Exam 2 (A2)

Class: Bayesian Statistics
Student: Lucas Machado Moschen
Instructor: Luiz Max Carvalho

02/06/2021

Turn in date: until 09/06/2021 at 23:59h Brasilia Time.

- Please read through the whole exam before starting to answer;
- State and prove all non-trivial mathematical results necessary to substantiate your arguments;
- Do not forget to add appropriate scholarly references *at the end* of the document;
- Mathematical expressions also receive punctuation;
- You can write your answer to a question as a point-by-point response or in "essay" form, your call;
- Please hand in a single, **typeset** (LaTeX) PDF file as your final main document. Code appendices are welcome, *in addition* to the main PDF document.
- You may consult any sources, provided you cite **ALL** of your sources (books, papers, blog posts, videos);
- You may use symbolic algebra programs such as Sympy or Wolfram Alpha to help you get through the hairier calculations, provided you cite the tools you have used.
- The exam is worth 100 marks.

Background

This exam covers applications, namely estimation, prior sensitivity and prediction. You will need a working knowledge of basic computing tools, and knowledge of MCMC is highly valuable. Chapter 6 in Robert (2007) gives an overview of computational techniques for Bayesian statistics.

Inferring population sizes – theory

Consider the model

$$x_i \sim \text{Binomial}(N, \theta),$$

with **both** N and θ unknown and suppose one observes $\boldsymbol{x} = \{x_1, x_2, \ldots, x_K\}$. Here, we will write $\xi = (N, \theta)$.

a) (10 marks) Formulate a hierarchical prior (π_1) for N, i.e., elicit F such that $N \mid \alpha \sim F(\alpha)$ and $\alpha \sim \Pi_A$. Justify your choice;

Solution. The only restriction prior to the data we have is that $N \in \mathbb{Z}^+$. Therefore we are looking for a distribution in this space. We can follow the assumption made by Raftery (1988) where $N \sim \text{Poisson}(\mu)$. Other approach I will use is to consider $N \sim \text{Geometric}(\nu)$.

The first approach has the advantage in calculations. When

$$x_i | N, \theta \sim \text{Binomial}(N, \theta)$$

and

$$N \sim \text{Poisson}(\mu),$$

we have that $x_i \sim \text{Poisson}(\mu \cdot \theta)$, as proved in Appendix A. And we can transform prior information about the mean of x_i in probability distributions for $\mu \cdot \theta$.

The second approach serves as a comparative. The geometric distribution is a simple distribution used for discrete survival distribution, but it is interesting because of its simplicity and the compact domain of the parameter since $\nu \in [0, 1]$. From that, we can build a model that caries the correlation between ν and θ whenever necessary in a direct way: we transform the variables to logistic space and defines a bivariate normal distribution.

Priors to the hyperparameters

Now we shall define the priors to the hyperparameters. Define $\lambda = \mu \cdot \theta$. Given that it is more likely that previous research made statements about x_i, we use the same idea as Raftery considering the prior over (λ, θ). I will assume they are independent from now on with

$$\lambda \sim \text{Gamma}(\alpha, \beta),$$

and

$$\theta \sim \text{Beta}(a, b).$$

This simplification may be carried, given that the data brings information to the correlation, and it simplifies the math a lot. The other approach will have two settings. First I suppose ν and θ independents with

$$\nu \sim \text{Beta}(\alpha_1, \beta_1)$$

and

$$\theta \sim \text{Beta}(\alpha_2, \beta_2).$$

I choose the Beta distribution because it has a flexible shape with a good intuition behind it. Other point is that the Beta distribution forms a conjugate family for the Geometric distribution. Another set up is to consider the correlated case. We do it in the following way:

$$\begin{pmatrix} \text{logit}(\nu) \\ \text{logit}(\theta) \end{pmatrix} \sim \text{Normal}(\eta, \Sigma).$$

This choice is intrinsically linked to the fact the the normal distribution is a good approximation to a series of events, and it has a very good interpretation of the parameters. The problem with this approach is that it is harder to codify prior information. We necessarily need information about N, θ, and how they relate.

From these three approaches, I will call these approaches in the text *Raftery approach*, *Geometric and independent approach*, and *Geometric and correlated approach*, respectively.

b) (5 marks) Using the prior from the previous item, write out the full joint posterior kernel for all unknown quantities in the model, $p(\xi \mid \boldsymbol{x})$. *Hint:* do not forget to include the appropriate indicator functions!;

Solution. For the Geometric and correlated approach, it may be impossible to find a simple expression.

