1 Palindromic symmetry of a shape vector and stiffness matrix

As seen in Exercise 2 of Session 5, the mean vector $\mu(S)$ an the stiffness matrix $K(S)$ of the cgDNA Gaussian distribution $\rho(\mathbf{w}, S)$, satisfy the following symmetric properties: $\mu(S) = E_n\mu(\overline{S})$, and $K(S) = E_nK(\overline{S})E_n$, where $\overline{S}$ is the complementary sequence to S and the matrix E_n is defined in Session 5. For palindromic sequences, i.e. when $S = \overline{S}$, these properties become $\mu(S) = E_n\mu(S)$ and $K(S) = E_nK(S)E_n$.

1. Load the file `muABC_S3.mat` with the DNA 18-mer data used previously in Session 10 . Construct the matrix E_n for $n = 18$, and check if the palindromic properties of the shape vector and the covariance matrix are satisfied.

2. Before using the MD data to fit a cgDNA parameter set, in order to minimize modelling errors, it is useful to first get shape and stiffness estimates satisfying palindromic symmetry conditions, if a sequence is palindromic. Symmetrize the shape estimate by computing $\bar{\mathbf{w}}_{sym} = \frac{1}{2}(\bar{\mathbf{w}} + E_n\bar{\mathbf{w}})$. Plot the difference between $\bar{\mathbf{w}}$ and $\bar{\mathbf{w}}_{sym}$.

3. The covariance matrix was obtained as $C = \frac{1}{M}\sum_{j=1}^{M}\mathbf{w}^{[j]} \otimes \mathbf{w}^{[j]} - \bar{\mathbf{w}} \otimes \bar{\mathbf{w}}$. Symmetrize $D := \frac{1}{M}\sum_{j=1}^{M}\mathbf{w}^{[j]} \otimes \mathbf{w}^{[j]}$ and then compute the symmetrized covariance matrix using D_{sym} and $\bar{\mathbf{w}}_{sym}$. Compute the symmetrized stiffness K_{sym} as the inverse of C_{sym}. Using `plot2DMatrix.m`, plot the differences between C and C_{sym} and between K and K_{sym}.

4. Download this scripts: `http://lcvmwww.epfl.ch/teaching/modelling_dna/public_files/MaxEntropy.m`. Use the code `MaxEntropy.m` to get the maximum entropy fit to the symmetrized covariance computed in the previous point.

5. Compute the following Kullback-Leibler divergences:

i) $D(\rho_{obs}^{sym}(S), \rho_{obs}(S))$,
ii) $D(\rho_{band}^{sym}(S), \rho_{obs}^{sym}(S))$,
iii) $D(\rho_{cgDNA}(S), \rho_{band}^{sym}(S))$.

Compare this error with the modelling errors (Kullback-Leibler divergences computed in Exercise 1 of Session 11). Is the convergence error big or small compared to the modelling errors?

2 Gaussian Integral III

Let $\hat{\mathbf{x}} \in \mathbb{R}^n$ and a symmetric, positive - definite matrix $K = \Sigma^{-1} \in \mathbb{R}^{n\times n}$ be given. Show that a conditional of a Gaussian distribution is also a Gaussian distribution: if $\mathbf{x} \sim N(\hat{\mathbf{x}}, \Sigma)$,

$$\mathbf{x} = \begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix}, \quad \hat{\mathbf{x}} = \begin{bmatrix}\hat{\mathbf{x}}_1\\ \hat{\mathbf{x}}_2\end{bmatrix}, \quad \mathbf{x}_1, \hat{\mathbf{x}}_1 \in \mathbb{R}^k, \quad \mathbf{x}_2, \hat{\mathbf{x}}_2 \in \mathbb{R}^m,$$

$$\Sigma = \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{12}^T & \Sigma_{22}\end{bmatrix}, \quad \Sigma_{11} = \Sigma_{11}^T \in \mathbb{R}^{k\times k}, \quad \Sigma_{12} \in \mathbb{R}^{k\times m},$$

$$\Sigma_{22} = \Sigma_{22}^T \in \mathbb{R}^{m\times m}, \quad \text{and} \quad k + m = n,$$

then $(\mathbf{x}_1|\mathbf{x}_2 = \mathbf{a}) \sim N(\overline{\mathbf{x}}, \overline{\Sigma})$, where

$$\overline{\mathbf{x}} = \hat{\mathbf{x}}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{a} - \hat{\mathbf{x}}_2), \quad (1)$$

$$\overline{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \quad (2)$$

[Hint: Use the definition for the conditional density function for continuous random variable and the solution of Exercise 1 of Session 9.]

