Bayesian Statistics

Group exercise 3

Read before you start

For these group exercises, you are expected to deliver the following:

- R & JAGS code for each sub-exercise.
- A report (PDF) containing the answers to the questions, including plots/figures where necessary, and referring to the code you wrote. Make sure figures etc. are properly visible and have informative labels.

Apart from the book by Kruschke, you may also need the JAGS manual[1] to look up the details of the distributions you will need to use.

1 Is the Euro a fair coin?

For this exercise, some computations should be done in R, as they involve some *evidence* terms for the beta-Bernoulli model. Recall however, that you have used these in past exercises, so be efficient and re-use your code.

Shortly after the Euro was introduced as a currency, the following statement appeared in *The Guardian*, on January 4th, 2002:

> When spun on edge 250 times, a Belgian one-euro coin came up heads 140 times and tails 110. 'It looks very suspicious to me,' said Barry Blight, a statistics lecturer at the London School of Economics. 'If the coin were unbiased, the chance of getting a result as extreme as that would be less than 7%.'

Let's see whether this bold claim makes sense.

1. (20pts) According to dr. Blight, this coin is suspect of being heads-biased. To conclude this, dr. Blight used a *two-sided binomial test*, which got him this particular p-value. First, replicate his frequentist analysis by computing the p-value in R (without using `binom.test`). Note that the probability of one particular observed value k using the binomial distribution is

$$p(z = k \mid N, \theta) = \binom{N}{k} \theta^k (1-\theta)^{N-k} \ ,$$

   but we are interested in the two-sided test of an equal or more extreme observation. What is the p-value up to 4 decimals (it should indeed be close to 0.07 or you have made a mistake!)?[2]

2. (10pts) What assumptions on data collection did Dr. Blight make?

Now let's investigate how these results hold up to a Bayesian analysis for different priors.

3. (40pts) Consider two models. The first, m_0, states that the coin is fair, i.e. $\theta = 0.5$.[3] Essentially, this is a spike prior defined as

$$p(\theta \mid m_0) = \begin{cases} 1, & \text{if } \theta = 0.5 \\ 0, & \text{otherwise.} \end{cases}$$

[1] http://www.stats.ox.ac.uk/ nicholls/MScMCMC15/jags$_u ser_m anual.pdf$

[2] Note that the approach to compute the p-value is described on pages 331 and 332 of the book, but that formulation only works for $\theta = 0.5$, not for asymmetric binomial distributions.

[3] If you want to be annoying you can say 'well the coin will never be *exactly* perfect!'. In that case, our prior could be defined as $\theta \in [0.499, 0.501]$.

The alternative, m_1, considers a uniform prior probability on θ, e.g. $\theta \mid m_1 \sim \text{beta}(\theta \mid 1, 1)$. Compute the evidence $p(z, N \mid m_0)$, the evidence $p(z, N \mid m_1)$ and the Bayes factor

$$BF_{10} = \frac{p(z, N \mid m_1)}{p(z, N \mid m_0)} \ ,$$

using the analytical expressions for these terms. Show how you got your result. To actually compute the results you will have to use R, but you don't need (and shouldn't) use JAGS/MCMC. What does this Bayes factor mean for the fairness of the coin?

4. (30pts) Dr. Blight just read your answer and tells you 'No, no, your uniform prior is silly. Only small biases around θ are sensible — of course I'm not really considering $\theta = 0.01$, for example!'. Well, you think, fair enough, let's adjust our prior beliefs to be less uniform. So let's replace the prior by $\theta \mid m_1 \sim \text{beta}(\theta \mid \alpha, \alpha)$ (so a and b are both set to α. For $\alpha = 1$ we recover the prior from the previous question.) For larger α, our alternative hypothesis will peak more and more around $\theta = 0.5$ (because the mode of the beta distribution is $a/(a+b) = \alpha/2\alpha = 1/2$). Now try different values for α, starting at $\alpha = 1$ and ending at some large value for α[4]. Plot the Bayes factor of the model comparison (as in the previous question) as a function of α and superimpose on the plot a straight line that shows $BF = 1$; the case where the two models are equally likely. What is the maximum Bayes factor you find? What does this mean for the comparison of the two models?