Generally speaking, by Bayes' Theorem,

$$\begin{aligned} p(\xi|\boldsymbol{x}) &\propto l(\xi|\boldsymbol{x}) \cdot \pi(\xi) \\ &= \left(\prod_{i=1}^{n} \binom{N}{x_i} \theta^{x_i} (1-\theta)^{N-x_i} \mathbb{1}(N \geq x_i) \right) \cdot \pi(\xi) \\ &= \left(\prod_{i=1}^{n} \binom{N}{x_i} \right) \theta^{S} (1-\theta)^{nN-S} \cdot \pi(\xi) \cdot \mathbb{1}(N \geq x_{\max}), \end{aligned} \tag{1}$$

where $S = \sum_{i=1}^{n} x_i$. We shall derive for each case the prior $\pi(\xi)$.

(1) **Raftery**:

$$\begin{aligned} \pi(\xi) &= \int_0^\infty \pi(\xi, \lambda)\, d\lambda \\ &= \int_0^\infty \pi(N|\theta, \lambda)\pi(\theta, \lambda)\, d\lambda \\ &= \int_0^\infty \frac{e^{-\lambda/\theta}(\lambda/\theta)^N}{N!} \pi(\lambda)\pi(\theta)\, d\lambda \\ &\propto \int_0^\infty \frac{e^{-\lambda/\theta}(\lambda/\theta)^N}{N!} \lambda^{\alpha-1} e^{-\beta\lambda} \theta^{a-1} (1-\theta)^{b-1} \mathbb{1}(0<\theta<1)\, d\lambda \\ &= \frac{\theta^{a-1-N}(1-\theta)^{b-1}}{N!} \mathbb{1}(0<\theta<1) \int_0^\infty \lambda^{\alpha+N-1} e^{-(\beta+1/\theta)\lambda}\, d\lambda \\ &= \frac{\theta^{a-1-N}(1-\theta)^{b-1}}{N!} \mathbb{1}(0<\theta<1) \cdot \frac{\Gamma(\alpha+N)}{(\beta+1/\theta)^{\alpha+N}}, \end{aligned}$$

since the integrand is the kernel of a gamma distribution. Therefore, rewriting,

$$\pi(\xi) \propto \frac{\Gamma(\alpha+N)\theta^{a-1-N}(1-\theta)^{b-1}}{(\beta+1/\theta)^{\alpha+N} N!} \mathbb{1}(0<\theta<1) \tag{2}$$

and

$$p(\xi|\boldsymbol{x}) \propto \left(\prod_{i=1}^{n} \binom{N}{x_i} \right) \frac{\Gamma(\alpha+N)\theta^{a+S-1-N}(1-\theta)^{b+nN-S-1}}{(\beta+1/\theta)^{\alpha+N} N!} \mathbb{1}(0<\theta<1, N \geq x_{\max}) \tag{3}$$

(2) **Geometric and independent**:

$$
\begin{aligned}
\pi(\xi) &= \int_0^1 \pi(\xi, \nu)\, d\nu \\
&= \int_0^1 \pi(N|\nu, \theta)\pi(\nu)\pi(\theta)\, d\nu \\
&\propto \theta^{\alpha_2 - 1}(1-\theta)^{\beta_2 - 1}\mathbb{1}(0 < \theta < 1) \int_0^1 (1-\nu)^N \nu \cdot \nu^{\alpha_1 - 1}(1-\nu)^{\beta_1 - 1}\, d\nu \\
&= \theta^{\alpha_2 - 1}(1-\theta)^{\beta_2 - 1}\mathbb{1}(0 < \theta < 1) \int_0^1 \nu^{\alpha_1}(1-\nu)^{N + \beta_1 - 1}\, d\nu \\
&= \theta^{\alpha_2 - 1}(1-\theta)^{\beta_2 - 1}\mathbb{1}(0 < \theta < 1) B(\alpha_1 + 1, N + \beta_1),
\end{aligned}
$$

since the integrand is the kernel of a Beta distribution. Rewriting,

$$
\pi(\xi) \propto B(\alpha_1 + 1, N + \beta_1)\theta^{\alpha_2 - 1}(1-\theta)^{\beta_2 - 1}\mathbb{1}(0 < \theta < 1) \tag{4}
$$

and

$$
p(\xi|\boldsymbol{x}) \propto \left(\prod_{i=1}^n \binom{N}{x_i}\right) B(\alpha_1 + 1, N + \beta_1)\theta^{S + \alpha_2 - 1}(1-\theta)^{nN + \beta_2 - S - 1}\mathbb{1}(0 < \theta < 1, N \geq x_{\max}) \tag{5}
$$

