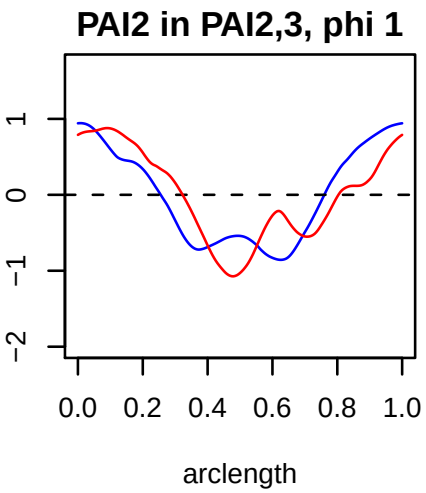


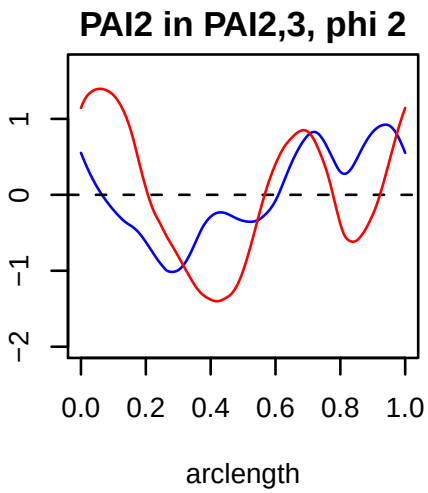


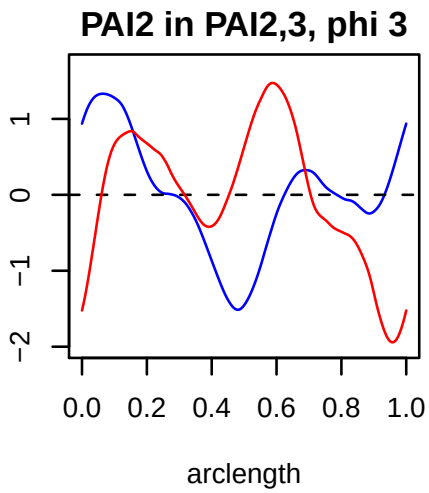
Outliers

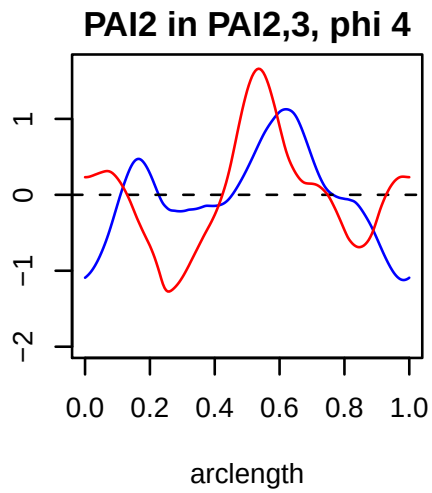

Principal Axis 3, TATA
Principal Axis 2, TATA
Principal Axis 1, TATA
Prin. Plane, TATA
Principal Axis 3, CAP
Principal Axis 2, CAP
Principal Axis 1, CAP
Prin. Plane, CAP
arclength
PAI2
PAI3
0.0
0.4
0.8
−0.2
0.0
0.1
0.2

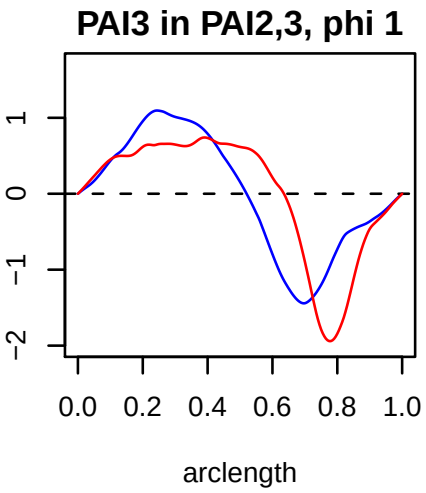
Joint PCA of PAI2,3: coordinates of eigenfunctions

