1 Principle of maximum entropy parameter estimation for banded stiffness matrices

Denote by $[[K]]_{\mathcal{N}}$ all the entries $(i,j) \in \mathcal{N}$ of K where $\mathcal{N}$ is a set of indices. For this exercise we will fix $\mathcal{N}$ to be the set of all indices associated to the cgDNA+ stiffness matrix (ie 42×42 block diagonal pattern with 18×18 overlaps in the interior of the sequence and 36×36 block diagonal pattern with 18×18 overlaps for ends). For sake of notation we will omit the $\mathcal{N}$, i.e, $[[K]]_{\mathcal{N}} = [[K]]$, for all $K \in \mathbb{R}^{24n-18 \times 24n-18}$, with $n \in \mathbb{N}$. Moreover with $[[\cdot]]^c$ we denote all the entries $(l,k) \in \mathcal{N}^c$, where $\mathcal{N}^c$ is the complement of $\mathcal{N}$.

Given $\mu \in \mathbb{R}^{24n-18}$ and $C \in \mathbb{R}^{24n-18 \times 24n-18}$ (the observed statistics, mean and covariance, of a n base–pair long molecule of DNA) define the following constraint set:

$$C = \left\{ \rho : \int_\Omega \rho dx = 1,\ \int_\Omega x_k \rho(x) dx = \mu_k,\ k = 1, \dots, 24n-18,\ \int_\Omega x_i x_j \rho(x) dx = c_{ij},\ (i,j) \in \mathcal{N} \right\}. \quad (1)$$

where $\Omega = \mathbb{R}^{24n-18}$. Using the principle of maximum entropy, prove that the maximum entropy distribution is a Gaussian, i.e, it can be written as

$$\rho_{ME}(x) = \frac{1}{Z(\mu, K)} \exp\left\{ -\frac{1}{2}(x-\mu) \cdot K_{ME}(x-\mu) \right\}, \quad (2)$$

where μ is the observed mean and K_{ME} is such that $[[K^{-1}]] = [[C]]$, and $[[K]]^c = 0$.

[Remark: From the exercise 1 & 2 of session 10 we know how to compute the matrix K_{ME} directly from the data $[[C]]$.]

2 Gaussian integrals II (Part I in Session 1)

Let $\beta > 0$, $n \geq 0$ and a symmetric, positive definite matrix $K = \Sigma^{-1} \in \mathbb{R}^{n \times n}$ be given. Show that a marginal of a Gaussian distribution is also a Gaussian distribution: if $\mathbf{x} \sim N(\hat{\mathbf{x}}, \Sigma)$,

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

then $\mathbf{x}_1 \sim N(\hat{\mathbf{x}}_1, \Sigma_{11})$, i.e.

$$\frac{\left(\frac{\beta}{\pi}\right)^{\frac{n}{2}}}{\sqrt{\det[K^{-1}]}} \int_{\mathbb{R}^m} e^{-\beta(\mathbf{x}-\hat{\mathbf{x}}) \cdot K (\mathbf{x}-\hat{\mathbf{x}})} \, d\mathbf{x}_2 = \frac{\left(\frac{\beta}{\pi}\right)^{\frac{k}{2}}}{\sqrt{\det \Sigma_{11}}} \, e^{-\beta(\mathbf{x}_1-\hat{\mathbf{x}}_1) \cdot \Sigma_{11}^{-1} (\mathbf{x}_1-\hat{\mathbf{x}}_1)}.$$

[Hint: Found explicitly the inverse of a symmetric, positive definite matrix of the following form $\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix}$, with $A = A^T$ and $C = C^T$ and compute its determinant.]

3 On the computation of marginals of the cgDNA+ probability distribution

In this exercise, we want to do marginals of a cgDNA+ Gaussian distribution at two levels. First, marginals over phosphate, thus the obtained marginalised pdf will be in inter and intra coordinates (this is equivalent to cgDNA/rigid base model coordinates). Second, marginals over phosphates as well as intra coordinates, thus the obtained marginalised pdf will only be in inter coordinates (this is equivalent to rigid base-pair model).

Given a sequence S and a parameter set $\mathcal{P}$, the cgDNA+ model is the Gaussian distribution:

$$\rho(x;S,\mathcal{P}) = \frac{1}{Z}\exp\left\{-\beta(x-\mu(S,\mathcal{P}))\cdot K(S,\mathcal{P})(x-\mu(S,\mathcal{P}))\right\} \tag{3}$$

where $\mu(S,\mathcal{P})$ and $K(S,\mathcal{P})$ are respectively the mean and the stiffness matrix. We recall that the covariance is $\Sigma = K^{-1}(S,\mathcal{P})$. For the computation consider the R. E. Dickerson palindromic dodecamer S_D = CGCGAATTCGCG. Consider the covariance $\Sigma_D = K_D^{-1}$. We stress on the fact that the stiffness matrix K_D has a specific pattern and is sparse while Σ_D is dense. Write a matlab code to compute the following marginal pdfs.

3.1 Marginalise over phosphates variables

To compute the marginals of a Gaussian pdf over some variables, remove the rows and columns corresponding to those variables from the covariance matrix and remove the corresponding variables from the mean. So, in cgDNA+ Gaussian pdf, to compute marginals over phosphates, remove the rows and columns corresponding to phosphate variables from the covariance matrix and phosphate variables from the mean. Then the corresponding marginalised stiffness matrix would be the inverse of this marginalised covariance matrix. What can you say about the sparsity of the marginalised stiffness matrix and compare it with the cgDNA+ stiffness sparsity pattern?

3.2 Marginalise over phosphates and intra-base-pair variables

Repeat the above exercise but for computing marginals over both phosphates and intras variables. What can you say?

4 A localized cgDNA+ model: marginals over the configurations of the flanking sequences

The following marginalisation could be useful to study the statistical mechanics property of a small segment of a potentially very long fragment of DNA. Begin by adding randomly 100 basepairs at each end of the Dickerson dodecamer, i.e, define $\tilde{S} = S_1 S_D S_2$ where S_1, S_2 are randomly chosen (but then fixed) 100 basepair long sequences. Do the following steps:

i) Reconstruct the stiffness matrix and the groundstate for $\tilde{S}$ using cgDNA+ matlab package.

ii) Invert the reconstructed stiffness matrix and extract the entries of the covariance that correspond to S_D. Invert them to obtain the marginalised stiffness of Dickerson dodecamer.

iii) Extract the entries of the groundstate corresponding to S_D.

What is the sparsity pattern of the marginalised stiffness? Compare the marginal stiffness and marginal groundstate with the corresponding cgDNA+ reconstruction of S_D (for example compute the Kullback-Leibler divergence between the two distributions). What happen if you change the flanking sequences?

[Note: Based on above method one can also marginalise over flanking sequences. By considering the following ensemble $\mathcal{S}(S_D) = \{S | S = S_1 S_D S_2,\ S_1, S_2 \text{ flanking sequences}\}$, one can compute the marginal of S_D over flanking sequences as the ensemble average of all the localized marginals of S_D computed for all $S \in \mathcal{S}$.]