CGS698C: Bayesian models & data analysis

End-semester Exam, 2023-24-II

27.04.2024

Instructions

Time: 180 minutes

Answer all seven questions. Be brief in your responses, and answer the question asked. The maximum number of points is 110.

You can refer to the reference material provided with this question paper some useful probability distributions.

Instructions on how to answer the questions:

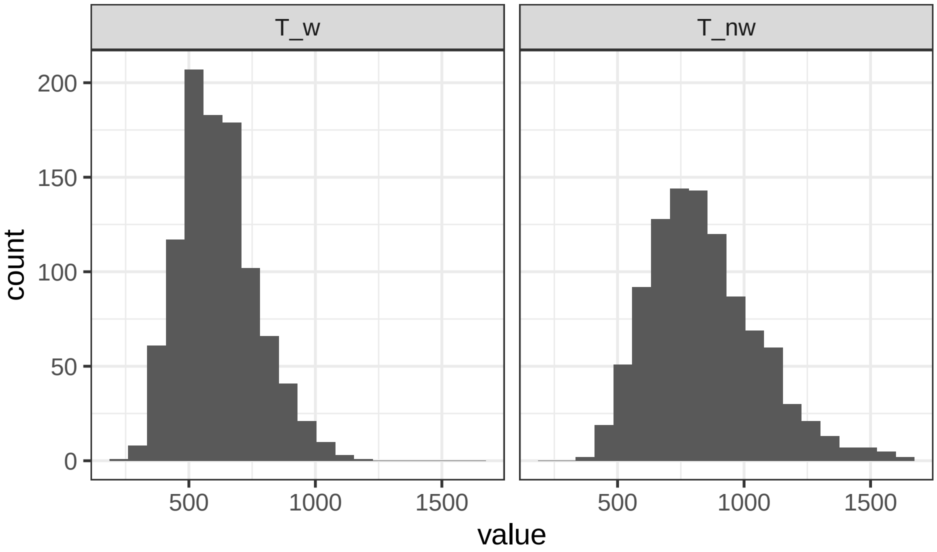
- When you are asked to write the likelihood and prior assumptions, write the assumptions using a set of mathematical statements. For example, $y_i \sim Normal(\mu, \sigma)$.
- When you are asked to described the model, write the likelihood and prior assumptions using a set of mathematical statements as mentioned above.
- When you are asked to write the code, you can write the psuedocode or R/python code.
- Always show your work so that I can give partial credit for answers that are in the right direction.
- When I ask you to write text, please keep your answers brief and to the point. I will not give points for the number of words used but by whether you answer the question. Few words that are to the point will get you full marks if you answer the question.
- When I ask you to draw a graph, you only need to draw a rough sketch on your answer sheet.
- You will get 10 extra minutes to read the question paper carefully.

Question 1

[10 points]

In a visual word recognition experiment, a participant has to recognize whether a string shown on the screen is a meaningful word (e.g., "book") or a non-word (e.g., "bktr"). The participant is asked to answer "yes" if the shown string is a meaningful word, and "no" if it is a meaningless non-word. Suppose a participant is shown n words and n non-words on the screen one by one and you record the recognition time for each word/non-word.

Say, T_w is the vector of word recognition times, and T_{nw} is the vector of non-word recognition times. The distribution of word recognition times and non-word recognition times is given in the following plot.



You ask the following question:

Does it take longer to recognize the non-words compared to the words?

More formally, you test the **lexical-access hypothesis** that the mean recognition time for the non-words is longer than the mean recognition time for the words.

1. Describe a model with reasonable likelihood and prior assumptions to test the above lexical-access hypothesis.

   (The model should be described mathematically using a set of likelihood and prior assumption statements.)

2. Write a psuedocode or R/python code to fit your lexical-access model to the data.

Question 2

[30 points]

You are given 5 independent and identically distributed datapoints that are assumed to come from a Binomial distribution with sample size 20 and probability of success θ:

10, 15, 14, 11, 14

Model:

Let y_i be the i^{th} datapoint,

$$y_i \sim Binomial(n = 20, \theta)$$

$$\theta \sim Beta(1, 1)$$

You can analytically derive the posterior distribution of θ. When the likelihood is Binomial and the prior on θ is a Beta distribution, the posterior will be a Beta distribution.

$$\theta | y \sim Beta(\alpha, \beta)$$

Use the above information to solve the following exercises.