(3) **Geometric and correlated**: Harder case. I have to derive $\pi(\nu, \theta)$ from $\pi(\text{logit}(\nu), \text{logit}(\theta))$. Define $f(x_1, x_2) = (\text{logit}^{-1}(x_1), \text{logit}^{-1}(x_2))$. This is an invertible function with $f^{-1}(y_1, y_2) = (\text{logit}(y_1), \text{logit}(y_2))$. By the change of variables,

$$
\pi(\nu, \theta) = \pi(f^{-1}(\nu, \theta)) \cdot \left|\det\left[\left.\frac{df^{-1}(z)}{dz}\right|_{z=(\nu,\theta)}\right]\right|.
$$

Observe that

$$
\det\left[\left.\frac{df^{-1}(z)}{dz}\right|_{z=(\nu,\theta)}\right] = \det\begin{bmatrix} \frac{d}{d\nu}\text{logit}(\nu) & 0 \\ 0 & \frac{d}{d\theta}\text{logit}(\theta) \end{bmatrix} = \frac{d}{d\nu}\text{logit}(\nu) \cdot \frac{d}{d\theta}\text{logit}(\theta).
$$

By the calculation of the Jacobian, we can join everything

$$
\begin{aligned}
\pi(\xi) &= \int_0^1 \pi(N|\nu, \theta)\pi(\nu, \theta)\, d\nu \\
&= \int_0^1 (1-\nu)^N \nu \pi(\text{logit}(\nu), \text{logit}(\theta))|\,\text{logit}'(\nu) \cdot \text{logit}'(\theta)|\, d\nu \\
&= \int_0^1 (1-\nu)^N \nu \pi(\text{logit}(\nu), \text{logit}(\theta)) \frac{1}{\nu(1-\nu)\theta(1-\theta)}\, d\nu \\
&= \frac{1}{\theta(1-\theta)} \int_0^1 (1-\nu)^{N-1} \pi(\text{logit}(\nu), \text{logit}(\theta))\, d\nu.
\end{aligned}
$$

Let $z = (\text{logit}(\nu), \text{logit}(\theta))$. Therefore,

$$
\pi(\xi) \propto \frac{1}{\theta(1-\theta)} \int_0^1 (1-\nu)^{N-1} \exp\left\{-\frac{1}{2}(z-\eta)^T \Sigma^{-1}(z-\eta)\right\}\, d\nu \tag{6}
$$

and, for $N \geq x_{\max}$,

$$
p(\xi|\boldsymbol{x}) \propto \left(\prod_{i=1}^n \binom{N}{x_i}\right) \theta^{S-1}(1-\theta)^{nN - S - 1} \int_0^1 (1-\nu)^{N-1} \exp\left\{-\frac{1}{2}(z-\eta)^T \Sigma^{-1}(z-\eta)\right\}\, d\nu. \tag{7}
$$

c) (5 marks) Is your model identifiable?

Solution. I shall verify the likelihood of the model is identifiable and it brings the information from the data to the parameter. I say a model is identifiable if the implication

$$P_{\theta_1} = P_{\theta_2} \implies \theta_1 = \theta_2$$

is true for the model P_θ. Suppose that

$$\binom{N_1}{x}\theta_1^x(1-\theta_1)^{N_1-x} = \binom{N_2}{x}\theta_2^x(1-\theta_2)^{N_2-x}$$

for every x between 0 and $\min(N_1, N_2)$. Suppose $N_1 < N_2$. Therefore the equality is false for $x = N_1 + 1$. If $N_1 > N_2$, the equality is also false. We conclude $N_1 = N_2$, and, therefore,

$$\frac{\theta_1}{1-\theta_1} = \frac{\theta_2}{1-\theta_2} \implies \theta_1 = \theta_2.$$

Hence, the binomial model is identifiable. For n observations, take $x_1 = ... = x_n = x$. We will have

$$f(x|\theta_1, N_1)^n = f(x|\theta_2, N_2)^n \implies f(x|\theta_1, N_1) = f(x|\theta_2, N_2)$$

and the prove follows as previously.

