1 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 $\mathbf{x} \sim N(\hat{\mathbf{x}}, \Sigma)$,

$$\begin{aligned}
\mathbf{x} &= \begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix}, \quad \hat{\mathbf{x}} = \begin{bmatrix}\hat{\mathbf{x}}_1\\ \hat{\mathbf{x}}_2\end{bmatrix}, \quad \mathbf{x}_1, \hat{\mathbf{x}}_1 \in \mathbb{R}^k, \quad \mathbf{x}_2, \hat{\mathbf{x}}_2 \in \mathbb{R}^m,\\
\Sigma &= \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{12}^T & \Sigma_{22}\end{bmatrix}, \quad \Sigma_{11} = \Sigma_{11}^T \in \mathbb{R}^{k\times k}, \quad \Sigma_{12} \in \mathbb{R}^{k\times m}, \Sigma_{22} = \Sigma_{22}^T \in \mathbb{R}^{m\times m},\\
K &= \begin{bmatrix}K_{11} & K_{12}\\ K_{12}^T & K_{22}\end{bmatrix}, \quad K_{11} = K_{11}^T \in \mathbb{R}^{k\times k}, \quad K_{12} \in \mathbb{R}^{k\times m}, K_{22} = K_{22}^T \in \mathbb{R}^{m\times m}, \quad \text{and} \quad k+m=n,
\end{aligned}$$

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

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

$$\overline{\Sigma}^{-1} = K_{11}. \tag{2}$$

Remark: For Gaussians: marginal shift is a sub-vector of original mean and stiffness is the inverse of a sub-block of the original covariance. For conditional, stiffness is a sub-block of original stiffness but shift requires a computation.

[Hint: Use the definition for Gaussian marginal density function and the solution of Qu 2 serie 13.]

2 On the computation of conditionals of the cgDNA+ probability distribution

For this exercise, first download (`http://lcvmwww.epfl.ch/teaching/modelling_dna/protected_files/codes_exercises/calc_conditional_shapes.m`) matlab scripts required for computing conditionals of cgDNA+ probability distribution. The aim of this exercise is to explain 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 model a protein that binds to a molecule of DNA as constraints 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 cgDNA+ land is Gaussian. Thanks to the previous exercise we know that a conditional distribution of a Gaussian distribution is still a Gaussian. Let us assume that the interaction of DNA-protein is reduced to a change in only the k^{th} Watson phosphate coordinate, where k represent the base-pair number/index in a given sequence. Let $\mathbf{w} = (y_1, y_1^{pC}, x_1, y_2^{pW}, y_2, y_2^{pC}, x_2, \ldots, x_{n-1}, y_n^{pW}, y_n) = (w_1, y_k^{pW}) \in \mathbb{R}^{24n-18}$ where $y_i \in \mathbb{R}^6$ are the intras, $x_i \in \mathbb{R}^6$ are the inters, and $y_i^{pC} \in \mathbb{R}^6, y_i^{pW} \in \mathbb{R}^6$ are Watson and Crick phosphates coordinates

respectively and $w_1 \in \mathbb{R}^{24n-24}$ are rest (except y_k^{pW}) of the coordinates for a sequence having n base-pairs and

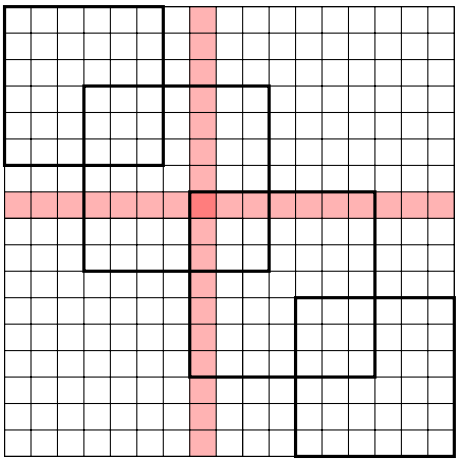
$$\begin{aligned}\rho(\mathbf{w}; S, \mathcal{P}) &= \frac{1}{Z}\exp\left\{-\frac{1}{2}(\mathbf{w}-\hat{\mathbf{w}}(S,\mathcal{P}))\cdot K(S,\mathcal{P})(\mathbf{w}-\hat{\mathbf{w}}(S,\mathcal{P}))\right\} \\ &= \frac{1}{Z}\exp\left\{-\frac{1}{2}(\mathbf{w}-\hat{\mathbf{w}}(S,\mathcal{P}))\cdot \Sigma^{-1}(S,\mathcal{P})(\mathbf{w}-\hat{\mathbf{w}}(S,\mathcal{P}))\right\} \\ &= \frac{1}{Z}\exp\left\{-\frac{1}{2}\begin{bmatrix} w_1-\hat{w_1} \\ y_k^{pW}-\hat{y}_k^{pW}\end{bmatrix}\begin{bmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22}\end{bmatrix}^{-1}\begin{bmatrix} w_1-\hat{w_1} \\ y_k^{pW}-\hat{y}_k^{pW}\end{bmatrix}\right\}, \end{aligned} \tag{3}$$