1. Graph the analytically-derived posterior distribution of θ. [5 points]

   (You can draw a rough approximate graph on paper; it should resemble the actual posterior in the mean value and the variance.)

2. Suppose you use the Markov chain Monte Carlo (MCMC) method to estimate the posterior distribution of θ. Write a psuedocode or R/python code to implement a MCMC sampler for the above model. [10 points]

3. Suppose in your MCMC sampling, the current state of Markov chain has a value 0.3, i.e., $\theta_i = 0.3$.

   Now, you draw a new proposal θ^* from the proposal density such that

   $$\theta^* \sim N(\theta_i, \sigma = .05)$$

   where σ is the step-size parameter.

   The probability of accepting the a new proposal θ^* is given by

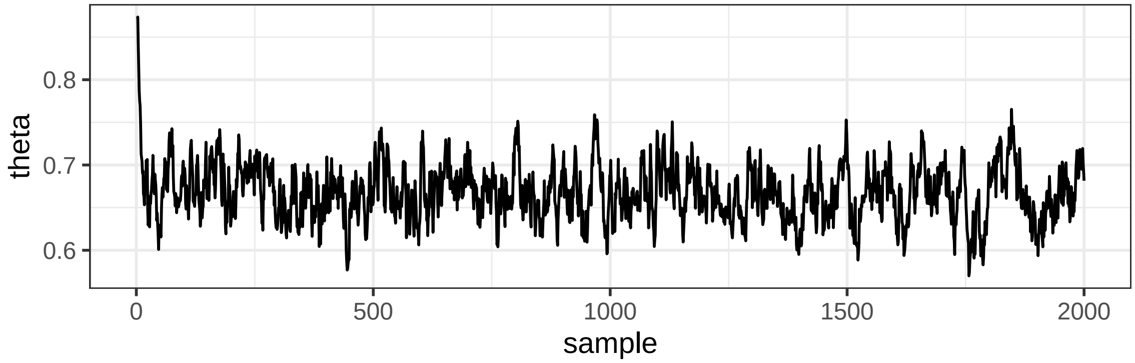
   $$P(\theta^* \rightarrow \theta_{i+1}) = min\Big(1, \frac{\mathcal{L}(\theta^*|y) \cdot p(\theta^*)}{\mathcal{L}(\theta_i|y) \cdot p(\theta_i) \cdot}\Big)$$

   where $\mathcal{L}(.|y)$ represents the likelihood, $p(.)$ is the prior density.

   What is the probability of accepting a new proposal θ^* when it has the following values? [5 points]

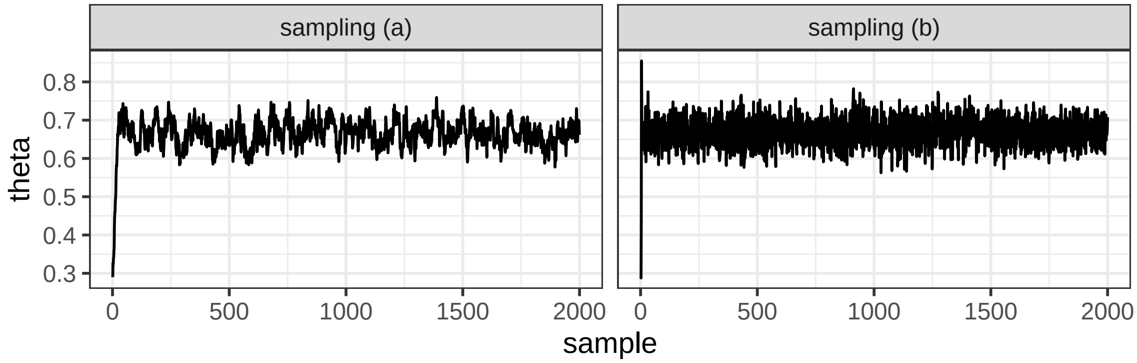
   - $\theta^* = 0.7$
   - $\theta^* = 0.15$

