1 Sampling amplitudes and timescales

Consider a simple amplitude scaling move using random number $c \sim U(0,1)$:

$$a' = (c + 0.5)a = ac + 0.5a$$

The proposal densities are then

$$\begin{aligned} p(a \to a') &= a \\ p(a' \to a) &= a' \end{aligned}$$

giving rise to the ratio:

$$\frac{p(a' \to a)}{p(a \to a')} = \frac{a'}{a} = c + 0.5\,.$$

Likewise we can sample with a different random process that keeps the amplitudes closer to 1 (probably better for acceptance).

2 Splitting and merging states

This is probably not needed for the fingerprints because fingerprints are estimated on a timescale grid. Anyway, for future reference....

2.1 State splitting probability

Like a direct Markov model [**?**], the estimated HMM will provide a consistent model of the stationary and kinetic properties of the data if the Markov states are sufficiently fine, i.e., if there are sufficiently many states such that the conformational dynamics is appropriately described by a Markov chain. We conducted tests of the model consistency by comparing directly estimated and HMM-generated FRET efficiency histograms on various timescales as described in Sec. "Time-binned FRET efficiency distributions", and, when this test fails we consider to split states that are most likely to be the course for the failure. This procedure is continued until the HMM test succeeded. To select candidates for splitting, we considered the lifetime distributions described in Sec. "Lifetime distributions" and identified those states whose lifetimes can clearly not be described by a single exponential.

Since the lifetime distributions are computed from a finite number of realizations, the decision whether an estimated lifetime distribution is single-exponential or not, must be based on statistics. Here, we develop a Markov Chain Monte Carlo (MCMC) algorithm that, for each estimated cumulative lifetime distribution $\hat{p}(t)$, performs a model selection between a single exponential generating model:

$$p_1(t) = \mathrm{e}^{-tk}$$

and a bi-exponential generating model:

$$p_2(t) = a\mathrm{e}^{tk_1} + (1-a)\mathrm{e}^{tk_2}.$$

The probability for either of these models to generate a sample of segment lengths $(t_1, ..., t_n)$ is given by:

$$\mathbb{P}(\lambda_x \mid t_1, ..., t_n) = \mathbb{P}_x(\lambda) \prod_i p_x(t_i)\,,$$

where we use Jeffrey's prior:

$$\mathbb{P}_1(k) = \frac{1}{k}$$

and

$$\mathbb{P}_2(a, k_1, k_2) = \frac{1}{ak_1 + (a-1)k_2}$$

When correctly defined, this MCMC algorithm will sample from each of the two models according to their respective probabilities to have generated the observed set of exit times. Such an MCMC algorithm requires at least four Monte Carlo steps: (1) a step that can sample new parameters k within the single-exponential model, (2) a step that can sample new parameters a, k_1, k_2 in the bi-exponential model, (3) a step to split a single-exponential model into a bi-exponential model, (4) a step to merge a bi-exponential model into a single-exponential model.

In order to implement the split and merge steps, we need to propose a rule by which the single-exponential parameter k and the bi-exponential parameters a, k_1, k_2 are related, and then compute the appropriate MCMC acceptance probabilities from this rule. Consider the following relation:

$$\begin{aligned} k &= ak_1 + (1-a)k_2 \\ &= a(k_1 - k_2) + k_2 \end{aligned} \tag{1}$$

and consider further the parametrization

$$k_1 = bk$$

with $a, b \in [0,1]$ and $k_1 \leq k_2$. We define the splitting move by generating a, b as uniform random numbers in $[0,1]$. We obtain

$$k_2 = \frac{1-ab}{1-a}k$$

In order to calculate the proposal probability of the splitting move consider the random number distributions

$$\begin{aligned} p(a) &= 1 \quad a\in[0,1] \\ p(b) &= 1 \quad b\in[0,1] \end{aligned}$$

we transform the variables (a, b) into (k_1, k_2):

$$\begin{aligned} a &= \frac{k-k_2}{k_1-k_2} \\ b &= \frac{k_1}{k} \end{aligned}$$

This involves the Jacobian:

$$\begin{aligned} \mid J \mid &= \left| \det \begin{pmatrix} \frac{d}{dk_1}a & \frac{d}{dk_2}a \\ \frac{d}{dk_1}b & \frac{d}{dk_2}b \end{pmatrix} \right| \\ &= \frac{k-k_1}{(k_1-k_2)^2}\frac{1}{k} \end{aligned}$$