5. (20pts) If you look at the plot, especially at its behaviour for large α, then it seems as if the Bayes factor converges to something. Why is that and does it make sense?

6. (30pts) The most extreme prior we can come up with, in terms of this likelihood, is a prior that places *all* probability mass exactly on the empirical mean (which we call the *maximum likelihood estimate* for θ) $\theta_{\text{MLE}} = z/N = 0.56$. Let's call this the prior for m_2. This is the spike prior we have seen earlier, now defined as:

$$p(\theta \mid m_2) = \begin{cases} 1, & \text{if } \theta = \theta_{\text{MLE}} = 0.56 \\ 0, & \text{otherwise.} \end{cases}$$

Note that this is a weird prior, as it assumes θ can *only* be *exactly* 0.56. That is some serious prior knowledge we're putting in, since 0.56 is a rather arbitrary number. This prior results in a Bayes factor that gives the strongest possible evidence *against* a fair coin, because it agrees most with the observations, which are not exactly at 50/50 heads and tails. Note that the evidence for m_2 is given by

$$p(z, N \mid m_2) = \int \mathrm{d}\theta \, p(z, N \mid \theta) p(\theta \mid m_2) \ ,$$

which has a very simple form, which you have seen and used before. What is this form (hint: note that the integral in this equations reads as 'for all θ', but the prior $p(\theta \mid m_2)$ has some very strong opinions about this...).

7. (10pts) Using this form, what is the Bayes factor $p(z, N \mid m_2)/p(z, N \mid m_0)$? Show how you got to your answer!

8. (20pts) Look at the Bayes factor interpretation table again. Also look at the quote by dr. Blight about the fairness of the coin. Now back to the Bayes factor table. Now back to the quote again... What do you conclude about the coin and dr. Blight's assessment of it? Motivate your answer.

9. (20pts) A tiny change in the data can make matters even worse. Let's assume $N = 250$ as before, but now $z = 141$, i.e. one more coin flip resulted in 'heads' compared to earlier. Compute again the p-value for this outcome using the approach of the first question. What are the implications of this result for a typical significance threshold of 0.05?

10. (10pts) Also compute the Bayes factors $p(z, N \mid m_1)/p(z, N \mid m_0)$ for a range of α using these slightly altered data. What is the maximum possible Bayes factor this time?

[4] Deciding what 'large' means requires looking at the resulting plot and then picking a value for which the plot is informative.

A general remark from David J. C. MacKay, from whom this exercise was adopted, on this material:

> Be warned! A p-value of 0.05 is often interpreted as implying that the odds are stacked about twenty-to-one against the null hypothesis. But the truth in this case is that the evidence either slightly favours the null hypothesis, or disfavours it by at most 2.3 to one, depending on the choice of prior.

Note finally that we have considered only the Bayes factors here. If we had instead computed the posterior odds, we would also need to define a prior odds. Our prior odds would most likely indicate our assumptions that serious coin factories employed by European countries produce fair coins — which would make the case for fairness of the coin even stronger!

Figure 1: A meme a day keeps the p-hacking away.

2 Climate change and Savage-Dickey approximation of Bayes factors

A recent study interviewed a large cohort of meteorologists, consisting of two groups. The first group ($N_E = 4241$) constituted meteorologists in Europe, the second group ($N_W = 412$) consisted of meteorologists of Westeros. For the latter group, $s_W = 277$ report that winter is coming, while for the first group we observe $s_E = 2545$ (s_i is the number of meteorologists reporting that winter is indeed coming). The number of meteorologists reporting that winter is coming is adequately modeled by a binomial distribution with rate parameters θ_E and θ_W, and the group sizes N_E and N_W, respectively. The study is of course interested in the difference $\delta = \theta_E - \theta_W$, i.e. are meteorologists from either realm more likely to predict that winter is coming, or are the groups equal? Let the null hypothesis be that these two groups are equal, and the alternative hypothesis indicates that there is a difference, i.e.:

$$H_0 : \delta = 0$$
$$H_1 : \delta \neq 0 \ .$$

1. (10pts) The model for the difference in binomial rates is identical to a model you have implemented earlier (in practice, you'll see that many different studies use essentially the same models, just like a t-test is applicable to many analyses). Which exercise/assignment was this?

