Stat
The ISI's Journal for the Rapid
Dissemination of Statistics Research
(wileyonlinelibrary.com) DOI: 10.100X/sta.0000

Covariance analysis for temporal data, with applications to DNA modelling

Ian L. Dryden[a]*, Blake C. Hill[b], Hao Wang[c], Charles A. Laughton[a]

Received 28 April 2017; Accepted 00 Month 2017

We introduce methodology for analysing the mean size-and-shape and covariance matrix of landmark data that are collected over time. Motivated by a study of DNA damage, we study some permutation based tests for investigating significant differences in the structure of the mean and the variability/covariance of size-and-shape of point sets which evolve over time. The covariance matrix tests make use of some recently introduced metrics for comparing covariance matrices. We demonstrate that the tests have the correct significance level in various simulation studies, and we also investigate the relative power of the tests. Finally we apply the procedures to the DNA datasets, providing practical insights into different types of DNA damage. Copyright © 0000 John Wiley & Sons, Ltd.

Keywords: Auto-regressive, covariance matrix, DNA, non-Euclidean, non-parametric, permutation test, size-and-shapes, temporal

1. Introduction

In many applications, it is of interest to test the equality of two covariance matrices of size-and-shape data collected from two populations. By 'size-and-shape' we mean the geometrical properties of an object that are unchanged by translation and rotation (Dryden and Mardia, 2016). Our motivating example is to understand the differences in size-and-shapes between damaged versus undamaged DNA molecules (Jiranusornkul and Laughton, 2008). The issue is important because it sheds light on a key question in molecular recognition, namely how DNA repair proteins locate a region of damage. The damaged regions may appear to have only very slight modifications in structure, within a vast excess of normal, and structurally dynamic DNA. Comparing size-and-shape distributions of damaged and undamaged DNA modelcules could also be useful for finding cures for diseases stemming from damaged DNA strands, such as amyotrophic lateral sclerosis (ALS), Huntington's disease (Ahmad, 2010) and Cockayne's syndrome (Friedberg et al., 1995).

[a] University of Nottingham, University Park, Nottingham, NG7 2RD, UK
[b] University of South Carolina, Columbia, SC 29208, USA
[c]Prime Quantitative Research LLC, East Lansing, MI 48824 USA.
***Email: ian.dryden@nottingham.ac.uk**

Prepared using staauth.cls [Version: 2012/05/12 v1.00]

The problem of testing the equality of two covariance matrices is, however, well known to be statistically challenging because the covariance matrix is positive definitive and often of high dimensionality. The traditional likelihood ratio tests may fail for high dimensional problems (Bai et al., 2009). Advances in comparing high-dimensional covariance matrices include Schott (2007); Srivastava and Yanagihara (2010); Li and Chen (2012); Cai et al. (2013). However, these tests may not be ideal for our DNA size-and-shape problem. For one reason, they are often built upon either the norms defined on the Euclidean space (e.g., Frobenius norm) or the assumption of normal populations. An alternative based on an L_∞-norm is given by Chang et al. (2017). The space of covariance matrices is most naturally described as non-Euclidean and we consider a variety of different metrics, taking into account temporal correlations.

The motivating DNA dataset comes from the molecular dynamics (MD) simulations conducted by Jiranusornkul and Laughton (2008) to simulate the structure and dynamics of DNA duplexes. Our analysis focuses on six undamaged DNA molecules, labeled AGA, AGC, AGG, TGA, TGC, and TGT and their respective damaged counterparts AFA, AFC, AFG, TFA, TFC, and TFT. Here the four letters A, C, G, and T in undamaged DNA molecules represent the four nucleotide bases of a DNA strand: adenine, cytosine, guanine, thymine. These six undamaged molecules encode different genetic information: TGC and TGT for the amino acid Cysteine, AGA and AGG for the amino acid Arginine, AGC for the amino acid Serine, and TGA, as a stop codon, does not form an amino acid but signals the termination of the polypeptide chain (Petsko and Ringe, 2004). Given an undamaged DNA molecule, its damaged counterpart is generated by replacing guanine with FapydG, which is one of the most prevalent guanine-derived lesions formed under O2-deficient conditions by ionizing radiation and other agents that produce reactive oxygen species (Jiranusornkul and Laughton, 2008; Douki et al., 1997; Pouget et al., 2000; Crespo-Hernández and Arce, 2004). Take the pair of TGC and TFC as an example. The initial dodecamer DNA duplex of TGC consists of 2×12 bases $d(\mathrm{CTTTTGCAAAAG})_2$. The damaged counterpart TFC is then generated by replacing the middle base G with F such that the dodecamer becomes $d(\mathrm{CTTTTFCAAAAG}) \cdot d(\mathrm{CTTTTGCAAAAG})$. Other pairs are simulated in a similar manner.

The biological question of interest is: does the introduction of a site of damage into each of these different DNA sequences produce a consistent structural or dynamic 'signature' that the repair protein could use to recognise it?

The configuration of DNA is recorded for every picosecond (1×10^{-12} seconds) and consists of the xyz-coordinates of 22 landmarks which are located at the phosphorous atoms in duplex DNA (Dryden et al., 2002). Thus, each observation can be represented as a 22×3 matrix whose rows and columns correspond to landmarks and dimensions respectively. A total of 2500 one-picosecond observations are collected for each molecule. In summary, the dataset consists of times series of 22×3 data matrices $\mathbf{X}_i^M$ where $i = 1, 2, \ldots, 2500$ indexes observations and $M \in \{$AGA, AGC, AGG, TGA, TGC,TGT, AFA, AFC, AFG, TFA, TFC, TFT$\}$ indexes molecules.

2. Background

2.1. Generalized Procrustes analysis and tangent space

As our interests lie in the differences in size-and-shape rather than the relative rotations and locations, we eliminate the differences due to arbitrary translation and rotation via generalized Procrustes analysis (GPA). GPA involves minimizing the sum of squares (Gower, 1975; Goodall, 1991; Dryden and Mardia, 2016):

$$G(\mathbf{X}_1, \ldots, \mathbf{X}_n) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=i+1}^{n} ||(\mathbf{X}_i \boldsymbol{\Gamma}_i + \mathbf{1}_k \boldsymbol{\gamma}_i^T) - (\mathbf{X}_j \boldsymbol{\Gamma}_j + \mathbf{1}_k \boldsymbol{\gamma}_j^T)||^2, \quad (1)$$

over rotations Γ_i and translations $\boldsymbol{\gamma}_i$, $i = 1, \ldots, n$. It can be also shown that the minimization problem of (1) is equivalent to minimizing $\sum_{i=1}^{n} ||\mathbf{X}_i\boldsymbol{\Gamma}_i + \mathbf{1}_k\boldsymbol{\gamma}_i^T - \boldsymbol{\mu}||^2$ with respect to $\boldsymbol{\Gamma}_i$, $\boldsymbol{\gamma}_i$, and $\boldsymbol{\mu}$, where $\boldsymbol{\mu}$ is the mean size-and-shape. Let $\hat{\boldsymbol{\Gamma}}_i$ and $\hat{\boldsymbol{\gamma}}_i$ be the Procrustes estimators of rotation and translation and let $\mathbf{X}_i^R = \mathbf{X}_i\hat{\boldsymbol{\Gamma}}_i + \mathbf{1}_k\hat{\boldsymbol{\gamma}}_i^T$ be the i^{th} registered observation after Procrustes analysis.