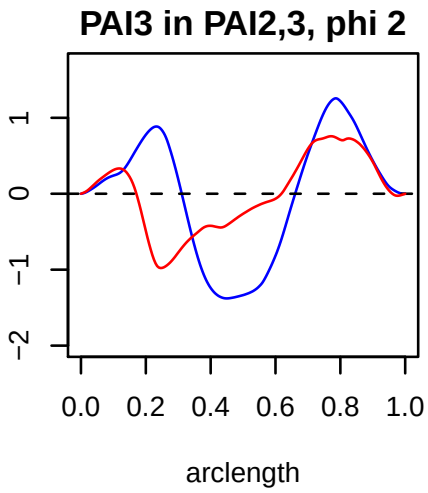


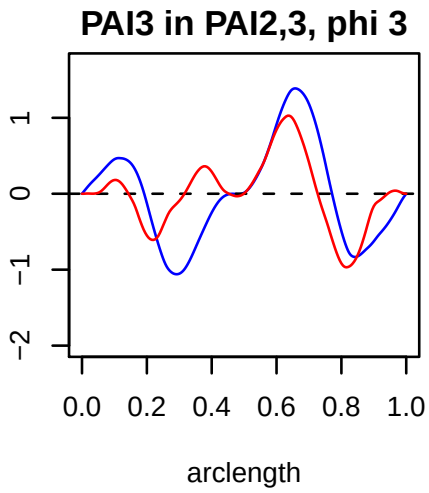


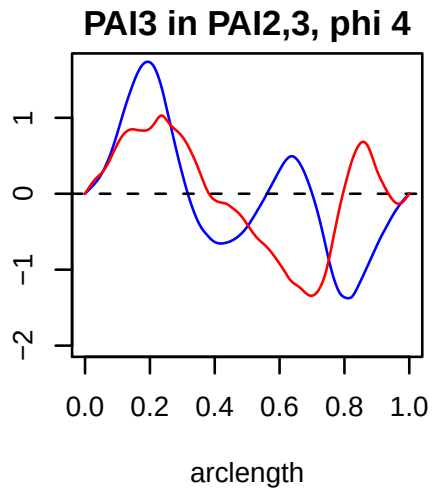




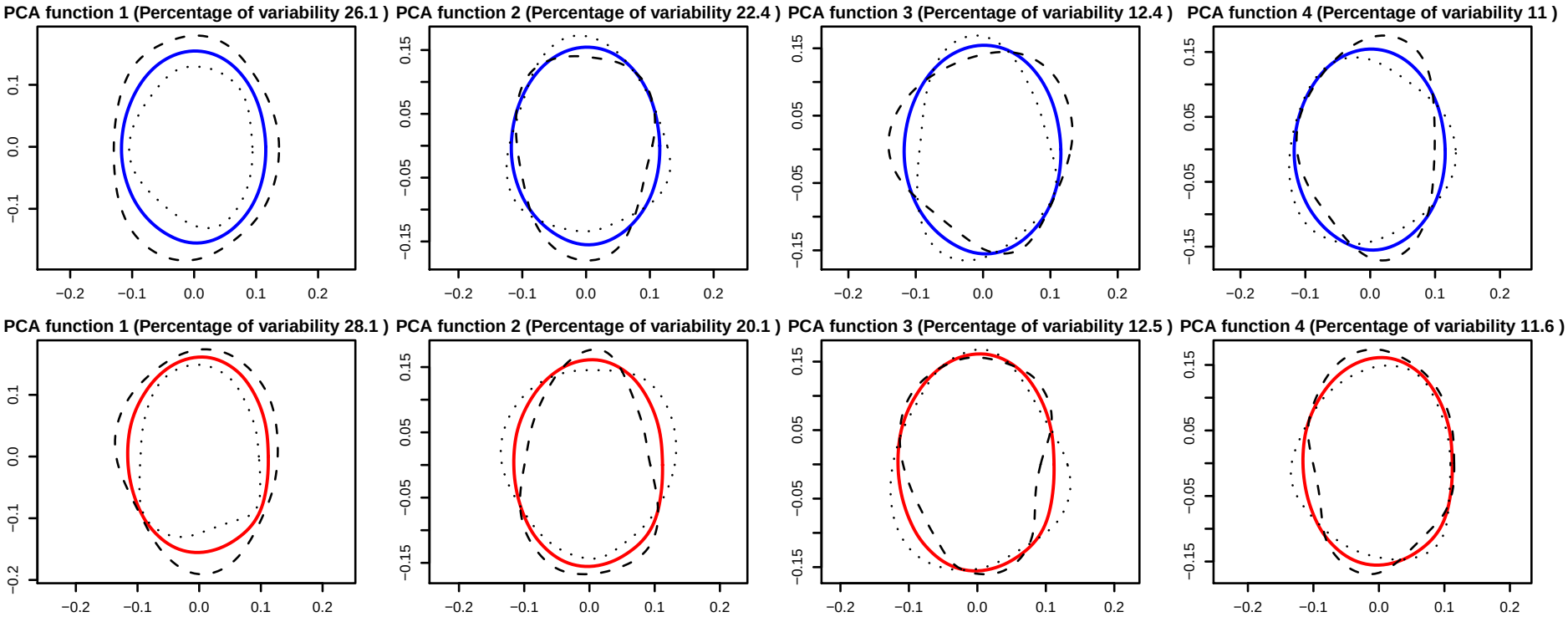




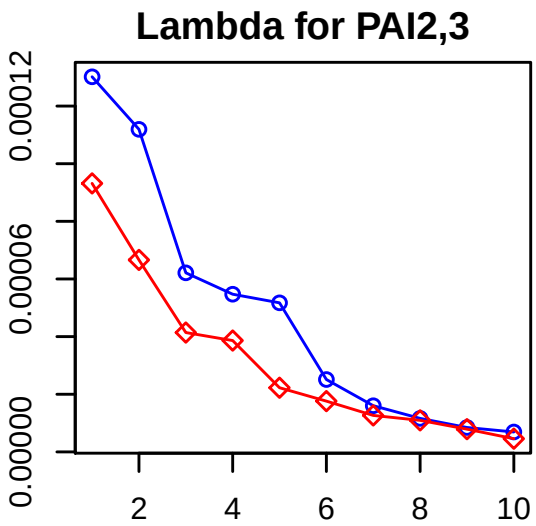


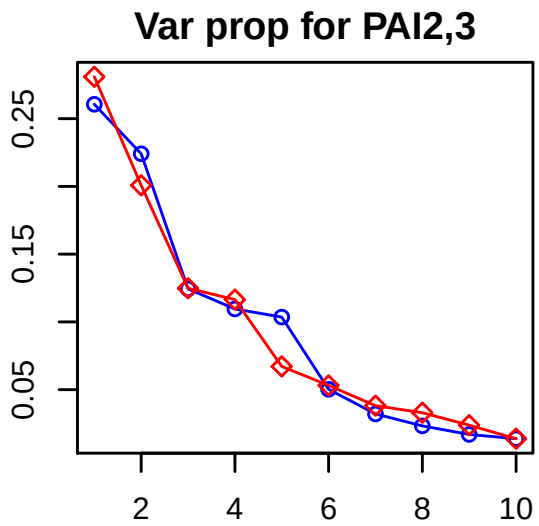


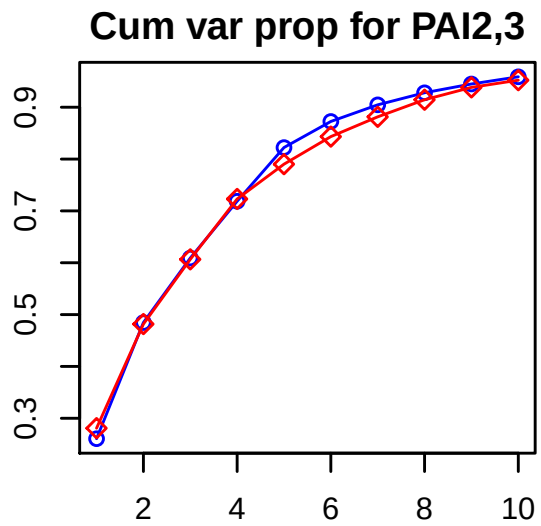
Joint PCA of PAI2,3: eigencircles



Joint PCA of PAI2,3: eigenvalues