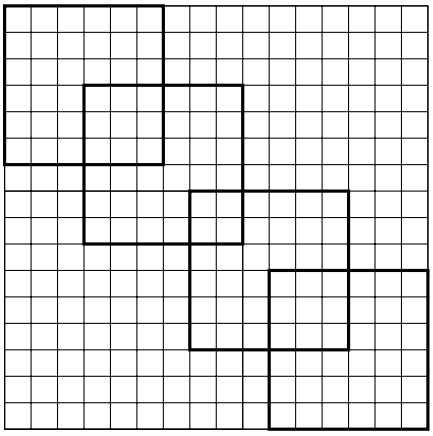
a cgDNA+ Gaussian for the sequence S and the parameter set $\mathcal{P}$. Here $K = \Sigma^{-1}$, block Σ_{22} is 6×6 matrix corresponding to the k^{tk} Watson phosphate coordinates and Σ_{11} is rest of the block and Σ_{12} represent coupling between k^{tk} Watson phosphates and rest of the blocks. Imagine now that a protein is binding to the k^{th} Watson phosphate, thus it constraints $y_k^{pW} = \mathbf{a} \in \mathbb{R}^6$.

1. By using the Exercise 1 of this serie find the conditional mean $(\overline{w}_1)$
2. Complete the lines 52 in matlab script `calc_conditional_shapes.m` with your findings $(\overline{w}_1)$ and run it with the following input arguments :
   - sequence (S) : ATCGCGAATGCGAGCCTGTA ;
   - cond_index : 10 ;
   - condition : [0.1 0.5 0.5 0.3 0.6 0] $(= \delta\mathbf{a})$. In the following code we consider $\mathbf{a} = \hat{y}_k^{pW} + \delta\mathbf{a}$.

   Be aware that you have add the path of cgDNA+ folder and cgDNAp2dplot while using the script `calc_conditional_shapes.m`.
3. Plot the cgDNA+ groundstate and computed conditional groundstate in one plot. What can you say?

Note: Conditioning one Watson phosphate (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 (matrix in the right) can be seen as an overlapping block diagonal matrix.





3 On the average of rotation matrices sharing a common (deterministic) axis (provides a proof of a result used in computation of persistence length in Chapter 2)

Let $\mathcal{Q} = \{Q_k\}_{k=1}^N \subset \mathrm{SO}(3)$ an ensemble of rotation matrices sharing a common axis of rotation denoted by $\mathbf{u}$. Show that $\|\langle Q\rangle\| = 1$, where $\langle Q\rangle := \frac{1}{N}\sum_{k=1}^N Q_k$ and $\|A\| = \sup_{\|\mathbf{x}\|=1}\|Ax\|$. Moreover show that if at least one rotation matrix in $\mathcal{Q}$ has a different rotation axis, then $\|\langle Q\rangle\| < 1$.