The size-and-shape space is difficult to work with directly because of its nonhomogeneous nature (Kendall et al., 1999), in particular it is a non-Euclidean manifold with singularities. The Procrustes tangent space provides a Euclidean approximation to facilitate statistical inference, and this is appropriate when the data are reasonably concentrated as in our DNA application. The tangent coordinates for $\mathbf{X}_i$ at $\boldsymbol{\mu}$ denoted as $\mathbf{T}_i$ are given by (Dryden and Mardia, 2016)

$$\mathbf{T}_i = \mathbf{T}_{\widehat{\boldsymbol{\mu}}}(\mathbf{X}_i^R) = \mathbf{X}_i^R - \widehat{\boldsymbol{\mu}}, i = 1, 2, \ldots, n,$$

where $\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^R$ is the estimated $k \times m$ mean shape matrix. The tangent coordinates satisfy two constraints

$$\mathbf{T}_i^T\mathbf{1}_k = \mathbf{0}, \quad \widehat{\boldsymbol{\mu}}^T\mathbf{T}_i \quad = \mathbf{T}_i^T\widehat{\boldsymbol{\mu}}, \tag{2}$$

where the first equation contains the m translation constraints and the second equation contains the $\frac{1}{2}m(m-1)$ linear rotation constraints, leading to the tangent space of dimension $q = mk - \frac{1}{2}m(m-1) - m$.

After obtaining the tangent coordinates $\mathbf{T}_i$ of the size-and-shapes, we now wish to estimate the covariance matrix. However, there are two challenges in this task: the constraints due to both the tangent space and positive definiteness, and the proliferation of parameters due to high dimension.

Let $\mathbf{u}$ be a $km \times q$ matrix with orthonormal columns in the directions of strictly positive variability in the tangent space at the mean $\widehat{\boldsymbol{\mu}}$. We can project into the lower q-dimensional sub-space of the tangent space which has full rank covariance matrix by pre-multiplying the tangent co-ordinate km-vector $\text{vec}(\mathbf{T}_i)$ by the transpose of $\mathbf{u}$, where $\text{vec}(\mathbf{T}_i)$ is the km-vector formed by stacking the columns of $\mathbf{T}_i$. For example, $\mathbf{u}$ could be formed with columns as eigenvectors of the covariance matrix of the tangent co-ordinates corresponding to the q directions of strictly positive variability. Alternatively, we can select q of the tangent co-ordinates in order to work with full-rank covariance matrices, and this will be our approach. For our 3D DNA dataset there are $k = 22$ landmarks and hence we work with the full rank covariance matrices from $q = 3k - 6 = 60$ co-ordinates.

2.2. Factored model

A factored covariance structure given by $\boldsymbol{\Sigma}_r \otimes \boldsymbol{\Sigma}_c$ where $\boldsymbol{\Sigma}_r$ is a $r \times r$ covariance matrix and $\boldsymbol{\Sigma}_c$ is a $c \times c$ covariance matrix. Such covariance structures can be estimated using the following algorithm (Dutilleul, 1999), which gives the maximum likelihood estimate for the factored covariance matrix assuming multivariate normal data.

0) Find $\mathbf{W}_i = \text{vec}_3^{-1}(\text{vec}(\mathbf{T}_i))$ and $\text{vec}_m^{-1}(\mathbf{y})$ makes a matrix of m columns and r rows from the $mr \times 1$ vector $\mathbf{y}$ by making the elements $1, \ldots, r$ of $\mathbf{y}$ column 1, elements $(r+1), \ldots, (2r)$ of $\mathbf{y}$ column 2, $\ldots$, and elements $((m-1)r+1), \ldots, (mr)$ column m of $\text{vec}_m^{-1}(\mathbf{y})$.

1) Let $\Sigma_{c^*} = \mathbf{I}_c$

2) Let $\boldsymbol{\Sigma}_{r^*} = \frac{1}{nc}\sum_{i=1}^{n}(\mathbf{W}_i - \overline{\mathbf{W}})\boldsymbol{\Sigma}_{c^*}^{-1}(\mathbf{W}_i - \overline{\mathbf{W}})^T$

3) Let $\Sigma_{c^+} = \frac{1}{nr}\sum_{i=1}^{n}(\mathbf{W}_i - \overline{\mathbf{W}})^T\boldsymbol{\Sigma}_{r^*}^{-1}(\mathbf{W}_i - \overline{\mathbf{W}})$

4) Let $\boldsymbol{\Sigma}_{r^+} = \frac{1}{nc}\sum_{i=1}^{n}(\mathbf{W}_i - \overline{\mathbf{W}})\boldsymbol{\Sigma}_{c^+}^{-1}(\mathbf{W}_i - \overline{\mathbf{W}})^T$

5) If $||\mathbf{\Sigma}_{c^+} - \mathbf{\Sigma}_{c^*}|| > \epsilon$ or $||\mathbf{\Sigma}_{r^+} - \mathbf{\Sigma}_{r^*}|| > \epsilon$, then let $\mathbf{\Sigma}_{c^*} = \mathbf{\Sigma}_{c^+}$ and $\mathbf{\Sigma}_{r^*} = \mathbf{\Sigma}_{r^+}$ and repeat steps 3-5

Once the algorithm has converged, the maximum likelihood estimates of $\mathbf{\Sigma}_c$ and $\mathbf{\Sigma}_r$ are $\widehat{\mathbf{\Sigma}}_c$ and $\widehat{\mathbf{\Sigma}}_r$, respectively. Since the maximum likelihood estimates can only be found up to a scale factor, each matrix is normalized by so that $\mathrm{trace}(\mathbf{\Sigma}_c) = c$ and the scale factors cancel in the product. We will use the factored model in analysis of the 3D DNA Procrustes registered data by decomposing using $r = 20$ and $c = 3$ dimensions for a factored covariance structure in $q = 60$ dimensions of the tangent space to size-and-shape space.