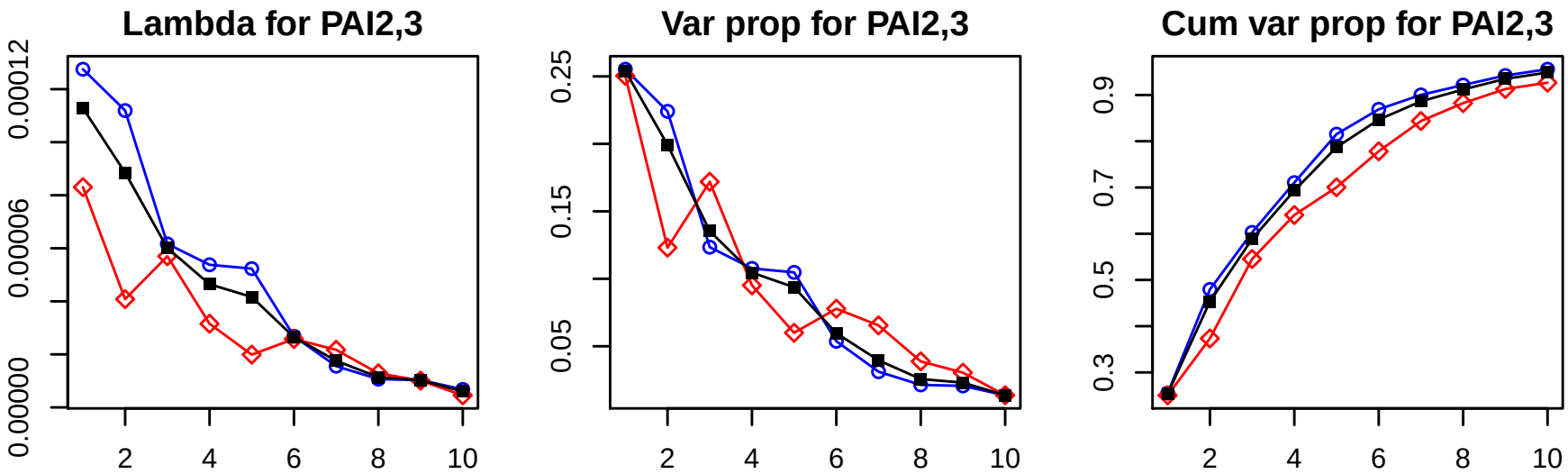






Selection of K (joint analysis of PAI2,3)

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



- Plots suggest $K = 6$ or 7
- Automatic choice $K = 7$

Test results

Test statistic: off-diagonal, transformed, χ^2 approximation

	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^{S}	0.022	0.886	0.051
5	0.001^{A}	0.053^{S}	0.555	0.009
6	0.010	0.087	0.327	0.005^{S}
7	0.019	0.098^{A}	0.360	0.023^{A}
8	0.046	0.173	0.148	0.094

(S = Selection based on scree plots, A = Automatic selection)

Outline

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 analaysis (outlier detection, order selection, testing, . . .)

- Outlook
  - Robustification
  - Normality testing
  - . . .