0.1 Comparing bilateral models

Up to this point we have largely argued that the mixing parameter makes intuitive sense because of the thought experiment, where we moved the primary tumor from a clearly lateralized posititon closer and closer to the mid-sagittal plane, until it was perferctly symmetric w.r.t. that plane. However, we now need to actually test whether or not our arguments hold. For that, we decided to compare three models:

- Model $\mathcal{M}_{\text{ag}}$, which is agnostic to the tumor's extension e over the mid-sagittal plane and treats the contralateral base spread in the same way for all patients.
- Model $\mathcal{M}_{\alpha}$ that uses the linear combination of the ipsilateral base probabilities and the contralateral ones for the patients without mid-plane extension to describe the spread for tumors which do extend over that plane.
- Model $\mathcal{M}_{\text{full}}$, going even further by defining a completely independent set of contralateral base probabilities for the patients whose tumor extends over the mid-sagittal plane.

Essentially, we now want to know which of these three models does the best job of describing the data. Intuitively, one would argue that it must be $\mathcal{M}_{\text{full}}$, but this model is also more complex than the other two. A natural choice for a metric that incorporates both the accuracy of the model and a penalty for model complexity – often also called *Occam's razor* – is the *model evidence* [1].

Model evidence and Bayes factor

In Bayesian terms, we would like to know which model $\mathcal{M}$ has the highest probability $P(\mathcal{M} \mid \boldsymbol{\mathcal{D}})$ given a dataset $\boldsymbol{\mathcal{D}}$. This probability is given by

$$P(\mathcal{M} \mid \boldsymbol{\mathcal{D}}) = \frac{P(\boldsymbol{\mathcal{D}} \mid \mathcal{M})\, P(\mathcal{M})}{P(\boldsymbol{\mathcal{D}})} \tag{1}$$

If a priori all models we want to consider have the same probability $P(\mathcal{M})$ and we only make pairwise comparisons between models, then we can restrict ourselves to computing the *Bayes factor*:

$$K_{1\text{v}2} = \frac{P(\mathcal{M}_1 \mid \boldsymbol{\mathcal{D}})}{P(\mathcal{M}_2 \mid \boldsymbol{\mathcal{D}})} = \frac{P(\boldsymbol{\mathcal{D}} \mid \mathcal{M}_1)\, P(\mathcal{M}_1)}{P(\boldsymbol{\mathcal{D}} \mid \mathcal{M}_2)\, P(\mathcal{M}_2)} = \frac{P(\boldsymbol{\mathcal{D}} \mid \mathcal{M}_1)}{P(\boldsymbol{\mathcal{D}} \mid \mathcal{M}_2)} \tag{2}$$

On the right side in the above equation, we see the ratio of the two model's evidences, which are merely their respective likelihoods, marginalized over all parameters:

$$P(\boldsymbol{\mathcal{D}} \mid \mathcal{M}) = \int p(\boldsymbol{\mathcal{D}} \mid \theta, \mathcal{M})\, p(\theta \mid \mathcal{M}) d\theta \tag{3}$$

So, if we can compute this model evidence – commonly also called *marginal likelihood* or *partition function* Z from physics – for our models $\mathcal{M}_{\text{ag}}$, $\mathcal{M}_{\alpha}$ and $\mathcal{M}_{\text{full}}$, the respecive pairwise Bayes factors will indicate which of them is *most likely* to be the true one, given the observed data, in the probabilistic sense. Note that this

does not mean it *is* the true data-generating model and not even that we should *believe* it is the true one. But only that among the models investigated, this one is probably the best.

Harold Jeffreys gives a scale for interpreting values of the Bayes factor [3]:

K_{1v2}	$\ln K_{1v2}$	support for $\mathcal{M}_1$
$< 10^0$	< 0	negative evidence (supports $\mathcal{M}_2$)
10^0 to $10^{1/2}$	0 to 1.15	barely worth a mention
$10^{1/2}$ to 10^1	1.15 to 2.3	substantial
10^1 to $10^{3/2}$	2.3 to 3.45	strong
$10^{3/2}$ to 10^2	3.45 to 4.6	very strong
$> 10^2$	> 4.6	decisive

We have also listed the natural logarithm $\ln K_{1v2}$ of the Bayes factor here, because what we will actually be doing is compute differences in the log-evidences.