2.3. Distances between covariance matrices

A key step in comparing two covariance matrices is to choose distance metrics that quantify their differences. Our first dissimilarity measure is motivated by the classical Box's M test (Box, 1949) which is a likelihood ratio test of homogeneity of two or more covariance matrices. In the case of two covariances of dimension p, the Box's M test statistic M is given by

$$M = \gamma \sum_{i=1}^{2} (n_i - 1)\log(|\mathbf{S}_i^{-1}\mathbf{S}|), \quad \gamma = 1 - \frac{2p^2 + 3p - 1}{6(p+1)}\left(\sum_{i=1}^{2} \frac{1}{n_i - 1} - \frac{1}{n-2}\right),$$

where n_i and $\mathbf{S}_i$ are the sample size and the unbiased estimator of the covariance matrix of the ith population, $n = n_1 + n_2$ is the total sample size, and $\mathbf{S} = \sum_{i=1}^{2}(n_i - 1)\mathbf{S}_i/(n-2)$ is the pooled estimator. Box's M has an asymptotic chi-square distribution and often works well for lower-dimensional problems such as $p \leq 5$ (Mardia et al., 1979). In shape analysis, the dimension p is usually large, causing the approximation of the asymptotic distributions to fail. Nevertheless, we could regard the Box's M test statistic M as a dissimilarity measure between two covariance matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ and develop a permutation test based upon it. In the example of DNA size-and-shapes, the estimates from the factored models in Section 2.2 will be used as $\mathbf{S}_1$ and $\mathbf{S}_2$. Other metrics considered are: Riemannian distance, Procrustes distance, Procrustes shape distance, Cholesky distance, power distance, Euclidean distance, log Euclidean distance, and Riemannian Le distance. Given two covariance matrices $\mathbf{S}_1$ and $\mathbf{S}_2$, we give the definition of these distances in the following list. Note that $||\mathbf{F}|| = \sqrt{\mathrm{trace}(\mathbf{F}^T\mathbf{F})}$ denotes the Frobenius norm and $O(k)$ denotes the space of orthogonal matrix of size k.

1. The Riemannian distance (e.g., Pennec, 2006; Fletcher and Joshi, 2007):

$$d(\mathbf{S}_1, \mathbf{S}_2)_{\mathrm{Riem}} = ||\log(\mathbf{S}_1^{-1/2}\mathbf{S}_2\mathbf{S}_1^{-1/2})||,$$

   where $\log(\mathbf{S}_i)$ is the natural log of the matrix $\mathbf{S}_i$ and can be found by first applying the eigendecomposition $\mathbf{S}_i = \mathbf{U}_i\mathbf{\Lambda}_i\mathbf{U}_i^T$ and then setting $\log(\mathbf{S}_i) = \mathbf{U}\log(\mathbf{\Lambda}_i)\mathbf{U}_i^T$ with $\log(\mathbf{\Lambda}_i)$ being a diagonal matrix of the logarithm of the diagonal elements of Λ_i on the diagonal.

2. The Procrustes distance (Dryden et al., 2009):

$$d(\mathbf{S}_1, \mathbf{S}_2)_{\mathrm{Proc}} = \inf_{\mathbf{R}\in O(k)} ||\mathbf{S}_1^{1/2} - \mathbf{S}_2^{1/2}\mathbf{R}||,$$

   where the matrices $\mathbf{S}_i^{1/2} = \mathbf{U}_i\mathbf{\Lambda}_i^{1/2}\mathbf{U}_i^T$ are the symmetric square root matrices determined by the spectral decomposition $\mathbf{S}_i = \mathbf{U}_i\mathbf{\Lambda}_i\mathbf{U}_i^T$.

3. The Procrustes shape distance (Dryden et al., 2009):

$$d(\mathbf{S}_1, \mathbf{S}_2)_{\mathrm{PShape}} = \inf_{R\epsilon O(k), \beta > 0} ||\frac{\mathbf{S}_1^{1/2}}{||\mathbf{S}_1^{1/2}||} - \beta\mathbf{S}_2^{1/2}\mathbf{R}||.$$

4. The Cholesky distance (Wang et al., 2004):

$$d(\mathbf{S}_1, \mathbf{S}_2)_{\text{Chol}} = ||\text{chol}(\mathbf{S}_1) - \text{chol}(\mathbf{S}_2)||.$$

where $\text{chol}(\mathbf{S}_i) = \mathbf{L}_{S_i}$ is the Cholesky decomposition of $\mathbf{S}_i$ where $\mathbf{S}_i = \mathbf{L}_{S_i}\mathbf{L}_{S_i}^T$ and $\mathbf{L}_{S_i}$ is a lower triangular matrix with positive entries along the diagonal.

5. The Power distance (Dryden et al., 2009):

$$d(\mathbf{S}_1, \mathbf{S}_2)_{\text{Power}} = \frac{1}{\alpha}||\mathbf{S}_1^\alpha - \mathbf{S}_2^\alpha||, \quad \alpha \in (0, 1),$$

where the symmetric matrices $\mathbf{S}_i^\alpha = \mathbf{U}_i\mathbf{\Lambda}_i^\alpha\mathbf{U}_i^T$ are determined by the spectral decomposition $\mathbf{S}_i = \mathbf{U}_i\mathbf{\Lambda}_i\mathbf{U}_i^T$. The value of α used in this analysis is set at 0.5, the symmetric square root case.

6. The Euclidean distance (e.g., Meyer, 2000):

$$d(\mathbf{S}_1, \mathbf{S}_2)_{\text{Eucl}} = ||\mathbf{S}_1 - \mathbf{S}_2||.$$

7. The Log Euclidean distance (Arsigny et al., 2007):

$$d(\mathbf{S}_1, \mathbf{S}_2)_{\text{LE}} = ||\log(\mathbf{S}_1) - \log(\mathbf{S}_2)||.$$

8. The Riemannian Le distance (Su et al., 2011) :

$$d(\mathbf{S}_1, \mathbf{S}_2)_{\text{RiemLe}} = \frac{1}{2}||\log(\mathbf{C}^{-1/2}\mathbf{D}\mathbf{C}^{-1/2})||,$$

where $\mathbf{C} = \mathbf{S}_1\mathbf{S}_1^T$ and $\mathbf{D} = \mathbf{S}_2\mathbf{S}_2^T$.

Note that the Procrustes distances above could alternatively use the Cholesky decomposition instead of the symmetric square root (Dryden et al., 2009), although the Cholesky decomposition is not available for matrices with zero eigenvalues, which is a disadvantage. A few important properties of these distances are worth noting. First, the Riemannian distance is affine invariant. That is, for a general square matrix, $\mathbf{A}$, $d(\mathbf{A}\mathbf{S}_1\mathbf{A}, \mathbf{A}\mathbf{S}_2\mathbf{A}^T) = d(\mathbf{S}_1, \mathbf{S}_2)$. Second, the Riemannian distance, Log-Euclidean distance, Procrustes shape distance, and Box's M statistic are scale invariant meaning that for $c > 0$, $d(c\mathbf{S}_1, c\mathbf{S}_2) = d(\mathbf{S}_1, \mathbf{S}_2)$. Finally, except for the Cholesky distance, all other distance measures are rotation invariant, meaning that for an orthogonal matrix $\mathbf{R}$, $d(\mathbf{R}\mathbf{S}_1\mathbf{R}^T, \mathbf{R}\mathbf{S}_2\mathbf{R}^T) = d(\mathbf{S}_1, \mathbf{S}_2)$. The distances can all be computed using the function `distcov` in the `shapes` library in R (Dryden, 2017).

In this paper we will use the covariance distances to construct test statistics to investigate differences in the structure of size-and-shape variability between the undamaged and damaged molecules for the six types of DNA: AGA, AGC, AGG, TGA, TGC, TGT, where the damaged versions have the middle 'G' replaced by 'F'.