2. (10pts) Write down the generative model that describes this setting. You may assume uniform priors on θ_i ($i \in \{E, W\}$).

3. (20pts) Implement the model in JAGS and obtain an approximation of the posterior distribution $p(\delta \mid s_E, s_W, N_E, N_W)$. Use the template file `wintermightbecoming_template.R`

4. (30pts) As you have seen in the lecture, we can compute the Bayes factor between H_0 and H_1 using the Savage-Dickey density ratio method, provided the null hypothesis is a point estimate hypothesis (i.e. a specific value, like $\delta = 0$ here). But to do so, we also need the prior distribution of δ. Since δ is not a stochastic variable, we cannot simply sample from the corresponding distribution — or can we? Remember that if you do not associate parameters in a model with the observations, the samples you get for these parameters come from their prior. Here we exploit this explicitly. Create additional variables in your script and give them the suffix `.prior`, for δ and both θ's. If you now run your model and monitor both the 'normal' δ (which gives samples from the posterior), as well as the new δ_{prior} parameter, you obtain samples from both the posterior and the prior in one go. Run your model and use the code in `wintermightbecoming_template.R` to make a plot of the collected samples. Explain the what the lines of code indicated with `# ***` do and why they are necessary.

5. (20pts) To compute the Savage-Dickey density ratio, you need the prior and posterior probability densities at $\delta = 0$. You can obtain these from the histograms, but depending on how many bins you used to create the histogram this might be inaccurate (although valid in principle). Instead, the template file demonstrates how you can fit a curve to the histograms, which we can then use to read off the probability density at $\delta = 0$. In general, the probability density at a desired point x may then be obtained using the `dlogspline(`x`, <fitted line object>)` function. Use the available code two create a figure showing the fitted curves superimposed on the histograms, and a plot containing only the curves and black filled circles on these curves at the exact position of $\delta = 0$ (cf. the slides of the lecture)[5].

6. (30pts) Repeat the previous exercise, but now zoom in (by changing `xlim`) so that you can clearly see the probability densities at the points you marked with the two black circles and compute BF_{10} using the Savage-Dickey density ratio. Pay attention to not mix up BF_{10} and BF_{01}! What is the conclusion for this study about meteorologists around Europe and Westeros; do they give different predictions? Be sure to consult the Bayes factor interpretation table if you haven't developed an intuition for this yet.

7. (20pts) The binomial rate difference model is an example of a model where the Bayes factor can be computed analytically. We have

$$BF_{01} = \frac{p(D \mid H_0)}{p(D \mid H_1)} = \frac{\binom{N_E}{s_E}\binom{N_W}{s_W}}{\binom{N_E+N_W}{s_E+s_W}} \frac{(N_E+1)(N_W+1)}{N_E+N_W+1} \quad , \tag{1}$$

with $D = (s_E, s_W, N_E, N_W)$. Note: the binomial coefficient $\binom{n}{k}$ is available in R (`choose(n,k)`), but as the numbers may become exceedingly large, you might run into numerical trouble (e.g. $\binom{N_E}{s_E} = \infty$, according to R). To deal with this, you should actually compute the *logarithm* of (1) and use `lchoose(n,k)`=$\log\binom{n}{k}$ instead, and then take the exponential of your result as a final step, since $BF_{01} = \exp(\log(BF_{01}))$. Compute the analytical Bayes factor using this approach and compare it with your MCMC & Savage-Dickey density ratio approximation. Although the result should be in the same general ballpark, it's probably not exactly identical. Why is this?

8. (20pts) Look up the guidelines for a proper Bayesian report in the lecture material and conclude this study according to those guidelines.

3 Modeling time series data

For classical null hypothesis significance testing, it is relatively straightforward to come up with a Bayesian analysis that answers the same question by either estimating the relevant parameter, or by translating the hypotheses into generative models and doing model comparison. In fact, the latter can also be applied to more complicated settings where a frequentist alternative would be more difficult to construct. In this exercise we'll examine such a case. The first models also have

