# BE/Bi 103: Data Analysis in the Biological Sciences

Justin Bois

Caltech

Fall, 2017

# 8 Parallel tempering MCMC

In this lecture, we will discuss parallel tempering Markov chain Monte Carlo (PTMCMC). This technique allows for effective sampling of multimodal distributions and it avoids getting trapped on local maxima of the posterior. Perhaps even more importantly, it allows us to perform model selection.

## 8.1 The basic idea

Recall that the posterior distribution we seek to sample in the model selection problem is

$$g(\theta_i \mid D, M_i) \propto g(\theta_i \mid M_i) f(D \mid \theta_i, M_i). \tag{8.1}$$

Now, we define

$$g_{\text{hot}}(\theta_i \mid D, M_i, \beta) = \frac{1}{Z_i(\beta)}\, g(\theta_i \mid M_i)\, [f(D \mid \theta_i, M_i)]^{\beta} \tag{8.2}$$

$$= \frac{1}{Z_i(\beta)}\, g(\theta_i \mid M_i) \exp\left[\beta \ln f(D \mid \theta_i, M_i)\right]. \tag{8.3}$$

Here, $\beta \in (0, 1]$ is an "inverse temperature" in analogy to statistical mechanics, where the negative log likelihood, $-\ln f(D \mid \theta_i, M_i)$, is an energy. Keeping with the analogy, the normalization constant $Z_i(\beta)$, given by

$$Z_i(\beta) = \int \mathrm{d}\theta_i\, g(\theta_i \mid M_i)\, [f(D \mid \theta_i, M_i)]^{\beta}\,, \tag{8.4}$$

is called a **partition function**. We will call the distribution $g_{\text{hot}}(\theta_i \mid D, M_i, \beta)$ a **hot posterior** because it is the posterior with a high temperature.

If $\beta$ is close to zero (the "high temperature" limit), we are just sampling the prior. If $\beta = 1$, we are sampling our target posterior, the so-called "cold distribution." So, lowering $\beta$ has the effect of flattening the posterior distribution. Therefore, walkers at higher temperature (lower $\beta$) are not trapped at local maxima. By occasionally swapping walkers from adjacent temperatures, we can effectively sample a broader swath of parameter space.

In practice, we choose a set of $\beta$'s with $\beta = [\beta_0, \beta_1, \ldots, \beta_m]$, with $\beta_{i+1} < \beta_i$ and $\beta_0 = 1$. We propose a swap roughly every $n_s$ steps and accept it based on criteria that guarantees the posterior is a stationary distribution of the transition kernel. To do this in practice, we choose a uniform random number on $[0, 1]$ every iteration and propose a swap when this random number is less than $1/n_s$. When we do propose a

swap, we randomly pick a temperature $\beta_j$ from $\{\beta_1, \beta_2, \ldots \beta_m\}$. We then compute

$$r = \min\left(1, \frac{g_{\text{hot}}(\theta_{i,j} \mid D, M_i, \beta_{j-1})}{g_{\text{hot}}(\theta_{i,j-1} \mid D, M_i, \beta_{j-1})} \frac{g_{\text{hot}}(\theta_{i,j-1} \mid D, M_i, \beta_j)}{g_{\text{hot}}(\theta_{i,j} \mid D, M_i, \beta_j)}\right). \tag{8.5}$$

Here, we have defined $\theta_{i,j}$ as the value of parameter $i$ for a walker at temperature $\beta_j$. Note that this calculation does not require calculation of any partition functions; the $Z_i(\beta)$ cancel out in the expression for $r$. We then draw another uniform random number on $[0, 1]$ and accept the swap is that number if less than $r$.

This useful technique is implemented the package ptemcee (pronouced tem-see; the p is silent). Conveniently, it automatically chooses reasonable values of $\beta$ and swapping rate, though you can specify these as well. It also has a bit more sophistication that what I have described here, using adaptive parallel tempering.

## 8.2 Model selection with PTMCMC

We will now do some clever ticks to see how we can use PTMCMC to do model comparison without making the approximations we have thus far. In fact, we do not necessarily need parallel tempering with swapping; we only need samples of $g_{\text{hot}}(\theta_i \mid D, M_i, \beta)$ for various values of $\beta$. Recall the statement of Bayes's theorem for the model comparison problem, equation (5.3).

$$g(M_i \mid D) = \frac{f(D \mid M_i)\, g(M_i)}{f(D)}. \tag{8.6}$$

The likelihood in the model selection problem is given by the evidence, a.k.a. fully marginalized likelihood, from the parameter estimation problem, as we derived in equation (5.5). Thus,

$$g(M_i \mid D) = \frac{g(M_i)}{f(D)} \left[\int \mathrm{d}\theta_i\, g(\theta_i \mid M_i) f(D \mid \theta_i, M_i)\right]. \tag{8.7}$$

We recognize the bracketed term as $Z_i(1)$. Our goal is to calculate this quantity.