3. Exploratory data analysis

3.1. Canonical variate analysis

We initially visualize the data using canonical variate analysis (Timm, 2002) carried out by the `shapes.cva` function in the R `shapes` package (Dryden, 2017). The left hand panel of Figure 1 displays the first two canonical

variate scores for each of the six pairs of DNA molecules. As can be seen, the differences in mean shape structure between the damaged and undamaged molecules appear to be substantial for the AFC-AGC, TFT-TGT, and TFA-TGA molecule pairs (right column) but smaller for the AFA-AGA, AFG-AGG, and TFC-TGC molecule pairs (left column). The pattern is more apparent in the right hand panel of Figure 1 where the changes of the mean canonical variate scores from the undamaged to the damaged molecules are shown. In fact, the damaged molecule means are all lower in terms of the second canonical variate score, and those beginning with T are higher in the first canonical variate scores. As expected there are strong temporal correlations in the data, and inspection of the autocorrelation function (ACF) and partial ACF suggest that first order autoregressive models may be reasonable for the dominant principal component (PC) scores (figures not shown).

3.2. Mean shape differences

In a classical two-sample multivariate normal model, James' statistic can be used to test the mean differences without assuming equal covariances between the two populations (James, 1954; Seber, 1984). The test statistic is equivalent to the Hotelling T^2 test statistic (Hotelling, 1931) when the sample sizes are equal. However, the test is based on some strict parametric assumptions about the population and can be often improved by nonparametric procedures (Wasserman, 2006; Hall and Wilson, 1991; Amaral et al., 2007). For our DNA application we use permutation tests (Good, 1994), with sub-sampling to account for temporal correlation.

Let $\mathbf{X}_{i,D}^R$ and $\mathbf{X}_{i,U}^R$ be the ith registered damaged and undamaged molecule sample, respectively. The tangent coordinates are $\mathbf{T}_{i,U} = \mathbf{X}_{i,U}^R - \widehat{\boldsymbol{\mu}}$ and $\mathbf{T}_{i,D} = \mathbf{X}_{i,D}^R - \widehat{\boldsymbol{\mu}}$, where $\widehat{\boldsymbol{\mu}}$ is the estimated mean shape of the combined damaged and undamaged group data. We compare the means of the vectorized tangent coordinates $\mathbf{v}_{i,D} \equiv \mathrm{vec}(\mathbf{T}_{i,D})$ and $\mathbf{v}_{i,U} \equiv \mathrm{vec}(\mathbf{T}_{i,U})$ using the permutation test based on James' statistic:

$$J^2 = (\bar{\mathbf{v}}_D - \bar{\mathbf{v}}_U)^T \left(\frac{\mathbf{S}_{\mathbf{v}_D}}{n_1} + \frac{\mathbf{S}_{\mathbf{v}_U}}{n_2} \right)^{\dagger} (\bar{\mathbf{v}}_D - \bar{\mathbf{v}}_U), \tag{3}$$

where $\bar{\mathbf{v}}_D$ and $\mathbf{S}_{\mathbf{v}_D}$ ($\bar{\mathbf{v}}_U$ and $\mathbf{S}_{\mathbf{v}_U}$) are the sample mean and covariance matrix of $\{\mathbf{v}_{i,D}\}$($\{\mathbf{v}_{i,U}\}$), and † represents the Moore-Penrose generalized inverse (Amaral et al., 2007). Specifically, we first evaluate (3) for the observed samples $\{\mathbf{T}_{i,D}\}$ and $\{\mathbf{T}_{i,U}\}$ and denote this value as J_0^2. Next, we permute the tangent coordinates between groups to form two new sets of tangent coordinates, evaluate (3) again but for the permuted samples, and denote its value as J_1^2. After repeating the permutation procedure h times and obtain $J_2^2, J_3^2, \ldots, J_h^2$, we compare the observed value J_0^2 with the simulated values of $\{J_v^2\}_{v=1}^h$ and compute the p-value as $\sum_{v=1}^h I(J_v^2 \geq J_0^2)/(h+1)$, where $I(J_v^2 \geq J_0^2) = 1$ if $J_v^2 \geq J_0^2$ and 0 otherwise. To account for the temporal autocorrelation we thin out the data so that we select every 100th, 50th, and 25th observation, which leads to sample sizes of size $n_1 = n_2 \in \{25, 50, 100\}$ respectively, and we take $h = 1000$. In Table 1 we see that there are significant size-and-shape differences in all but the AFA-AGA pair when $n_i = 100$ at the 5% level. When $n_i = 50$ the pairs AFA-AGA, AFG-AGG and TFC-TGC are not significantly different, and these are the pairs in the left hand column of Figure 1. Clearly AFA-AGA is the pair with the smallest mean size-and-shape difference between damaged and undamaged. Naturally there is a trade-off between removing correlation and reducing power when thinning.

4. Permutation tests for covariance matrices

4.1. Tests for covariance matrices with unequal mean shapes and temporal dependence

We now develop a procedure for testing the difference of covariance matrices in the presence of unequal means and temporal correlation. A test without considering the autocorrelations ignores the decrease of the effective sample size, thus underestimates parameter uncertainty, causing higher false rejections. To account for the correlation, we build time series models, which enable a pre-whitening method to remove the autocorrelation (Dryden et al., 2010). The residuals might be regarded as "uncorrelated" and can be analyzed by permutation test methods.

The pre-whitening step starts with the registration of DNA shapes in pairs based on the GPA described in Section 2.1. We then find the first PC score of the tangent coordinates and fit an AR(1) model to it. Let ψ be the estimated AR coefficient and Σ_n be $n \times n$ implied autocorrelation matrix. Then, the inverse of Σ_n can be expressed:

$$\mathbf{\Sigma}_n^{-1} \equiv \frac{1}{1-\psi^2} \begin{pmatrix} 1 & -\psi & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ -\psi & 1+\psi^2 & -\psi & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & -\psi & 1+\psi^2 & -\psi & 0 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & -\psi & 1+\psi^2 & -\psi \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & -\psi & 1 \end{pmatrix}. \tag{4}$$

In other applications other temporal structures may be appropriate, e.g. a more general ARMA model. After obtaining Σ_n^{-1}, we multiply the tangent coordinates by $\mathbf{\Sigma}_n^{-1/2}$, where $\mathbf{\Sigma}_n^{-1/2}$ is the square root of $\mathbf{\Sigma}_n^{-1}$, which can be obtained for example, through Cholesky decomposition. Assume that all coordinates follow the same $\mathrm{AR}(1)$ process as the first PC. Then those transformed tangent coordinates can be regarded as serially uncorrelated (approximately). Although the temporal structure may appear to vary among tangent coordinates, we prefer this single $\mathrm{AR}(1)$ model for all coordinates because it has the least amount of parameters to fit and also appear to be adequate in capturing most of the dependence structure in the DNA data. Admittedly, residual dependence structure might remain in the pre-whitened data, and so a block permutation step is implemented to safeguard against the remaining autocorrelations (as described below in step 8). The permutation test for testing equal covariances operates on the pre-whitened data.