[5] See the `points` function and its documentation

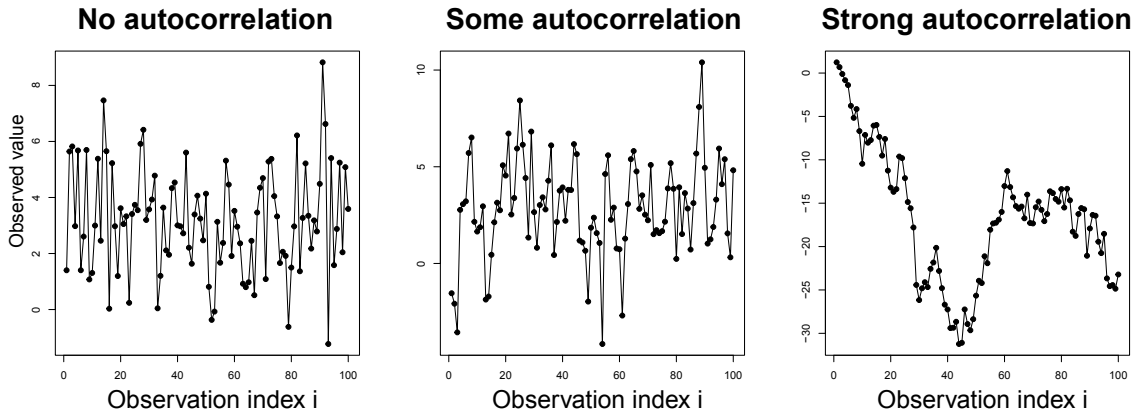

Figure 2: From left to right: data with no autocorrelation, some autocorrelation and strong autocorrelation. For the leftmost plot, the data ordering seems arbitrary, while for the rightmost plot we can clearly see that successive points have similar values.

equivalent frequentist *tests* (although that is not the same as parameter estimation, as you now), but as the models become more complex, Bayesian modeling is far more intuitive.

3.1 A time series model with correlation observations

So far, we have always assumed that individual data points were *independent* of each other. In reality, there are many cases in which this assumption is not true. For example, the stock market value for a particular currency is *autocorrelated*, which means data point y_i is usually close to y_{i-1}. Other examples include neuroimaging data, weather phenomena, sound waves and many, many, others, including the weight gain exercise you saw earlier in this course.

We can look for an autocorrelation effect by creating two models, one with autocorrelation and one without, and then applying model comparison to decide between the two. Consider our null model m_0, defined as follows:

$$\begin{aligned} \tau &\sim \text{Gamma}(0.1, 0.1) \\ \sigma \mid \tau &= 1/\sqrt{\tau} \\ \mu &\sim \text{Gaussian}(0, 1) \\ y_i \mid \mu, \tau &\sim \text{Gaussian}(\mu, \tau), i = 1, \ldots, n \ , \end{aligned}$$

with n the number of observations. This is essentially a linear regression model with an absent slope (i.e. $w_1 = 0$ and $w_0 = \mu$). In this model, each observation has the *same* expectation, $\mathbb{E}[y_i] = \mu$. This means that individual observations have no direct relationship (only indirectly through μ). For example, the heights of students could be modeled this way (provided we have subtracted the mean of the heights so that they are 0 on average). A basic alternative model m_1 is as follows:

$$\begin{aligned} \tau &\sim \text{Gamma}(0.1, 0.1) \\ \sigma \mid \tau &= 1/\sqrt{\tau} \\ \mu &\sim \text{Gaussian}(0, 1) \\ y_1 \mid \mu, \tau &\sim \text{Gaussian}(\mu, \tau) \\ y_i \mid \mu, \tau, y_{i-1} &\sim \text{Gaussian}(y_{i-1}, \tau), i = 2, \ldots, n \ , \end{aligned}$$

in which each observation is similar to the *previous observation*, i.e. $\mathbb{E}[y_i] = y_{i-1}$, with some standard deviation determined by σ. In such a model, successive observations have similar values, which corresponds to a 'smoother' series of data. Examples of the two extremes are shown in Fig. 2.

In the template file `timeseries_template` you'll find a simple data set with $n = 100$ observations as well as some code to help you.