Strictly speaking, given that the prior is proper, and hence the posterior is proper, all the parameters are identifiable. The choice of a suitable prior can resolve the non-identifiability Xie and Carlin (2006). We can also see in the three models, the posterior of ξ has additional information of the data, coming from the productory and S.

Remark 1 *Throughout the development of the work, I observed a problem in the practical identifiability. In figure 1, we observe three different combinations of parameters which generate similar likelihoods, despite being mathematically different. This may be problematic when we assume the independence of these parameters.*

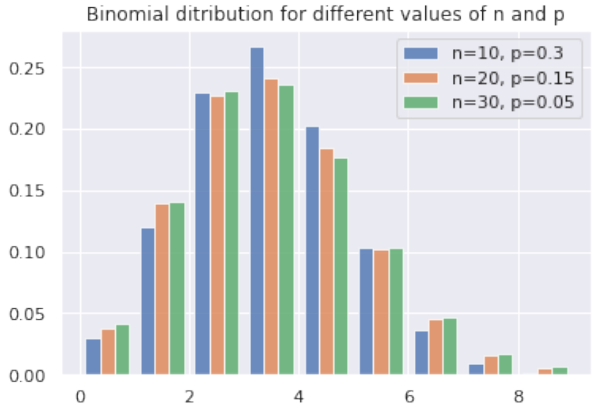

Figure 1: Binomial distribution for different parameters n and p.

d) (5 marks) Exhibit the marginal posterior density for N, $p_1(N \mid \boldsymbol{x})$;

Solution. We separate the marginalized posterior:

(1) **Raftery:** Consider the equation 3 and integrate over θ. For $N \geq x_{\max}$,

$$p_1(N|\boldsymbol{x}) \propto \left(\prod_{i=1}^{n}\binom{N}{x_i}\right)\frac{\Gamma(\alpha+N)}{N!}\int_0^1 \frac{\theta^{a+S-1-N}(1-\theta)^{b+nN-S-1}}{(\beta+1/\theta)^{\alpha+N}}\,d\theta. \tag{8}$$

(2) **Geometric independent:** Consider the equation 5 and integrate over θ. For $N \geq x_{\max}$,

$$\begin{aligned} p_1(N|\boldsymbol{x}) &\propto \left(\prod_{i=1}^{n}\binom{N}{x_i}\right) B(\alpha_1+1, N+\beta_1)\int_0^1 \theta^{S+\alpha_2-1}(1-\theta)^{nN+\beta_2-S-1}\,d\theta \\ &= \left(\prod_{i=1}^{n}\binom{N}{x_i}\right) B(\alpha_1+1, N+\beta_1)B(S+\alpha_2, nN-S+\beta_2). \end{aligned} \tag{9}$$

(3) **Geometric correlated:** Consider the equation 7 and integrate over θ. For $N \geq x_{\max}$,

$$\begin{aligned} p_1(N|\boldsymbol{x}) \propto \left(\prod_{i=1}^{n}\binom{N}{x_i}\right) \int_0^1\int_0^1 \theta^{S-1}(1-\theta)^{nN-S-1}(1-\nu)^{N-1} \\ \times \exp\left\{-\frac{1}{2}(z-\eta)^T\Sigma^{-1}(z-\eta)\right\}\,d\nu\,d\theta. \end{aligned} \tag{10}$$

e) (5 marks) Return to point (a) above and consider an alternative, uninformative prior structure for ξ, π_2. Then, derive $p_2(N \mid \boldsymbol{x})$;

Solution. For an uninformative prior, for instance, we could consider in the Raftery approach $\alpha, \beta \to 0$. When we do that, we obtain

$$\pi(\lambda) \propto \lambda^{-1},$$

an improper prior and $a = b = 1$, what implies θ uniformly distributed on the interval $(0, 1)$. In that regard,

$$p_2(N|\boldsymbol{x}) \propto \left(\prod_{i=1}^{n}\binom{N}{x_i}\right)\frac{\Gamma(N)}{N!}\int_0^1 \frac{\theta^{S-\cancel{N}}(1-\theta)^{nN-S}}{\cancel{\theta^{N}}}\,d\theta = \left(\prod_{i=1}^{n}\binom{N}{x_i}\right)\frac{B(S+1, nN-S+1)}{N}.$$