This yields the splitting proposal density

$$p(k \to a, k_1, k_2) \propto \frac{k-k_1}{(k_1-k_2)^2}\frac{1}{k} = \frac{1-b}{(k_1-k_2)^2} \quad k_1 \leq k$$

while the merging proposal density is given by:

$$p(a, k_1, k_2 \to k) = 1$$

yielding the splitting acceptance probability:

$$p_{acc}^{\text{split}} = \frac{\mathbb{P}_2(a, k_1, k_2 \mid t_1, ..., t_n)}{\mathbb{P}_1(k \mid t_1, ..., t_n)} \frac{(k_1-k_2)^2}{(1-b)}$$

and the merging acceptance probability:

$$p_{acc}^{\text{merge}} = \frac{\mathbb{P}_1(k \mid t_1, ..., t_n)}{\mathbb{P}_2(a, k_1, k_2 \mid t_1, ..., t_n)} \frac{(1-b)}{(k_1-k_2)^2}$$

For the MCMC steps that change the parameters within a given model, we consider the straightforward and the uniform move $a \sim U(0,1)$, and the rate scaling move:

$$k' = (c + 0.5)k = kc + 0.5k$$

with random number $c \sim U(0,1)$. The proposal densities are

$$\begin{aligned} p(k \to k') &= k \\ p(k' \to k) &= k' \end{aligned}$$

giving rise to the ratio:

$$\frac{p(k' \to k)}{p(k \to k')} = \frac{k'}{k} = c + 0.5\,.$$

This results in the sampling algorithm 5. When Algorithm 5 returns $n > 1.5$, the corresponding hidden state should be split.

2.2 Performing a state splitting

Without restriction of generality we consider that the n'th state will be split, and we generate a new set of state parameters for the childs $(n, n+1)$ as described below. The new parameters will serve as an input to an EM algorithm, in which the new full parameter set is optimized to convergence.

We find a separation of the nth exit time distribution in terms of

$$p(\tau) = a\mathrm{e}^{-k_n\tau} + (1-a)\mathrm{e}^{-k_{n+1}\tau}\,.$$

Given a transition element Q_{ii} we have the relationship

$$k_i = -\ln Q_{i,i}$$

suggesting diagonal matrix elements

$$Q_{i,i} = \mathrm{e}^{-k_i}\,.$$

We start with matrix $\mathbf{T}$ and stationary distribution $\boldsymbol{\pi}$, for which the corresponding correlation matrix is defined as

$$\mathbf{C} = \mathbf{\Pi T}$$

with $\mathbf{\Pi} = \mathrm{diag}(\boldsymbol{\pi})$. Furthermore, let us assume we want to split the last state, n, such that the new states have diagonal elements given by

$$\begin{aligned} Q_{n,n} &= \mathrm{e}^{-k_n} \\ Q_{n+1,n+1} &= \mathrm{e}^{-k_{n+1}}\,, \end{aligned}$$

and the relative probabilities are given by

$$\frac{\pi_n}{\pi_{n+1}} = \frac{a}{1-a}$$

An new correlation matrix $\mathbf{D}$ with a modified state n and and additional state $n+1$ is obtained from the original matrix $\mathbf{C}$ as

$$\mathbf{D} = \begin{pmatrix} C_{11} & \cdots & C_{1,n-1} & D_{1,n} & D_{1,n+1} \\ \vdots & & \vdots & \vdots & \\ C_{n-1,1} & \cdots & C_{n-1,n-1} & \vdots & \vdots \\ D_{n,1} & \cdots & \cdots & D_{n,n} & \\ D_{n+1,1} & & \cdots & & D_{n+1,n+1} \end{pmatrix}$$

with the constraints:

1. $Q_{n,n} = \frac{D_{n,n}}{\sum_i D_{n,i}} = \mathrm{e}^{-k_n}$

2. $Q_{n+1,n+1} = \frac{D_{n+1,n+1}}{\sum_i D_{n+1,i}} = \mathrm{e}^{-k_{n+1}}$

3. $\frac{\pi'_n}{\pi'_{n+1}} = \frac{\sum_i D_{n,i}}{\sum_i D_{n+1,i}} = \frac{a}{1-a}$

4. $\pi'_n + \pi'_{n+1} = \pi_n$

5. $D_{ij} = D_{ji} \quad \forall i$

We intend to fulfill all these contraints, and additionally staying "close" to $D_{n,i} + D_{n+1,i} = C_{n,i} \quad i \in \{1, n-1\}$. From 3 and 4 we get the row sums:

$$\begin{aligned}
\pi'_{n+1} &= \pi_n - \pi'_n \\
\pi'_n &= \pi'_{n+1}\frac{a}{1-a} \\
&= (\pi_n - \pi'_n)\frac{a}{1-a} \\
\pi'_n &= \pi_n \frac{\frac{a}{1-a}}{(1+\frac{a}{1-a})} = \pi_n a
\end{aligned}$$

From the row sums and constraints 1 and 2 we get the diagonals:

$$\begin{aligned}
D_{n,n} &= \pi'_n \mathrm{e}^{-k_n} = d_1 \\
D_{n+1,n+1} &= \pi'_{n+1}\mathrm{e}^{-k_{n+1}} = d_2
\end{aligned}$$