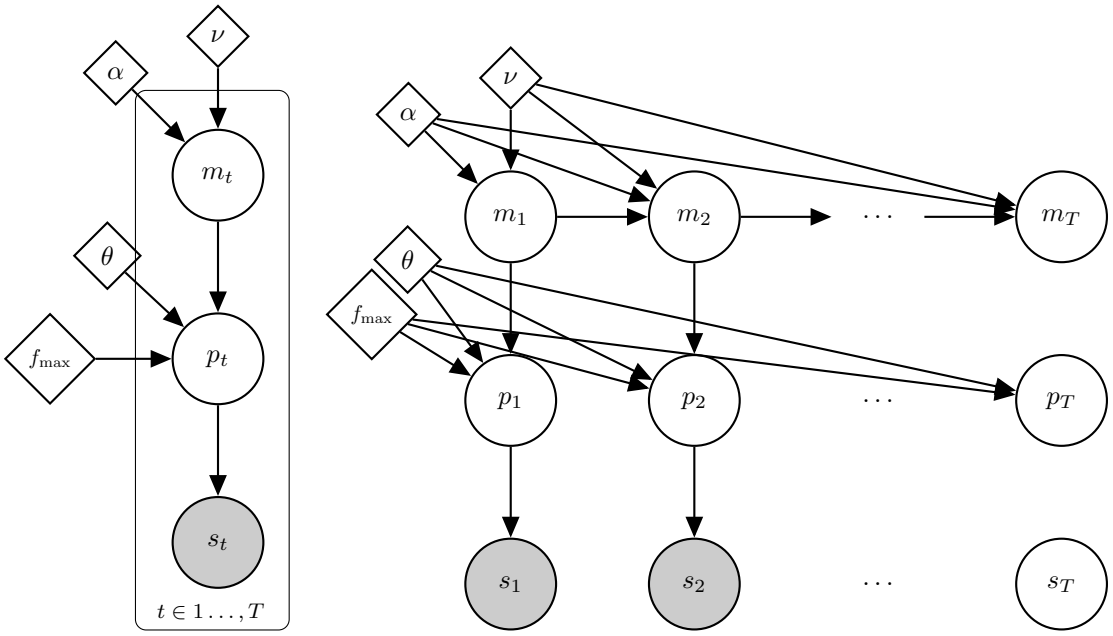

Figure 3: Two models for neuronal spiking behaviour based on the neuron's membrane potential. On the left, each membrane potential is independent of its predecessors. On the right, each membrane potential m_t depends on the previous potential m_{t-1}. Note that in the latter model, plate notation was omitted to make explicit that the first observation is treated different than the successive observations.

1. (10pts) The second model appears more complex, but appearances can be deceiving. How many parameters (and which ones) need to be estimated in each of the models?

2. (40pts) Load the first dataset and implement both models separately and plot the expectations of $\mathbb{E}[y_i]$, for $i = 1, \ldots, n$. You can use the function `rep(x, n)` to create a vector containing n repetitions of x. You can furthermore use the function `plot_sd(vector_of_means, vector_of_sigmas)` to plot the amount of noise around this expectation, as an *interval* around the expected value. Which model do you think models the data best, and why?[6]

3. (30pts) Implement the Bayesian model comparison between the two alternatives using an additional hierarchical categorical variable m as you've seen before. What is the Bayes factor?

3.2 An example from computational neuroscience

Rather than having *observations* themselves be correlated, we often encounter that *latent variables* are correlated, and our observations follow from these variables. In this exercise we examine two models, one without and one with this correlation between latent variables. The model is used to describe *spike trains*, i.e. the discrete events of a neuron firing, which is a function of the membrane potential of that neuron. What is interesting is that each of the parameters in this model have actual biophysical meaning. This demonstrates that the way we think about biological mechanisms can often be expressed almost directly in terms of a Bayesian generative model.

The first 'baseline' model is as follows, with T the total number of time points (see also Fig. 3

[6]Tip: if you want to obtain the mean of several columns of a matrix in R, you can use the function `colMeans`.