Simplifying, we have

$$p_2(N|\boldsymbol{x}) \propto \left(\prod_{i=1}^{n}\binom{N}{x_i}\right)\frac{(nN-S)!}{(nN+1)!N}. \tag{11}$$

We need to prove p_2 defines a probability distribution with a correct constant. Observe that,

$$\begin{aligned} \left(\prod_{i=1}^{n}\binom{N}{x_i}\right)\frac{(nN-S)!}{(nN+1)!N} &= \frac{1}{N\prod_{i=1}^{n}x_i!}\frac{\prod_{i=1}^{n}\prod_{j=1}^{x_i}(N-x_i+j)}{\prod_{k=1}^{S+1}(nN-S+k)}\frac{N^{-(S+1)}}{N^{-(S+1)}} \\ &\propto \frac{1}{N^2}\frac{\prod_{i=1}^{n}\prod_{j=1}^{x_i}(1-(x_i-j)/N)}{\prod_{k=1}^{S+1}(n-(S-k)/N)} \end{aligned}$$

given that the numerator has $x_1 + ... + x_n = S$ factors, while the dominator has $S + 1$. Observe that, taking $n \geq 2$, for every i, j, and k if $N \geq S$,

$$1 - (x_i - j)/N < 1$$

and $n - (S-k)/N \geq n - 1 + 1/N > 1$. Therefore, if $N \geq S$

$$\frac{1}{N^2}\frac{\prod_{i=1}^{n}\prod_{j=1}^{x_i}(1-(x_i-j)/N)}{\prod_{k=1}^{S+1}(n-(S-k)/N)} \leq \frac{1}{N^2}.$$

and

$$\sum_{N=x_{\max}}^{\infty}\left(\prod_{i=1}^{n}\binom{N}{x_i}\right)\frac{(nN-S)!}{(nN+1)!N} \leq \sum_{x_{\max}}^{S-1}\left(\prod_{i=1}^{n}\binom{N}{x_i}\right)\frac{(nN-S)!}{(nN+1)!N} + \sum_{N=S}^{\infty}\frac{1}{N^2} < +\infty,$$

what proves $p_2(N|x)$ is a probability distribution given the right constant. If $n = 1$, we have that

$$\binom{N}{x}\frac{(N-x)!}{(N+1)!N} \propto \frac{1}{N(N+1)} \leq \frac{1}{N^2}$$

also convergent.

f) (10 marks) Formulate a third prior structure on ξ, π_3, that allows for the closed-form marginalization over the hyperparameters α – see (a) – and write out $p_3(N \mid \boldsymbol{x})$;

Solution. The structure of the Geometric independent approach allowed the closed-form marginalization over ν and θ, as expressed in equation 9. I will denote from now on p_4 the marginalized distribution of Geometric correlated approach, given by expression 10.

g) (10 marks) Show whether each of the marginal posteriors considered is proper. Then, derive the posterior predictive distribution, $g_i(\tilde{x} \mid \boldsymbol{x})$, for each of the posteriors considered ($i = 1, 2, 3, 4$).

Solution. p_1 (Raftery), p_3 (Geometric independent), and p_4 (Geometric correlated) are proper since the prior is proper, by Bayes' theorem. For a prove, consult the second question from A1 exam Moschen (2021). We proved that the posterior p_2 is proper in exercise (e). Now we derive the posterior predictive distribution

$$g(\tilde{x}|\boldsymbol{x}) = \sum_{N=x_{\max}}^{\infty} \int_0^1 f(\tilde{x}, \boldsymbol{x}|\xi)\pi(\xi) d\,\theta$$

for each of the four priors. Let $x_{n+1} = \tilde{x}$ for notation simplification.

(1) **Raftery:** Consider the prior from expression 2.
Let $S' = S + x_{n+1}$ and $x'_{\max} = \max(x_{\max}, x_{n+1})$.

$$\begin{aligned} g_1(x_{n+1}|\boldsymbol{x}) &\propto \sum_{N=x'_{\max}}^{\infty} \int_0^1 \left\{ \prod_{i=1}^{n+1} \binom{N}{x_i} \right\} \theta^{S'}(1-\theta)^{(n+1)N-S'} \frac{\Gamma(\alpha+N)\theta^{a-1-N}(1-\theta)^{b-1}}{(\beta+1/\theta)^{\alpha+N}N!} d\,\theta \\ &= \sum_{N=x'_{\max}}^{\infty} \left\{ \prod_{i=1}^{n+1} \binom{N}{x_i} \right\} \frac{\Gamma(\alpha+N)}{N!} \int_0^1 \frac{\theta^{a+S'-1-N}(1-\theta)^{(n+1)N-S'+b-1}}{(\beta+1/\theta)^{\alpha+N}} d\,\theta. \end{aligned}$$