Next we fill the lower right block. In the ideal case we can maintain counts of the split diagonal elements.

$$\begin{aligned}
C_{nn} &= D_{n,n} + 2D_{n+1,n} + D_{n+1,n+1} \\
D_{n+1,n} &= \frac{C_{nn} - D_{n,n} - D_{n+1,n+1}}{2}
\end{aligned}$$

However, if that would results in $D_{n+1,n} < 0$ or $D_{n+1,n} \geq \min\{\pi'_n - D_{n,n}, \pi'_{n+1} - D_{n+1,n+1}\}$ this solution can't be used, and instead we resort to some number

$$D_{n+1,n} = \epsilon < \min\{\pi'_n - D_{n,n}, \pi'_{n+1} - D_{n+1,n+1}\}$$

We now have remaining counts to be distributed:

$$\begin{aligned}
r_n &= \pi'_n - D_{n,n} - D_{n+1,n} \\
r_{n+1} &= \pi'_{n+1} - D_{n+1,n+1} - D_{n+1,n}
\end{aligned}$$

and set the remaining elements as:

$$\begin{aligned}
D_{n,i} &= r_n \frac{C_{n,i}}{\pi_n} \\
D_{n+1,i} &= r_{n+1}\frac{C_{n,i}}{\pi_n}
\end{aligned}$$

Finally, we normalize $\mathbf{D}$ row-wise to get a transition matrix.

The new state parameters are used as an input for an EM algorithm, in which the HMM parameters are again optimized to convergence.

Algorithm 1 Number_of_exponentials$(t_1, ..., t_n)$

Input: A set of lifetimes $t_1, ..., t_n$

Output: $(n,\ k, a, k_1, k_2)$ where $n \in [1, 2]$ is the estimated number of exponentials required to fit the data, k is the rate parameter of the single-exponential model, and a, k_1, k_2 are the parameters of the bi-exponential model.

1. $n_{exp} = 1$, $n_1 = 0$, $k_{1,sum} = 0$, $a_{sum} = 0$, $k_{21,sum} = 0$, $k_{22,sum} = 0$
2. For $i = 1, ..., N_{sample}$

2.1. $r_1 \sim U(0, 1)$

2.2. If $n_{exp} = 1$

If $r_1 < 0.5$

Propose rate change $k \to k' = (r_2 + 0.5)k$ with $r_2 \sim U(0, 1)$. Accept with

$$p_{acc}^{k} = (r_2 + 0.5)\frac{\mathbb{P}_1(k' \mid t_1, ..., t_n)}{\mathbb{P}_1(k \mid t_1, ..., t_n)}$$

If $r_1 \geq 0.5$

$a \sim U(0, 1)$, $b \sim U(0, 1)$. Propose split $k \to \left(a,\ k_1 = bk,\ k_2 = \frac{1-ab}{1-a}k\right)$. Accept with:

$$p_{acc}^{\text{split}} = \frac{\mathbb{P}_2(a, k_1, k_2 \mid t_1, ..., t_n)}{\mathbb{P}_1(k \mid t_1, ..., t_n)}\frac{(k_1 - k_2)^2}{(1 - b)}$$

Else if $n_{exp} = 2$

If $r_1 < 0.5$

$r_3 \sim U(0, 1)$

If $r_3 < \frac{1}{3}$: Propose new amplitude $a' \sim U(0, 1)$. Accept with

$$p_{acc} = \frac{\mathbb{P}_2(a', k_1, k_2 \mid t_1, ..., t_n)}{\mathbb{P}_2(a, k_1, k_2 \mid t_1, ..., t_n)}$$

Else: propose rate change $k_{1/2} \to k'_{1,2} = (r_2 + 0.5)k_{1/2}$ with $r_2 \sim U(0, 1)$. Accept with

$$p_{acc}^{k} = (r_2 + 0.5)\frac{\mathbb{P}_1(k'_{1,2} \mid t_1, ..., t_n)}{\mathbb{P}_1(k_{1,2} \mid t_1, ..., t_n)}$$

If $r_1 \geq 0.5$

Propose merge $(a, k_1, k_2 \to k = ak_1 + (1 - a)k_2)$. Accept with:

$$p_{acc}^{\text{merge}} = \frac{\mathbb{P}_1(k \mid t_1, ..., t_n)}{\mathbb{P}_2(a, k_1, k_2 \mid t_1, ..., t_n)}\frac{(1 - b)}{(k_1 - k_2)^2}$$

3. Return:

$$\begin{aligned} n &= 1 + n_1/N_{sample} \\ k &= k_{1,sum}/N_{sample} \\ a &= a_{1,sum}/N_{sample} \\ k_1 &= k_{21,sum}/N_{sample} \\ k_2 &= k_{22,sum}/N_{sample} \end{aligned}$$