4 On the parametrization of junction displacement using quaternions

Nowadays it is rather fast to sample from multivariate distributions particularly ones with banded stiffness matrix because Cholesky decomposition is also banded. Thus the slowest portion in a Monte Carlo code can be the evaluation of the chosen deterministic function. In the cgDNApmc code we have made two different choice of functions of the cgDNA+ coordinates:

1. $\phi(\mathbf{x}) = (R_1^T R_n)_{(3,3)}$,
2. $\phi(\mathbf{x}) = R_1^T(r_n - r_1)$,

where $(A)_{(3,3)}$ means the (3,3) entry of a matrix $A \in \mathbb{R}^{3\times 3}$, (R_1, r_1) is a fixed base-pair frame chosen to be the first (but not necessarily the first one of the DNA fragment), and (R_n, r_n) is the nth base-pair frame after the fixed one. For both of the above functions, in cgDNApmc we have to perform many matrix multiplications in $SO(3)$ in order to be able to evaluate the functions for each sampled configuration. For efficiency, in the cgDNApmc code, these multiplications are implemented using quaternion multiplication. We have already seen how to parametrize a rotation matrix using three numbers or the Cayley vectors. In this exercise we will study the parametrization of a rotation matrix by four numbers called *Euler-Rodrigues parameters* or *quaternions*.

Any vector $q = (q_0, q_1, q_2, q_3) \in \mathbb{S}^3 = \{x \in \mathbb{R}^4 | x \cdot x = 1\}$ can be interpreted as a right-handed rotation in $\mathbb{R}^3$ through an angle θ and around a unit axis $\mathbf{w} \in \mathbb{R}^3$, where θ and $\mathbf{w}$ solve :

$$\cos\frac{\theta}{2} = q_0, \quad \text{and} \quad \mathbf{w}\sin\frac{\theta}{2} = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix} = \mathbf{q}. \tag{4}$$

1. Let $Q \in \mathrm{SO}(3)$ a rotation matrix about a unit axis $\mathbf{w}$ through an angle $0 \le \theta < \pi$. Let $u = Cay(Q) \in \mathbb{R}^3$ be the Cayley parametrisation of Q. Find the quaternion parametrisation of Q in term of the Cayley vector u. [Hint: We recall that $\|u\| = \tan\frac{\theta}{2}$].
2. Using the previous part, show that the Euler-Rodrigues formula (2) of exercise 1.2 session 3 implies the following quaternion parametrisation:

$$Q(q) = \begin{bmatrix} q_1^2 - q_2^2 - q_3^2 + q_0^2 & 2(q_1q_2 - q_3q_0) & 2(q_1q_3 + q_2q_0) \\ 2(q_1q_2 + q_3q_0) & -q_1^2 + q_2^2 - q_3^2 + q_0^2 & 2(-q_1q_0 + q_2q_3) \\ 2(q_1q_3 - q_2q_0) & 2(q_1q_0 + q_2q_3) & -q_1^2 - q_2^2 + q_3^2 + q_0^2 \end{bmatrix} \tag{5}$$

3. From a computational point of view, the interest of using quaternions instead of rotation matrices lies in the following equivalence: Let $Q_i = Q(q_i) \in \mathrm{SO}(3)$, for $i = 1, 2, 3$, we have

$$Q_3 = Q_1Q_2 \Longleftrightarrow q_3 = q_1 \circ q_2, \tag{6}$$

where the symbol $\circ$ mean the multiplication operator for quaternions. For two quaternions $q = (q_0, \mathbf{q})$ and $p = (p_0, \mathbf{p})$ we define

$$q \circ p = \left(q_0 p_0 - \mathbf{q} \cdot \mathbf{p}, q_0 \mathbf{p} + p_0 \mathbf{q} + \mathbf{q} \times \mathbf{p}\right). \tag{7}$$

We could derive (7) and prove the equivalence (6) but here we will just check them numerically. Infact, using (7) is significantly faster than multiplying the matrices. For that purpose use the cgDNA+ package to reconstruct the groundstate of a short fragment of DNA (10-12 base-pair). Then, check the following

$$R_3 = R_2 Q_2 \Longleftrightarrow q^{R_3} = q^{R_2} \circ q^{Q_2}, \tag{8}$$

where R_i is the orientation of the ith base-pair, and Q_2 is the rotational part of the second junction displacement, and q^{M_i} is the quaternion related to the rotation matrix M_i.

Remark: In the cgDNA+ model the Cayley vectors are scaled in a way that their norm equal $10 \tan \frac{\theta}{2}$, where θ is the angle of the rotation. Thus, you have to rescale the cgDNA+ Cayley vectors such that their norm are $\tan \frac{\theta}{2}$ if you want to use the relation between Cayley vectors and quaternions.