3 On the computation of conditionals of the cgDNA probability distribution

For this exercise you should download the following scripts: `http://lcvmwww.epfl.ch/teaching/modelling_dna/public_files/Conditional_scripts.zip`.
The aim of this exercise is to develop a basic statistical model for modelling the interaction between DNA and proteins. DNA-binding proteins are proteins which have an affinity with DNA (see for more details the Wikipedia article : DNA-binding protein), which can bind to the DNA in either the major or minor groove. In the context of the cgDNA model one can see the protein that binds to a molecule of DNA as a constrains on some of the internal coordinates describing the DNA segment. Thus, from a statistical mechanics point of view the interaction between DNA and protein can be modelled as a conditional distribution of the density function related to the DNA fragment, which in the cgDNA land is Gaussian. Thanks to the previous exercise we know that a conditional distribution of a Gaussian distribution still be a Gaussian. Let us assume that the interaction DNA-protein is reduced to a change in only an intra coordinate. Let $\mathbf{w} = (y_1, x_1, y_2, \ldots, x_i, y_{i+1}, x_{i+1}, \ldots, y_n) = (w_1, y_i, w_2) \in \mathbb{R}^{12n-6}$ where $y_j \in \mathbb{R}^6$ are the intras and $x_k \in \mathbb{R}^6$ are the inters, and

$$\begin{aligned}\rho(\mathbf{w}; S, \mathcal{P}) &= \frac{1}{Z}\exp\left\{-\frac{1}{2}(\mathbf{w} - \hat{\mathbf{w}}(S,\mathcal{P}))\cdot K(S,\mathcal{P})(\mathbf{w} - \hat{\mathbf{w}}(S,\mathcal{P}))\right\} \\ &= \frac{1}{Z}\exp\left\{-\frac{1}{2}\begin{bmatrix} w_1 - \hat{w}_1 \\ y_i - \hat{y}_i \\ w_2 - \hat{w}_2 \end{bmatrix}\begin{bmatrix} A & B & 0 \\ B^T & C & D^T \\ 0 & D & E \end{bmatrix}\begin{bmatrix} w_1 - \hat{w}_1 \\ y_i - \hat{y}_i \\ w_2 - \hat{w}_2 \end{bmatrix}\right\}, \end{aligned} \quad (3)$$

a cgDNA Gaussian for the sequence S and the parameter set $\mathcal{P}$. Imagine now that a protein is binding to the i−th basepair, thus it constraints $y_i = \mathbf{a} \in \mathbb{R}^6$.

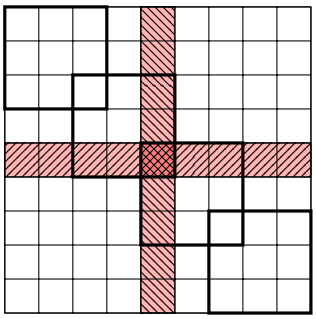
1. By using the Exercise 2 of this sheet find the conditional mean $\overline{\mathbf{w}} = (\overline{w}_1, \overline{w}_2)$.

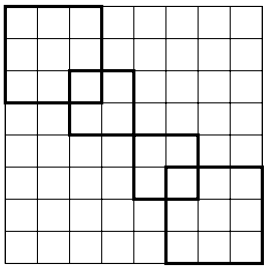
2. Complete the lines 53 and 54 in `calc_conditional_shapes.m` with your findings. and run it with the following input arguments :

   - sequence : ATCGCGAATGCGAGCCTGTA ;
   - cond_index : 10 ;
   - cond : [0.1 0.5 0.5 0.3 0.6 0] (= $\delta\mathbf{a}$) . In the following code we consider $\mathbf{a} = \hat{y}_i + \delta\mathbf{a}$.