Thermodynamic integration

Due to the integration over all model parameters, the quantity eq. (3) is usually impossible to calculate by brute force integration, even for models with only around a dozen parameters, as is the case for ours. Unless analytical solutions exist – which is rarely the case – it is often prohibitively expensive to compute the model evidence. For this reason, a large amount of approximation methods has been developed; [2] names only a few of those methods that can be used in the context of Markov-chain Monte Carlo (MCMC). Another method that is applicable in the context of MCMC is thermodynamic integration (TI), which is very well introduced in [1] and only roughly sketched out in this section.

The concept of TI originates from the field of statistical mechanics and can be motivated from that standpoint. And although this path is certainly more educational and might convey a deeper understanding w.r.t. thermodynamics and information theory, we will take a more direct approach by starting with what we want to compute and subtracting a 0 from it:

$$\begin{aligned} \ln Z \coloneqq \ln p(\boldsymbol{\mathcal{D}} \mid \mathcal{M}) &= \ln \int p\left(\boldsymbol{\mathcal{D}} \mid \theta, \mathcal{M}\right) p(\theta \mid \mathcal{M}) d\theta - \ln 1 \\ &= \ln \int p\left(\boldsymbol{\mathcal{D}} \mid \theta, \mathcal{M}\right) p(\theta \mid \mathcal{M}) d\theta - \underbrace{\ln \int p(\theta \mid \mathcal{M}) d\theta}_{\ln Z_0} \end{aligned} \tag{4}$$

Writing it as this difference between two different log-evidences $\ln Z$ and $\ln Z_0$ itself does not get us far. But if we could somehow parametrize a differentiable path between the two, then maybe the integration

$$\ln Z - \ln Z_0 = \int_0^1 \frac{d}{d\beta} \ln Z_\beta d\beta \tag{5}$$

we end up with can actually be computed. Just by inspection of eq. (4) and eq. (5), one can see that on such differentiable path could be built using what we are going

to call the *power posterior* $p_\beta(\theta \mid \boldsymbol{\mathcal{D}}, \mathcal{M})$:

$$\begin{aligned} \ln Z_\beta &= \ln \int p_\beta(\theta \mid \boldsymbol{\mathcal{D}}, \mathcal{M}) d\theta \\ &= \ln \int p\left(\boldsymbol{\mathcal{D}} \mid \theta, \mathcal{M}\right)^\beta p(\theta \mid \mathcal{M}) d\theta \end{aligned} \tag{6}$$

with the derivative

$$\begin{aligned} \frac{d}{d\beta} \ln Z_\beta &= \int \frac{p\left(\boldsymbol{\mathcal{D}} \mid \theta, \mathcal{M}\right)^\beta p(\theta \mid \mathcal{M})}{Z_\beta} \ln p(\boldsymbol{\mathcal{D}} \mid \theta, \mathcal{M}) d\theta \\ &= \mathbb{E}\left[\ln p(\boldsymbol{\mathcal{D}} \mid \theta, \mathcal{M})\right]_{p_\beta(\theta \mid \boldsymbol{\mathcal{D}}, \mathcal{M})} \\ &\approx \frac{1}{S} \sum_{i=1}^{S} \ln p(\boldsymbol{\mathcal{D}} \mid \hat{\theta}_{\beta i}, \mathcal{M}) = A_{\mathrm{MC}}(\beta) \end{aligned} \tag{7}$$

The solution to computing the evidence now lies in sight: Using MCMC, we can draw samples from the power posterior p_β and use those samples to compute the expectation over the (unmodified) likelihood. Doing this for a sufficient number of steps in the interval $[0, 1]$ and integrating over the resulting $A_{\mathrm{MC}}(\beta)$ will then yield an approximation to the log-evidence.

$$\ln Z \approx \frac{1}{2} \sum_{j=1}^{N-1} (\beta_{j+1} - \beta_j) \left(A_{\mathrm{MC}}(\beta_{j+1}) + A_{\mathrm{MC}}(\beta_j)\right) \tag{8}$$

This approximation gets better with larger values for S and N. But also how the β_j are chosen is crucial for computing a good estimate: Usually, the $A_{\mathrm{MC}}(\beta)$ – which can be seen as accuracy terms – rise steeply for increasing β close to 0, while levelling off towards $\beta = 1$. It therefore makes sense to distribute the ladder of these values unevenly. A common choice that we employed as well was $\beta_j = x_j^5$ where the x_j are linearly spaced within the interval $[0, 1]$. This yields a very fine resolution for the first steps and gets successively coarser towards the end of the interval.