1. Carry out pooled GPA on the combined dataset of the damaged/undamged pair to obtain the registered data $\{\mathbf{X}_{i,D}^R\}$ and $\{\mathbf{X}_{i,U}^R\}$.
2. Extract the first PC score of the vectors of $\{\mathrm{VEC}(\mathbf{X}_{i,D}^R), \mathrm{VEC}(\mathbf{X}_{i,U}^R)\}$; find the estimated $\mathrm{AR}(1)$ coefficient $\widehat{\psi}$ based on the first PC score; and define $\widehat{\Sigma}_T^{-1}$ by replacing ψ in Equation (4) with $\widehat{\psi}$.
3. Pre-whiten the damaged DNA shapes by computing $\mathbf{X}_{\mathrm{PW},D} = (\mathbf{\Sigma}_n^{-1/2} \otimes \mathbf{I}_k)^T \mathbf{Y}_D$ where $\mathbf{I}_k$ is the $k \times k$ identity matrix, $\mathbf{Y}_D$ is the $(nk) \times m$ matrix built by stacking the array slices of $\mathbf{X}_{i,D}^R$, k is the number of landmarks, m is the number of dimensions, and n is the number of observations. The ith pre-whitened configuration matrix $\mathbf{X}_{\mathrm{PW},i,D}$ is then the ith submatrix of size $k \times r$ in $\mathbf{X}_{\mathrm{PW},D}$. Pre-whiten the undamaged DNA shape data in a similar manner and denote the pre-whitened observations as $\mathbf{X}_{\mathrm{PW},i,U}$.

4. Obtain the pooled group mean $\widehat{\boldsymbol{\mu}}_{\text{pool}}$ of $\{\mathbf{X}_{\text{PW},i,D}\}$ and $\{\mathbf{X}_{\text{PW},i,U}\}$.
5. Find the pooled tangent coordinates $\mathbf{v}_{i,D} = T_{\widehat{\boldsymbol{\mu}}_{\text{pool}}}(\mathbf{X}_{\text{PW},i,D})$ and $\mathbf{v}_{i,U} = T_{\widehat{\boldsymbol{\mu}}_{\text{pool}}}(\mathbf{X}_{\text{PW},i,U})$ where $T_{\boldsymbol{\mu}}(X)$ denotes the tangent coordinates at $\boldsymbol{\mu}$.
6. In order to remove the effect of different means we centre each group by subtracting the group tangent mean from each $\mathbf{v}_{i,D}$ and $\mathbf{v}_{i,U}$ to get $\mathbf{w}_{i,D} = \mathbf{v}_{i,D} - \bar{\mathbf{v}}_D$ and $\mathbf{w}_{i,U} = \mathbf{v}_{i,U} - \bar{\mathbf{v}}_U$.
7. Work out the factored covariance matrix estimation for each group based on $\{\mathbf{w}_{i,D}\}$ and $\{\mathbf{w}_{i,U}\}$ to obtain $\widehat{\Sigma}_D$, $\widehat{\Sigma}_U$, and evaluate the test statistic $T_0 = d(\widehat{\Sigma}_D, \widehat{\Sigma}_U)$ based on a distance metric.
8. For $v = 1, \ldots, h$, swap/permute the blocks of size b between $\{\mathbf{X}_{\text{PW},i,D}\}$ and $\{\mathbf{X}_{\text{PW},i,U}\}$ at random to give $\{\mathbf{X}^v_{\text{PW},i,D}\}$ and $\{\mathbf{X}^v_{\text{PW},i,U}\}$; repeat Steps 4–8 to generate Monte Carlo samples of the test statistic T_v.
9. Compute the estimated p-value as $\sum_{v=1}^{h} I(T_v \geq T_0)/h$ where $I(T_v \geq T_0) = 1$ if $T_v \geq T_0$ and $I(T_v \geq T_0) = 0$ otherwise.

4.2. Simulation study

We carry out a simulation study generating the synthetic configuration matrix time series $\{\mathbf{X}_{i,D}\}$ and $\{\mathbf{X}_{i,U}\}$ for damaged and undamaged DNA from AR(1) models:

$$\mathbf{X}_{i,D} = \rho(\mathbf{X}_{i-1,D} - \boldsymbol{\mu}_D) + \boldsymbol{\mu}_D + \mathbf{e}_{i,D}, \quad \mathbf{X}_{i,U} = \rho(\mathbf{X}_{i-1,U} - \boldsymbol{\mu}_U) + \boldsymbol{\mu}_U + \mathbf{e}_{i,U}, \tag{5}$$

where the residuals $\{\mathbf{e}_{i,D}\}$ and $\{\mathbf{e}_{i,U}\}$ are serially independent and normally distributed with zero mean and covariance matrices $\boldsymbol{\Sigma}_D$ and $\boldsymbol{\Sigma}_U$, respectively. For the mean size-and-shapes $\boldsymbol{\mu}_D$ and $\boldsymbol{\mu}_U$, we allow them to be unequal and close to the real DNA size-and-shapes, for example, of the TGC and TFC molecules. Let $\mathbf{A}$ and $\mathbf{B}$ be the mean size-and-shapes of the authentic undamaged TGC and damaged TFC samples. We register $\mathbf{A}$ and $\mathbf{B}$ using OPA, treating the undamaged mean $\mathbf{A}$ as the reference size-and-shape and the damaged mean $\mathbf{B}$ as the size-and-shape to be transformed. After registration, we let $\widehat{\mathbf{A}}$ be the centered mean size-and-shape and $\widehat{\mathbf{B}}$ be the Procrustes registered shape and then set $\boldsymbol{\mu}_U = \widehat{\mathbf{A}}$ $\boldsymbol{\mu}_D = \widehat{\mathbf{B}}$ for the simulation. For the covariance matrices Σ_U and Σ_D, we set $\Sigma_U = \Sigma_D = \Sigma^*_U$ where Σ^*_U is the covariance matrix of the undamaged TGC in the size study and choose them to be different in the power study. We consider $\rho \in \{0, 0.5, 0.9, 0.99\}$, which covers a wide range of autocorrelation levels. In each simulation, we generate 2500 observations from (5) and then apply a block sampling method with block size equal to $b = 100$ and permutation replication $h = 300$. We repeat the simulation 1000 times to empirically estimate size and power.

Table 2 displays the simulated size for the landmark size $k = 22$ at a 5% significance level for different correlations. All simulated test sizes are close to the nominal value of 0.05, suggesting that the pre-whitening method generally produces the correct size regardless of the strength of temporal dependence. To investigate power we set Σ_U and Σ_D in the data generating process (5) to be different by letting $\boldsymbol{\Sigma}_D = \boldsymbol{\Sigma}^*_U + \kappa(\boldsymbol{\Sigma}^*_D - \boldsymbol{\Sigma}^*_U)$ and $\Sigma_U = \Sigma^*_U$ where $\boldsymbol{\Sigma}^*_D$ and $\boldsymbol{\Sigma}^*_U$ are the observed covariance matrices for molecules TGC and TFC and $\kappa \in (0, 1)$. Figure 2 plots the power functions against κ under various scenarios differing in ρ, with $k = 22$ landmarks. All curves are upward sloping, indicating that the power increases as the value of κ increases. This is because the assumed value for $\boldsymbol{\Sigma}_D$ departs further from $\boldsymbol{\Sigma}_U$ as κ becomes larger. The distances that give tests with the highest power are the Riemannian, log-Euclidean and Box's M tests with the Riemannian Le very close too. Next best are the Procrustes distance, the Procrustes shape and the Power distance based tests. The Cholesky distance and finally the Euclidean distances are the least powerful here. Note that the simulated data were multivariate normal so it is not surprising that Box's M test is the most powerful here. The test based on the Riemannian distance gives almost the same results in our simulations as that based on Box's M test, and an explanation is given in the following.