4. The following graph shows the progress of the markov chain. Do you neeed to increase the step size parameter? Why? [2 points]



5. In MCMC, the acceptance rate is the proportion of proposals that were accepted out of total proposals evaluaed during sampling. Suppose you accepted N samples and rejected r number of samples, the acceptance rate is $\frac{N}{N+r}$. [5 points]

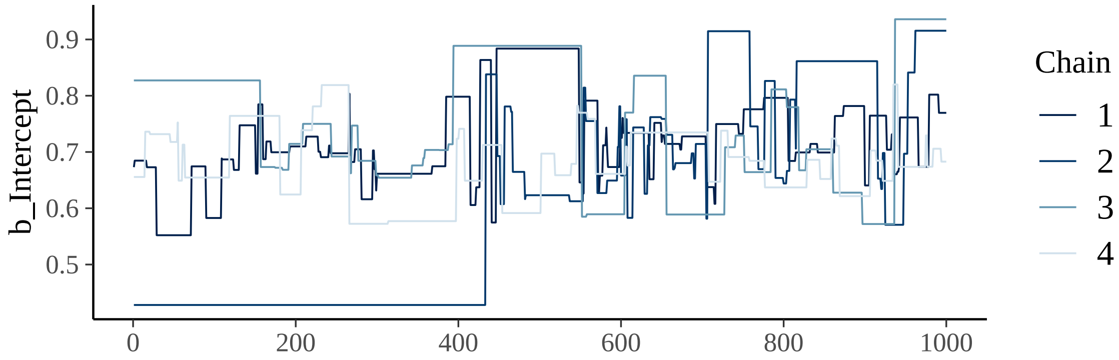
   - What is the relationship between the acceptance rate the step-size parameter σ?
   - Look at the markov chains (a) and (b) that represent the sampling under different values of step-size σ_a and σ_b respectively. Among (a) and (b), which sampling has larger acceptance rate?



6. Now suppose, we use the brms package, which is based on Hamiltonian Monte Carlo, to estimate the posterior distribution of θ. We use the following piece of code to fit the model. [3 points]

```r
mfit <- brm(formula, data,family = binomial(),
            warmup = 1000,iter = 2000, chains = 4,
            control=list(adapt_delta=0.15))
```

The markov chains show poor convergence (see the figure below). What are the possible reasons that might cause this poor convergence? What would you change in the above code to improve the sampling?



Question 3

[20 points]

We have reading time data (in milliseconds) from a repeated measures reading study where the predictability of a particular word in a sentence was manipulated: the word was either predictable (coded as +1) or not predictable (coded as -1). A researcher hypothesizes the reading time will be faster when the word is predictable (when *pred* is +1).

The data was collected from 100 subjects and 36 items (sentence variants). The data-frame looks like this:

```
##   subj item pred     rt
## 1    1    1    1 296.77
## 2    1    3    1 396.88
## 3    1    5    1 481.90
## 4    1    7    1 328.97
## 5    1    9    1 261.22
## 6    1   11    1 307.70
```

The researcher wants to fit a hierarchical model with by-subjects varying intercept and varying slopes. The following code was used to fit the model:

```
priors <- c(set_prior("normal(6, 1.5)", class = "Intercept"),
            set_prior("normal(0, 1)", class = "b",coef = "pred"),
            set_prior("normal(0, 1)", class = "sd"),
            set_prior("normal(0, 1)", class = "sigma"))

m1 <- brm(rt ~ 1+pred + (1+pred||subj),
          data=pred_dat,
          family=lognormal(),prior=priors)
```