Now, we're going to do a usual trick in statistical mechanics: we will differentiate the log of the partition function (analogous to the derivative of a free energy).

$$\begin{aligned}
\frac{\partial}{\partial\beta} \ln Z_i(\beta) &= \frac{1}{Z_i(\beta)} \frac{\partial Z_i}{\partial\beta} \\
&= \frac{1}{Z_i(\beta)} \int \mathrm{d}\theta_i \frac{\partial}{\partial\beta} \exp\left[\ln g(\theta_i \mid M_i) + \beta \ln f(D \mid \theta_i, M_i)\right] \\
&= \frac{1}{Z_i(\beta)} \int \mathrm{d}\theta_i \ln f(D \mid \theta_i, M_i) \exp\left[\ln g(\theta_i \mid M_i) + \beta \ln f(D \mid \theta_i, M_i)\right]
\end{aligned}$$

$$= \frac{1}{Z_i(\beta)} \int \mathrm{d}\theta_i \, \ln f(D \mid \theta_i, M_i) \, g(\theta_i \mid M_i) \, [f(D \mid \theta_i, M_i)]^{\beta} \, . \tag{8.8}$$

We recognize this as the average of the log likelihood $\ln f(D \mid \theta_i, M_i)$ over the distribution $g_{\mathrm{hot}}(\theta_i \mid D, M_i, \beta)$. We denote this as

$$\frac{\partial}{\partial \beta} \ln Z_i(\beta) = \langle \ln f(D \mid \theta_i, M_i) \rangle_{g_{\mathrm{hot}}(\theta_i | D, M_i, \beta)} \, . \tag{8.9}$$

Note that this average if done for each specific value of $\beta$ we are considering, and that the derivative of the log partition function is thus a function of $\beta$. Now, we can integrate both sizes of this equation to give

$$\begin{aligned} \int_0^1 \mathrm{d}\beta \, \frac{\partial}{\partial \beta} \ln Z_i(\beta) &= \ln Z_i(1) - \ln Z_i(0) \\ &= \int_0^1 \mathrm{d}\beta \, \langle \ln f(D \mid \theta_i, M_i) \rangle_{g_{\mathrm{hot}}(\theta_i | D, M_i, \beta)} \, . \end{aligned} \tag{8.10}$$

Now, if the prior is normalized, as it should be,

$$Z_i(0) = \int \mathrm{d}\theta_i \, g(\theta_i \mid M_i) = 1, \tag{8.11}$$

which means $\ln Z_i(0) = 0$. Thus, we get a fully marginalized likelihood of

$$\begin{aligned} \ln Z_i(1) &= \int \mathrm{d}\theta_i f(D \mid \theta_i, M_i) \, g(\theta_i \mid M_i) \\ &= \int_0^1 \mathrm{d}\beta \, \langle \ln f(D \mid \theta_i, M_i) \rangle_{g_{\mathrm{hot}}(\theta_i | D, M_i, \beta)} \, . \end{aligned} \tag{8.12}$$

Fortunately, if we have done PTMCMC, we have sampled out of the distribution $g_{\mathrm{hot}}(\theta_i \mid D, M_i, \beta)$ for various values of $\beta$. We can then compute the integrand in the above equation for each $\beta$ at which we sampled.

$$\langle \ln f(D \mid \theta_i, M_i) \rangle_{g_{\mathrm{hot}}(\theta_i | D, M_i, \beta)} = \frac{1}{n_{\mathrm{samples}}} \sum_{\mathrm{samples}} \ln f(D \mid \theta_i, M_i). \tag{8.13}$$

We just have to compute the log likelihood (*not* the hot log-likelihood) for each MCMC sample for a given inverse temperature $\beta$, and we have all we need. We then perform numerical quadrature across the values of $\beta$ that we sampled to get the integral. We therefore can compute the odds ratio of two models $M_i$ and $M_j$ as

$$O_{ij} = \frac{g(M_i \mid I)}{g(M_j \mid I)} \frac{Z_i(1)}{Z_j(1)} \tag{8.14}$$

$$= \frac{g(M_i \mid I)}{g(M_j \mid I)} \exp \left[ \frac{\int_0^1 \mathrm{d}\beta \ \langle \ln f(D \mid \theta_i, M_i) \rangle_{g_{\mathrm{hot}}(\theta_i | D, M_i, \beta)}}{\int_0^1 \mathrm{d}\beta \ \langle \ln f(D \mid \theta_j, M_j) \rangle_{g_{\mathrm{hot}}(\theta_i | D, M_i, \beta)}} \right],$$

where the last ratio is computed via numerical quadrature on results computed directly from our PTMCMC traces using equation (8.13). Note that we have made no approximations at all in the model. The calculation is only approximate to the extent that the PTMCMC sampler takes a finite number of samples and numerical quadrature is not exact.