4.3. Approximate linear relationship

Box's M test and the Riemannian distance test provide very similar p-values in our simulation studies, and we now show that the two statistics are approximately linearly related. We study the case with two groups of equal sample size. Consider the singular value decomposition $\mathbf{S}_1^{-1}\mathbf{S}_2 = \mathbf{UDV}^T$, where $\mathbf{D}$ is diagonal with strictly positive singular values $s_1, \ldots, s_p$ and $\mathbf{U}, \mathbf{V}$ are orthogonal matrices. The singular values of $\mathbf{S}_2^{-1}\mathbf{S}_1$ are s_i^{-1}, $i = 1, \ldots, p$, as we can write $\mathbf{S}_2^{-1}\mathbf{S}_1 = \mathbf{VD}^{-1}\mathbf{U}^T$. Box's M statistic is given by $M = \gamma \sum_{j=1}^{g}(n_j - 1)\log|\mathbf{S}_{uj}^{-1}\mathbf{S}_u|$ where $\gamma = 1 - \frac{2p^2+3p-1}{6(p+1)(g-1)}(\sum \frac{1}{n_j-1} - \frac{1}{n-g})$, $\mathbf{S}_{uj} = \frac{n_j}{n_j-1}\mathbf{S}_j$, and $\mathbf{S}_u = \frac{n}{n-g}(n_1\mathbf{S}_1 + \cdots + n_g\mathbf{S}_g)$ where $n = \sum n_j$. If there are $g = 2$ groups with n_j equal, then $M \propto \sum_{j=1}^{2} \log|\frac{1}{2}\mathbf{S}_j^{-1}(\mathbf{S}_1 + \mathbf{S}_2)|$. This is the same as writing

$$\begin{aligned} M &\propto \log|\frac{1}{2}\mathbf{I} + \frac{1}{2}\mathbf{S}_1^{-1}\mathbf{S}_2| + \log|\frac{1}{2}\mathbf{I} + \frac{1}{2}\mathbf{S}_2^{-1}\mathbf{S}_1| \\ &= \sum_{i=1}^{p} \log(1 + (s_i - 1)/2) + \log(1 + (s_i^{-1} - 1)/2) + O((1 - s_i)^2) + O((1 - s_i^{-1})^2) \\ &= \sum_{i=1}^{p} (s_i - 1)/2 + (s_i^{-1} - 1)/2 + O((1 - s_i)^2) + O((1 - s_i^{-1})^2) \\ &= \sum_{i=1}^{p} Q_i + O((1 - s_i)^2) + O((1 - s_i^{-1})^2) =: T_1 \end{aligned} \tag{6}$$

where $Q_i = \frac{1}{2}s_i + \frac{1}{2}s_i^{-1} - 1$, and we use $\log(1 + x) = x + O(x^2)$. The square of the Riemannian distance

$$d_R^2 = \sum_{i=1}^{p} \{\log(s_i)\}^2 =: T_2 \tag{7}$$

Proposition 1
If $s_i \approx 1$, then $T_2 = 2T_1 + p + \sum_i O((1 - s_i)^2) + \sum_i O((1 - s_i^{-1})^2)$.

Hence, if s_i are close to 1, with $g = 2$ groups and $n_1 = n_2$, then the Box's M statistic and the square of the Riemannian distance are linearly related to the first order terms, and permutation tests based on M and d_R will be very similar.

Proof
Using $\log(\frac{1}{x}) = -\log(x)$, T_2 can be expressed as,

$$\begin{aligned} T_2 &= \sum_{i=1}^{p} \{\log(s_i)\}^2 = \sum_{i=1}^{p} \frac{1}{2}\{\log(s_i)\}^2 + \frac{1}{2}\{\log(s_i^{-1})\}^2 \\ &= \sum_{i=1}^{p} \frac{1}{2}(1 + (s_i - 1) + O((s_i - 1)^2))^2 + \frac{1}{2}(1 + (s_i^{-1} - 1) + O((s_i^{-1} - 1)^2))^2 \\ &= \sum_{i=1}^{p} \frac{1}{2}(1 + 2(s_i - 1) + O((s_i - 1)^2) + \frac{1}{2}(1 + 2(s_i^{-1} - 1) + O((s_i^{-1} - 1)^2) \\ &= \sum_{i=1}^{p} s_i + s_i^{-1} - 1 + O((s_i - 1)^2) + O((s_i^{-1} - 1)^2 = \sum_{i=1}^{p} 2Q_i + 1 + O((s_i - 1)^2) + O((s_i^{-1} - 1)^2) \\ &= 2T_1 + p + O((s_i - 1)^2) + O((s_i^{-1} - 1)^2) \end{aligned} \tag{8}$$

as required. □

5. Application to DNA dataset

We now apply the proposed covariance matrix test procedure to the real DNA dataset of configurations described in Section 1. We first register the dataset for each individual molecule using GPA; choose an AR(1) model based on ACF and PACF plots; and then fit an $\mathtt{AR}(1)$ model to the first principal component scores of the vector tangent coordinates. The estimated AR coefficients are large, ranging from 0.9675 for AGC to 0.9999 for AFA. Based on these coefficients, we apply the test procedure with pre-whitening for comparing the covariance matrices between the pair of damaged and undamaged molecules.

Table 3 displays the p-values of the permutation tests under different distance metrics. These results are obtained based on 300 permutations and a block size of 100 observations. The covariance structures for the AFC-AGC, AFG-AGG, TFA-TGA, and TFC-TGC molecules pairs show differences that are statistically significant at the 0.05 level. The covariance structures for the AFA-AGA and TFT-TGT molecule pairs do not show differences. There is also a consistency between the permutation tests based on the different distances, with the results for the Riemannian metric and Box's M test being identical, as expected due to the approximate linear relationship.

To date, there seems to be no significant experimental study on how the rate of repair of this type of DNA damage depends on the sequence context in which it is found. This work suggests that damage to guanine bases in AGA and TGT contexts may be particularly hard for repair systems to detect (AGA in mean and covariance, TGT in covariance structure). Hence these observations are of particular biological significance.

6. Discussion

Some limitations of the current version of the test suggest interesting future work. We have used a factored model throughout in order to have a parsimonious model, which is sensible given the roles of landmarks being recorded in three dimensions. Other covariance structures, such as full covariance matrices, sparse covariance/inverse covariance matrices, or sparse models based on a low number of principal components could also be explored using the same types of non-Euclidean distances and non-parametric tests. Also, implementation could be quicker by directly computing the time series residuals of each individual tangent vector component.