for the graphical model):

$$\begin{aligned}
\alpha &\sim \text{beta}(10, 10) \\
\nu &\sim \text{Gamma}(1, 2) \\
f_{\max} &\sim \text{Uniform}(0, 1) \\
\theta &\sim \text{Gaussian}(2, 2) \\
m_t \mid \alpha, \nu &\sim \text{Gaussian}(\alpha, \nu), \quad t = 1, \ldots, T \\
p_t \mid f_{\max}, m_t, \theta &= f_{\max}\sigma(m_t + \theta), \quad t = 1, \ldots, T \\
s_t \mid p_t &\sim \text{Bernoulli}(p_t), \quad t = 1, \ldots, T \ ,
\end{aligned}$$

with $\sigma(x) = 1/(1+\exp(-x))$. The model should be read as follows. At each time point t, a neuron either spikes ($s_t = 1$) or it does not ($s_t = 0$). This depends on the probability p_t, which is scaled to have a maximum probability of $f_{\max}$. The probability p_t further depends on the baseline θ and the membrane potential m_t at time t. The membrane potential itself has a baseline of α and a precision (i.e. inverse variance) of ν. The priors on the hyperparameters are actually biologically plausible — they are based on previous studies[7]. For example, we are quite confident that $\alpha \approx 0.5$.

Typically, we do not observe the membrane potentials themselves, but the spike activations of the neuron, $\mathbf{s}$. However, here the data comes from a simulation, so $\mathbf{m}$ is also available. This allows us to estimate how well the model does in terms of recovering the actual underlying parameters.

With a very small change we can create autocorrelation in this model, but this time in the latent variable $\mathbf{m}$, rather than in the observations. Consider the extended model (see also Fig. 3 for the graphical model):

$$\begin{aligned}
\alpha &\sim \text{beta}(10, 10) \\
\nu &\sim \text{Gamma}(1, 2) \\
f_{\max} &\sim \text{Uniform}(0, 1) \\
\theta &\sim \text{Gaussian}(2, 2) \\
m_1 \mid \nu &\sim \text{Gaussian}(\alpha, \nu) \\
p_1 \mid f_{\max}, m_1, \theta &= f_{\max}\sigma(m_1 + \theta) \\
s_1 \mid p_1 &\sim \text{Bernoulli}(p_1) \\
m_t \mid \alpha, \nu, m_{t-1} &\sim \text{Gaussian}(\alpha m_{t-1}, \nu), \quad t = 2, \ldots, T \\
p_t \mid f_{\max}, m_t, \theta &= f_{\max}\sigma(m_t + \theta), \quad t = 2, \ldots, T \\
s_t \mid p_t &\sim \text{Bernoulli}(p_t), \quad t = 2, \ldots, T \ .
\end{aligned}$$

The substantial difference is only in the definition of m_t, which now depends on the previous membrane potential m_{t-1}, scaled by α. Because m_0 does not exist, we have to specifically define the m_1 case, but this is straightforward.

To get a feel for the data, use the code in `spikes_template.R` to plot both the spikes and the membrane potential of the neuron. You should see that spikes are more likely when the membrane potential is high — but spikes can also occur when the potential is low, although rarely.

1. (30pts) Implement the first model in JAGS and plot the expected posterior membrane potential. The code in the template file allows you to plot this together with an estimate of the variability in the estimation, as well as with the ground truth membrane potential and the observed spikes. Use the template file `spikes_template.R`.

2. (10pts) Did the model recover the ground truth membrane potential well? At which time points/range is the approximation most wrong?[8] Compute the correlation between the real membrane potential and the posterior expectation of your approximation.

3. (30pts) Now implement the extended model and repeat the questions from before. Does the correlation between the true and estimated membrane potential increase, when using the more advanced model?

[7] Roughly speaking.

[8] Be aware that the model has to learn a *function* based only on a few discrete events and some plausible prior distributions. This is a difficult task, in terms of the limited amount of data available to solve the puzzle. There's no need for disappointment if your posterior expectation is not exactly the same as the ground truth. As a sanity check, you should see that when there is a spike, the membrane potential should be raised.

4. (30pts) Implement the model comparison between the two models, using an additional categorical variable[9]. What is the Bayes factor for this comparison?

Good to know: in this exercise you have worked with simulated data for a single trial of a single neuron. In a real study, multiple trials would have been available, and neurons would influence each other as well. Furthermore, we have assumed that the variance in our estimate was stationary (the same everywhere), while in reality when a few spikes occur in a row, we can be more certain about a raised membrane potential. But although this exercise simplifies a few things, very similar models are actually state-of-the-art in computational neuroscience and AI!

[9]For the categorical variable that determines the model you need another name than m, as we use that for membrane potential here. But don't use the word `model`, because that is a keyword in JAGS!