Mathematical Modelling of DNA
Prof. John Maddocks
Session 10 A. Patelli

For the exercise 2 please download the following dataset `http://lcvmwww.epfl.ch/teaching/modelling_dna/public_files/muABC_S3.mat` . The dataset consist in an oligomer-based statistics of a 18 basepairs long Palindrome. The MATLAB structure contains the following fields:

- `seq` : sequence,
- `nbp` : number of base pair,
- `nsnap` : total number of accepted snapshots from the MD simulation (= M).
- `shape` : ensemble mean ($\bar{\mathbf{w}} = \frac{1}{M}\sum_{j=1}^{M}\mathbf{w}^{[j]}$),
- `c1b` : ensemble covariance, ($C = \frac{1}{M}\sum_{j=1}^{M}(\mathbf{w}^{[j]} - \bar{\mathbf{w}}) \otimes (\mathbf{w}^{[j]} - \bar{\mathbf{w}})$)
- `stiff_me` : maximum entropy fit to `c1b`.

1 Relative entropy for Gaussians II

In the session 9, exercise 2.2, we showed the following formula for the relative entropy between two Gaussian density functions:

$$D(p,q) = \frac{1}{2}\left[\operatorname{tr}\left(K_2K_1^{-1}\right) - \ln\frac{\det K_2}{\det K_1} - N\right] + \frac{1}{2}(\widehat{x}_1 - \widehat{x}_2)\cdot K_2(\widehat{x}_1 - \widehat{x}_2), \tag{1}$$

where

$$\begin{aligned} p(x) &= \frac{1}{(2\pi)^{N/2}|K_1^{-1}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(x-\widehat{x}_1)\cdot K_1(x-\widehat{x}_1)\right\},\ x\in\mathbb{R}^N,\\ q(x) &= \frac{1}{(2\pi)^{N/2}|K_2^{-1}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(x-\widehat{x}_2)\cdot K_2(x-\widehat{x}_2)\right\},\ x\in\mathbb{R}^N, \end{aligned}$$

1. Let $M \in \mathbb{R}^{N\times N}$ and X_1 and X_2 the normal random variables with respectively probability density functions p and q. Define the change of variable $Y_i = MX_i$, $i = 1,2$.

i) What are the distributions of Y_1 and Y_2? Give explicitly the mean value and the covariance for each random variable.

ii) By denoting by $\tilde{p}$ and $\tilde{q}$ the density functions of, respectively, Y_1 and Y_2, show that $D(\tilde{p},\tilde{q}) = D(p,q)$.

2. In the Exercise 2.2 of the exercise session 9 we have also introduced:

$$D^{\dagger}(K_1,K_2) := \frac{1}{2}\left[\operatorname{tr}\left(K_2K_1^{-1}\right) - \ln\frac{\det K_2}{\det K_1} - N\right], \tag{2}$$

for K_i, symmetric positive defined matrices, $i = 1,2$. Define now the following generalised eigenvalue problem

$$K_2v_i = \mu_iK_1v_i,\ i = 1,\dots,N. \tag{3}$$

Show that using the generalised eigenvalue problem (3) the equation (2) becomes

$$D^{\dagger}(K_1,K_2) = \frac{1}{2}\sum_{i=1}^{N}(\mu_i - \ln\mu_i - 1). \tag{4}$$

3. The stiffness part of the relative entropy (2) is in general not symmetric, i.e, $D\dagger(K_1, K_2) \neq D^\dagger(K_2, K_1)$ for $K_1 \neq K_2$. We then define the symmetrized version of (2) in the following way:

$$D^\dagger_{sym}(K_1, K_2) \coloneqq \frac{1}{2}\left(D^\dagger(K_1, K_2) + D^\dagger(K_2, K_1)\right). \tag{5}$$

i) Find an explicit form for $D^\dagger_{sym}(K_1, K_2)$.

ii) Using the generalized eigenvalue problem (3) find the corresponding eigenvalue version form for $D^\dagger_{sym}(K_1, K_2)$.

[Note: Equation (4) can be useful to prove part 1, ii)]

2 Banded matrices and their inverses

Consider the following symmetrically partitioned matrix:

$$C = \begin{bmatrix} a & e & \mathrm{x} \\ e^T & b & d \\ \mathrm{x}^T & d^T & c \end{bmatrix},\ a = a^T,\ b = b^T,\ c = c^T, \tag{6}$$

where for simplicity we assume $C > 0$. We want to show that if $\mathrm{x} = eb^{-1}d$, then $K \coloneqq C^{-1}$ has zeros blocks in the $(1,3)$ and $(3,1)$ entries. For that prove the following statements:

1. Show that the matrix C can be decomposed as follow

$$C = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & \Psi & I \end{bmatrix}\begin{bmatrix} a & e & 0 \\ e^T & b & 0 \\ 0 & 0 & H \end{bmatrix}\begin{bmatrix} I & 0 & 0 \\ 0 & I & \Omega \\ 0 & 0 & I \end{bmatrix}, \tag{7}$$

where $\Omega = b^{-1}d$, $\Psi = d^T b^{-1}$, and $H = c - d^T b^{-1} d$.

2. Using the decomposition obtained on the previous point, compute the inverse of C.

3. Conclude that the blocks in the $(1,3)$ and $(3,1)$ entries of $K = C^{-1}$ are zero.

4. From the latter point derive an algorithm that will allow you to compute the matrix K that involve only the blocks a, b, c, d and e (but not x). Code your algorithm and test it on the two first blocks of the `c1b` matrix.

5. Can you capture the five blocks a, b, c, d and e of C from the five non zero blocks of K more simply than just inverting K?

This result, as well as the algorithm, can be generalized to matrices with multiple blocks and overlap without assuming equal dimension of the blocks and the overlaps. You can try by yourself to prove the general result using an induction argument, or you can just generalize your algorithm and test it on the `c1b` matrix in the dataset.

3 Kullback-Liebler divergence between : $\rho_{obs}(S)$, $\rho_{band}(S)$, $\rho_{cgDNA}(S, \mathcal{P})$

In this exercise we want to study the Kullback-Leibler divergence between the different steps of the approximation of the stiffness and the mean for a given sequence. We will use the following notation: $\rho_{obs}(S)$ is the Gaussian distribution which parameters are the ensemble mean and the ensemble covariance, $\rho_{band}(S)$ is the Maximum Entropy Gaussian distribution fitted to the ensemble mean and covariance, and $\rho_{cgDNA}(S, \mathcal{P})$ is the cgDNA reconstruction of S using the parameter set $\mathcal{P}$ (check that you are using the cgDNAparamset2).

1. Write a MATLAB script implementing the Kullback-Leibler divergence for Gaussian distributions (see Session 9, exercise 2.2). Allow your script to output the value of the Kullback-Leibler divergence, the value of the stiffness part, and the value of the mean part (Mahalanobis distance).

2. Using your script do the following computations:

i) $D(\rho_{band}(S), \rho_{obs}(S))$,

ii) $D(\rho_{cgDNA}(S, \mathcal{P}), \rho_{band}(S))$,

Compare the values of the Kullback-Leibler divergence as well as the values coming from the stiffness part and the mean part. What can you say?