The test does not indicate the source of the differences between covariance matrices. For example, it might be interesting to tell if the difference is due to variances or correlations. The analysis of correlation matrices would also be worth exploring for the detection of the correlation structure differences. Also, describing the patterns of differences in covariance matrices is of great interest.

Acknowledgements

The support of EPSRC grant EP/K022547/1 and Royal Society Wolfson Research Merit Award WM110140 are gratefully acknowledged.

References

Ahmad, S. I. (2010). *Diseases of Dna Repair*. Advances in experimental medicine and biology. Springer Science+Business Media.

Amaral, G. J. A., Dryden, I. L., and Wood, A. T. A. (2007). Pivotal bootstrap methods for k-sample problems in directional statistics and shape analysis. *J. Amer. Statist. Assoc.*, 102(478):695–707.

Arsigny, V., Fillard, P., Pennec, X., and Ayache, N. (2007). Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM J. Matrix Anal. Appl.*, 29(1):328–347 (electronic).

Bai, Z., Jiang, D., Yao, J.-F., and Zheng, S. (2009). Corrections to lrt on large-dimensional covariance matrix by rmt. *The Annals of Statistics*, pages 3822–3840.

Box, G. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika*, 36(3/4):317–346.

Cai, T., Liu, W., and Xia, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association*, 108(501):265–277.

Chang, J., Zhou, W., Zhou, W.-X., and Wang, L. (2017). Comparing large covariance matrices under weak conditions on the dependence structure and its application to gene clustering. *Biometrics*, 73(1):31–41.

Crespo-Hernández, C. and Arce, R. (2004). Formamidopyrimidines as major products in the low-and high-intensity uv irradiation of guanine derivatives. *Journal of Photochemistry and Photobiology B: Biology*, 73(3):167–175.

Douki, T., Martini, R., Ravanat, J., Turesky, R., and Cadet, J. (1997). Measurement of 2, 6-diamino-4-hydroxy-5-formamidopyrimidine and 8-oxo-7, 8-dihydroguanine in isolated dna exposed to gamma radiation in aqueous solution. *Carcinogenesis*, 18(12):2385–2391.

Dryden, I. L. (2017). *shapes: Statistical Shape Analysis*. R package version 1.2.

Dryden, I. L., Koloydenko, A., and Zhou, D. (2009). Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics*, pages 1102–1123.

Dryden, I. L., Kume, A., Le, H., and Wood, A. (2010). Statistical inference for functions of the covariance matrix in the stationary gaussian time-orthogonal principal components model. *Annals of the Institute of Statistical Mathematics*, 62(5):967–994.

Dryden, I. L., Kume, A., Le, H., Wood, A., and Laughton, C. (2002). Size-and-shape analysis of dna molecular dynamics simulations. In *Proceedings in Statistics of Large Datasets*, pages 23–26, University of Leeds.

Dryden, I. L. and Mardia, K. V. (2016). *Statistical Shape Analysis, with Applications in R. Second Edition.* Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Ltd., Chichester.

Dutilleul, P. (1999). The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, 64(2):105–123.

Fletcher, P. T. and Joshi, S. (2007). Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Processing*, 87(2):250–262.

Friedberg, E. C., Walker, G., and Siede, W. (1995). *DNA repair and mutagenesis*. ASM Press.

Good, P. (1994). *Permutation Tests*. Springer-Verlag, New York.

Goodall, C. (1991). Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society. Series B.*, 53(2):285–339.

Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika*, 40(1):33–51.

Hall, P. and Wilson, S. R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics*, pages 757–762.

Hotelling, H. (1931). The generalization of student's ratio. *Ann. Math. Statist.*, 2(3):360–378.

James, G. (1954). Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown. *Biometrika*, 41(1/2):19–43.

Jiranusornkul, S. and Laughton, C. A. (2008). Destabilization of dna duplexes by oxidative damage at guanine: implications for lesion recognition and repair. *Journal of The Royal Society Interface*, 5(Suppl 3):191–198.

Kendall, D. G., Barden, D., Carne, T. K., and Le, H. (1999). *Shape and Shape Theory*. Wiley Series in Probability and Statistics. Wiley.

Li, J. and Chen, S. X. (2012). Two sample tests for high-dimensional covariance matrices. *The Annals of Statistics*, 40(2):908–940.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate analysis*. Academic Press [Harcourt Brace Jovanovich Publishers], London. Probability and Mathematical Statistics: A Series of Monographs and Textbooks.

Meyer, C. (2000). *Matrix analysis and applied linear algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. With 1 CD-ROM (Windows, Macintosh and UNIX) and a solutions manual (iv+171 pp.).

Pennec, X. (2006). Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154. A preliminary appeared as INRIA RR-5093, January 2004.

Petsko, G. and Ringe, D. (2004). *Protein structure and function*. Primers in biology. NSP, New Science Press.

Pouget, J., Douki, T., Richard, M., and Cadet, J. (2000). Dna damage induced in cells by γ and uva radiation as measured by hplc/gc-ms and hplc-ec and comet assay. *Chemical Research in Toxicology*, 13(7):541–549.

Schott, J. R. (2007). A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. *Computational Statistics & Data Analysis*, 51(12):6535–6542.

Seber, G. A. F. (1984). *Multivariate observations*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York.

Srivastava, M. S. and Yanagihara, H. (2010). Testing the equality of several covariance matrices with fewer observations than the dimension. *Journal of Multivariate Analysis*, 101(6):1319–1329.

Su, J., Dryden, I. L., Klassen, E., Le, H., and Srivastava, A. (2011). Fitting optimal curves to time-indexed, noisy observations of stochastic processes on nonlinear manifolds. *Journal of Image and Vision Computing*, 30:428–442.

Timm, N. (2002). *Applied Multivariate Analysis*. Springer Texts in Statistics. Springer.

Wang, Z., Vemuri, B. C., Chen, Y., and Mareci, T. H. (2004). A constrained variational principle for direct estimation and smoothing of the diffusion tensor field from complex DWI. *Medical Imaging, IEEE Transactions on*, 23(8):930–939.

Wasserman, L. (2006). *All of nonparametric statistics*. Springer Texts in Statistics. Springer, New York.