   Be aware that you have also to complete lines 19 and 20 by adding to the path your cgDNA folder and your cgDNAviewer.

[Hint: Conditioning one intra (example on the left) leads to a specific decomposition of the stiffness matrix that can be seen in the matrix on the left (each little block is a 6 times 6 matrix). This implies that the conditional stiffness can be seen as a block diagonal matrix.]





4 Square roots in SE(3)

In the cgDNA model we have defined the basepair frame as the mid frame between two base frames by using the usual square root for the rotation part and the euclidean average for the $\mathbb{R}^3$ part. The latter way of defining the mid frame is also used to define the junction frames between two consecutive basepair frames. In this exercise we will study the difference between the cgDNA way of defining the mid frames and the square root in SE(3) which could be used as alternative way for the representation of the a mid frame. Let $G \in SE(3)$, then

$$G = \begin{bmatrix} Q & q \\ 0_3^T & 1 \end{bmatrix}, \tag{4}$$

where $Q \in SO(3)$, $q \in \mathbb{R}^3$. Let $B \in SE(3)$ such that

$$B = \begin{bmatrix} Q^{\frac{1}{2}} & x \\ 0_3^T & 1 \end{bmatrix}. \tag{5}$$

Find x such that $BB = G$. What can you say about B.

5 Statistical physics for ideal polymer chains (Optional)

Before doing the exercise please download and read the following document: `http://lcvmwww.epfl.ch/teaching/modelling_dna/public_files/Lecture_NotesLp.pdf`.

1) Let a stochastic chain made of N bond denoted $lt_n \in \mathbb{R}^3$ with $l > 0$ a constant bond length and $t_n \in \mathbb{R}^3$, with $\|t_n\| = 1 \ \forall n$, the bond direction. The positions in the chain, $r_n \in \mathbb{R}^3$, are given by the recursion relation $r_{n+1} = r_n + lt_n$, r_0 fixed. Define now the bond correlation by

$$c_{m,n} = \langle t_m \cdot t_n \rangle, \tag{6}$$

where $\langle \cdot \rangle$ stands for the expectation over a given ensemble of configurations.

   i) Deduce the expression for the values $\langle (r_n - r_0) \cdot t_0 \rangle$ and $\langle \|r_n - r_0\|^2 \rangle$ as a function of the values $c_{m,n}$.

   ii) Simplify these expression for the isotropic random–ϕ model presented in the notes, i.e,

$$c_{m,n} = \alpha^{|m-n|}, \ \alpha = \langle \cos\theta \rangle \tag{7}$$

[Hint: Use that, $\sum_{k=0}^{n-1} \sum_{l=0}^{k-1} \alpha^{k-l} = n\frac{\alpha - \alpha^n}{1-\alpha} - \alpha\frac{1-\alpha^n}{1-\alpha} + n^2\frac{\alpha^n}{1-\alpha}$]

2) Consider a chain $r_n \in \mathbb{R}^2$ in the plane. Suppose that the distribution of $(t_0, \ldots, t_N)$ is such that the most probable angle between t_n and t_{n+1}, $\forall n$, is $\hat{\theta}$ and the probability of θ is symmetric around $\hat{\theta}$, i.e, the probability of $\theta - \hat{\theta}$ is equal to $-(\theta - \hat{\theta})$. Show that

$$\langle t_{n+1} \cdot t_n \rangle = (\hat{t}_{n+1} \cdot \hat{t}_n) \langle \cos(\theta - \hat{\theta}) \rangle, \tag{8}$$

where $\hat{t}_{n+1} \cdot \hat{t}_n = \cos(\hat{\theta})$. The angle $\hat{\theta}$ is called static contribution while $\theta - \hat{\theta}$ is the dynamic contribution.

Exercise 2.4 iii) of Session 7 showed that this factorisation is also sensible in three dimensions and in the context of the cgDNA model.