Here is the output of the model fit. It shows the summary of posterior estimates for different parameters. The parameters `sd(Intercept)` and `sd(pred)` represent the standard deviation of subject-level (mean) intercepts and (mean) slopes respectively. For example, if population-level mean slope is β, then the mean slope for the j^{th} subject can be given by $\beta_j \sim Normal(\beta, sd(pred))$.

```r
summary(m1)
```

```
##  Family: lognormal
##   Links: mu = identity; sigma = identity
## Formula: rt ~ 1 + pred + (1 + pred || subj)
##    Data: pred_dat (Number of observations: 3600)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Group-Level Effects:
## ~subj (Number of levels: 100)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.01      0.01     0.00     0.02 1.00     1687     1679
## sd(pred)          0.01      0.01     0.00     0.03 1.00     1203     1949
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     5.90      0.00     5.89     5.91 1.00     6336     3151
## pred         -0.10      0.00    -0.11    -0.09 1.00     5629     3185
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     0.25      0.00     0.24     0.25 1.00     6912     2938
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

1. Based on the above output, what can you infer about the researcher's hypothesis? Does the posterior estimates support the researcher's hypothesis? [3 points]
2. From the above brms code, reconstruct the hierarchical model in terms of mathematical statements about the likelihood and the priors. (You can write the hierarchical model using sampling statements like $y_i \sim \mathcal{N}(\mu_i, \sigma)$, etc.) [10 points]
3. Based on the posterior estimates shown in the output above, answer the following: [7 points]
   (a) Do individual subjects differ in their average reading speed?
   (b) Out of total 50 subjects, approximately how many subjects would show positive estimate for the effect of *pred* on log reading times? Provide a rough approximation with a justification.

Question 4

[10 points]

Answer the following questions. Be as brief as possible in your answers.

1. Briefly state the difference between Markov Chain Monte Carlo (MCMC) and importance sampling methods of parameter estimation. [4 points]

2. What are the drawbacks of using Akaike Information Criteria (AIC) and Deviance Information Criteria (DIC) for evaluating a model's predictive performance? [2 points]
3. How the two model comparison methods, the cross validation and the Bayes factor, are different from each other? [4 points]

Question 5

[20 points]

You are given data from a corpus study on **dependency length minimization (DLM)**. The DLM hypothesis says that the average dependency length is shorter in natural languages compared to an artificial language.

The following dataframe shows the relevant variables; the `language_type` (natural/artificial) is encoded as $+1$ (for natural) and -1 (for artificial), `DL` represent dependency length, `treebank` represents a source natural language data. The DLM hypothesis predicts that the effect of `language_type` on `DL` will be negative, i.e., natural languages would have shorter DL, on average.

```
##   treebank language_type DL
## 1        1            -1  1
## 2        1             1  2
## 3        1            -1  2
## 4        1             1  3
## 5        1            -1  1
## 6        1             1  5
```

Suppose a researcher wants to quantify evidence for the DLM hypothesis given these data. They use Bayes factor analysis to compare models with and without the effect of language_type.

```
priors_dlm <- c(set_prior("normal(1, 1)", class = "Intercept"),
            set_prior("normal(0, p)", class = "b",ub=0),
            set_prior("normal(0, 1)", class = "sd"),
            set_prior("normal(0, 1)", class = "sigma"),
            set_prior("lkj(2)", class = "cor"))

priors_null <- c(set_prior("normal(1, 1)", class = "Intercept"),
            set_prior("normal(0, 1)", class = "sd"),
            set_prior("normal(0, 1)", class = "sigma"),
            set_prior("lkj(2)", class = "cor"))

m.dlm <- brm(DL ~ 1+language_type + (1+language_type|treebank),
          data=dl_dat,
          family=poisson(link="log"),
          prior=priors_dlm,
          iter=20000,warmup=2000)

m.null <- brm(DL ~ 1 + (1+language_type|treebank),
          data=dl_dat,
          family=poisson(link="log"),
          prior=priors_null,
          iter=20000,warmup=2000)

library(bridgesampling)
lkl.dlm <- bridge_sampler(m.dlm)
lkl.null <- bridge_sampler(m.null)
bf <- bayes_factor(lkl.dlm,lkl.null)
```

The Bayes factors were computed under different priors on the effect of `language_type`. Here is the output of the Bayes factor analysis:

```
##                    Prior    BF
## 1 Normal(0,0.005), ub=0 15.23
## 2  Normal(0,0.01), ub=0 18.57
## 3  Normal(0,0.05), ub=0 22.67
## 4   Normal(0,0.1), ub=0 21.78
## 5   Normal(0,0.5), ub=0 10.27
## 6     Normal(0,1), ub=0  2.45
## 7     Normal(0,5), ub=0  0.09
```

1. Given the above Bayes factor results, what would you infer about the DLM hypothesis? Is there sufficient evidence for the hypothesis? [5 points]
2. Now suppose you want to collect evidence for the DLM against the *anti-DLM hypothesis* that says average DL is longer in natural languages compared to artifical languages. What changes would you make in the above brms code to quantify Bayes factor in favor of DLM hypothesis against the anti-DLM hypothesis? [5 points]
3. Describe the anti-DLM model[1] as a set of likelihood and prior assumptions (using mathematical statements). [10 points]

Question 6

[10 points]

You are given 10 datapoints: $y_1, y_2, y_3, ..., y_{10}$.

Researcher A argues that these datapoints are generated by Model $\mathcal{M}1$ with likelihood $\mathcal{L}_1(\theta|y)$ and prior $p(\theta)$.

But Researcher B argues that these datapoints are generated by Model $\mathcal{M}2$ whose likelihood is $\mathcal{L}_2(\phi|y)$ and the prior is $p(\phi)$.

1. How would you compare the models $\mathcal{M}1$ and $\mathcal{M}2$ using leave-one-out cross validation? You can explain your answer either using a mathematical description of the method or using a piece of psuedocode or R/python code.

Question 7

[10 points]

Attempt either question 7a or 7b.

Question 7a

In an experiment, two dice a and b are rolled simultaneously. Rolling a dice can give you six possible outcomes: 1, 2, 3, 4, 5, 6.

Suppose X_a represents the outcome on the first dice and X_b represent the outcome on the second dice. Consider the following events:

$E_1 : X_a + X_b \geq 11$

$E_2 : X_a = 5$

$E_3 : X_b = 5$

$E_4 : X_a = X_b$

$E_5 : X_a + X_b < 3$

You are given that:

$P(E_1) = 0.125$

$P(E_2) = 0$

$P(E_3) = 0$

$P(E_4) = 0.208$

$P(E_5) = 0.042$

(Note: the dice is not fair; you **cannot** assume that the six outcomes have equal probabilities.)

1. Find the probability of an event where 6 appears on both the dice, i.e., the outcome is (6,6).
2. Find the probability of an event where the outcome is either (2,2), (3,3), or (4,4).
3. Find the probability $P(E_2 \cup E_4)$.

OR

Question 7b

1. Suppose a random variable X has binomial distribution with probability mass function $f(x) = \frac{n!}{k!(n-k)!}p^k(1-p)^k$.

   Prove that the expected value of X is equal to np.

[1] the model that implements the anti-DLM hypothesis

(The expected value of a discrete random variable X is given by: $E(X) = \sum_{i=1}^{n} x_i f(x_i)$ where $\{x_1, x_2, x_3, ..., x_n\}$ is the set of all possible values that the random variable X can take.)

2. A continuos random variable X has the following probability density function

   $f(x) = \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!} x^{\alpha-1}(1-x)^{\beta-1}$

   The domain of the function is $[0, 1]$ i.e., the random variable X can take any real number value between 0 and 1.

   Find the mode of the variable X given that $\alpha > 1$ and $\beta > 1$.

   The mode is the value of X at which the first derivative of the probability density function becomes zero, $\frac{d}{dx} f(x) = 0$.

Reference material

Probability distributions

	Type of Random variable	**Name of the distribution**	**Probability density function (PDF)** or **Probability mass function (PMF)**
1	Discrete	Binomial	PMF: $f(k; n, p) = \frac{n!}{k!(n-k)!} p^k (1-p)^k$
2	Discrete	Poisson	PMF: $f(k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$ where $\lambda > 0$
3	Continuous	Normal	PDF: $f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
4	Continuous	Beta	PDF: $f(x; \alpha, \beta) = \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!} x^{\alpha-1}(1-x)^{\beta-1}$
5	Continuous	Gamma	PDF: $f(x; \alpha, \beta) = \frac{\beta^\alpha}{(\alpha-1)!} x^{\alpha-1} e^{-\beta x}$