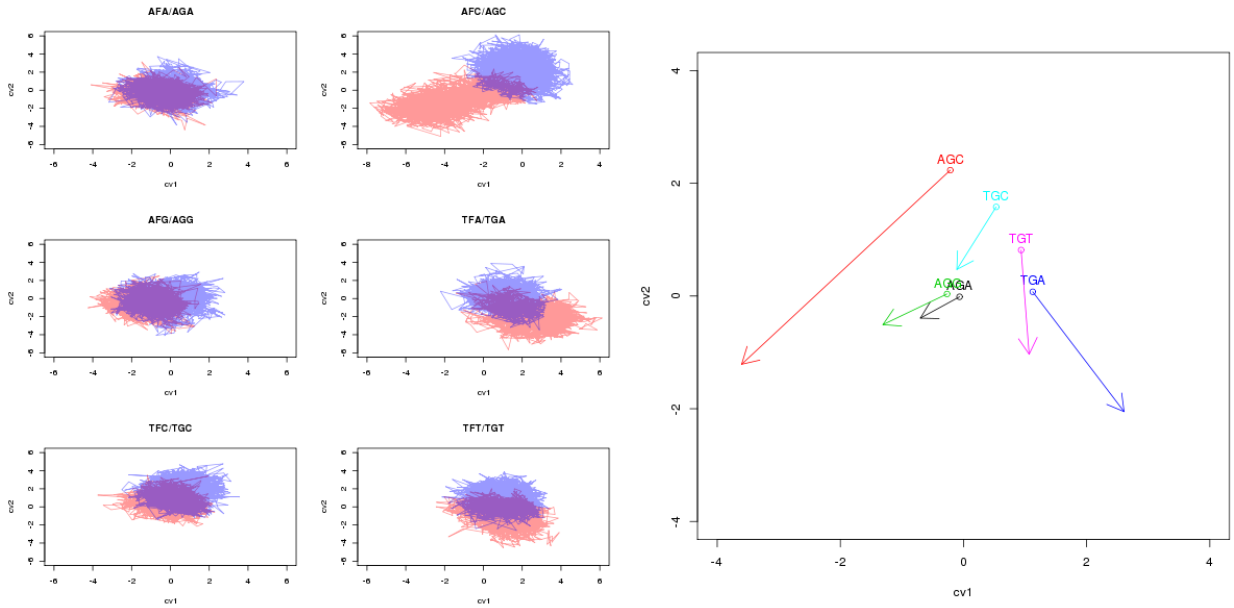

Figure 1. (left) Six pairs of DNA molecules. Plotted are the first two canonical variate scores of each DNA molecule. The undamaged molecules are indicated by XGX are shown in blue and the damaged molecules XFX are shown in red. (right) The mean CV scores for each pair of molecules. Arrows indicate the changes of the mean CV scores from an undamaged molecule to a damaged molecule.

Table 1. Mean size and shape test results of DNA molecules in Section 3.2. This table contains the results from the permutation tests for comparing the mean size and shape of the damaged-undamaged pairs of DNA molecules where the observations used in the tests were systematically selected from the DNA dataset for different sample sizes $n_1 = n_2 \in \{25, 50, 100\}$.

		p-value		
Damaged	Undamaged	$n_i = 25$	$n_i = 50$	$n_i = 100$
AFA	AGA	0.693	0.416	0.069
AFC	AGC	0.069	0.040	0.010
AFG	AGG	0.455	0.347	0.010
TFA	TGA	0.208	0.010	0.010
TFC	TGC	0.911	0.436	0.010
TFT	TGT	0.455	0.010	0.020

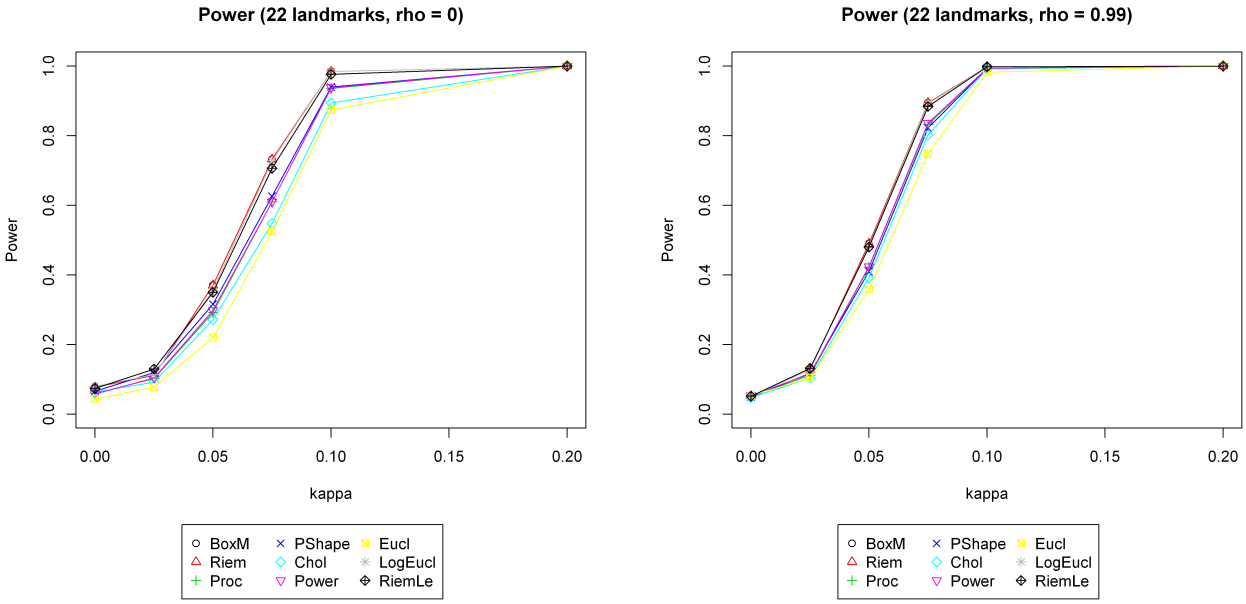

Figure 2. Pre-whitening power simulation results in Section 4.2. Plotted are the power functions against the tuning parameter κ for landmark size $k = 22$, and correlations $\rho = 0$ (left) and $\rho = 0.99$ (right).

Table 2. Pre-whitening size simulation results in Section 4.2. This table contains the simulated test size at the 5% significance level from the pre-whitening testing procedure under different specifications for $k = 22$ landmarks. Four autocorrelation levels and nine distance metrics are considered. The sample size is $n = 2500$.

ρ	BoxM	Riem	Proc	PShape	Chol	Power	Eucl	LogEucl	RiemLE
0	0.070	0.070	0.056	0.066	0.060	0.058	0.048	0.072	0.077
.5	0.048	0.048	0.046	0.053	0.045	0.043	0.044	0.053	0.056
.9	0.051	0.051	0.044	0.047	0.061	0.046	0.049	0.046	0.060
.99	0.060	0.060	0.060	0.062	0.067	0.061	0.057	0.063	0.052

Table 3. Test of size-and-shape covariance matrices of DNA molecules in Section 5. This table contains the p-values obtained from applying the permutation tests to the pre-whitened data for comparing difference in covariance matrices of damaged and undamaged DNA molecules.

Damaged	Undamaged	BoxM	Riem	Proc	PShape	Chol	Power	Eucl	LogEucl	RiemLE
AFA	AGA	0.213	0.213	0.300	0.273	0.173	0.320	0.483	0.183	0.287
AFC	AGC	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AFG	AGG	0.000	0.000	0.000	0.003	0.006	0.000	0.007	0.000	0.000
TFA	TGA	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
TFC	TGC	0.000	0.000	0.000	0.000	0.017	0.000	0.023	0.000	0.000
TFT	TGT	0.290	0.290	0.117	0.097	0.527	0.123	0.050	0.297	0.337

Prepared using staauth.cls