(2) **Uninformative:** This case is the same as Raftery, with $a = b = 1$ and $\alpha = \beta = 0$. Then,

$$\begin{aligned} g_2(x_{n+1}|\boldsymbol{x}) &\propto \sum_{N=x'_{\max}}^{\infty} \left\{ \prod_{i=1}^{n+1} \binom{N}{x_i} \right\} \frac{\Gamma(N)}{N!} \int_0^1 \frac{\theta^{S'}\cancel{\theta^{N}}(1-\theta)^{(n+1)N-S'}}{\cancel{(1/\theta)^{N}}} d\,\theta \\ &= \sum_{N=x'_{\max}}^{\infty} \left\{ \prod_{i=1}^{n+1} \binom{N}{x_i} \right\} \frac{1}{N} \int_0^1 \theta^{S'}(1-\theta)^{(n+1)N-S'} d\,\theta \\ &= \sum_{N=x'_{\max}}^{\infty} \left\{ \prod_{i=1}^{n+1} \binom{N}{x_i} \right\} \frac{1}{N} B(S'+1, (n+1)N - S' + 1) \\ &= \sum_{N=x'_{\max}}^{\infty} \left\{ \prod_{i=1}^{n+1} \binom{N}{x_i} \right\} \frac{S'![(n+1)N - S']!}{[(n+1)N+1]!N}. \end{aligned}$$

(3) **Geometric independent:** Consider the prior from expression 4.
Let $B_N = B(\alpha_1 + 1, N + \beta_1)$ to reduce the size of the expression. Then

$$\begin{aligned} g_3(x_{n+1}|\boldsymbol{x}) &\propto \sum_{N=x'_{\max}}^{\infty} \int_0^1 \left\{ \prod_{i=1}^{n+1} \binom{N}{x_i} \right\} \theta^{S'}(1-\theta)^{(n+1)N-S'} B_N \theta^{\alpha_2-1}(1-\theta)^{\beta_2-1} d\theta \\ &= \sum_{N=x'_{\max}}^{\infty} \left\{ \prod_{i=1}^{n+1} \binom{N}{x_i} \right\} B_N \int_0^1 \theta^{S'+\alpha_2-1}(1-\theta)^{(n+1)N+\beta_2-S'-1} d\theta \\ &= \sum_{N=x'_{\max}}^{\infty} \left\{ \prod_{i=1}^{n+1} \binom{N}{x_i} \right\} B(\alpha_1+1, N+\beta_1) B(S'+\alpha_2, (n+1)N+\beta_2-S'). \end{aligned}$$

(4) **Geometric correlated:** Consider the prior expression 6.
Denote $I_{\theta,N} = \int_0^1 (1-\nu)^{N-1} \exp\left\{-\frac{1}{2}(z-\eta)^T \Sigma^{-1}(z-\eta)\right\} d\nu$.

$$g_4(x_{n+1}|\boldsymbol{x}) \propto \sum_{N=x'_{\max}}^{\infty} \int_0^1 \left\{\prod_{i=1}^{n+1} \binom{N}{x_i}\right\} \theta^{S'}(1-\theta)^{(n+1)N-S'} \frac{1}{\theta(1-\theta)} \cdot I_{\theta,N}\, d\theta$$
$$= \sum_{N=x'_{\max}}^{\infty} \left\{\prod_{i=1}^{n+1} \binom{N}{x_i}\right\} \int_0^1 \theta^{S'-1}(1-\theta)^{(n+1)N-S'-1} \cdot I_{\theta,N}\, d\theta$$

h) (5 marks) Consider the loss function

$$L(\delta(\boldsymbol{x}), N) = \left(\frac{\delta(\boldsymbol{x}) - N}{N}\right)^2. \tag{12}$$

Derive the Bayes estimator under this loss.

Solution. The Bayes estimator is

$$\delta^\pi(\boldsymbol{x}) = \arg\min_d \mathbb{E}^\pi[L(d,N)|\boldsymbol{x}] = \arg\min_d \mathbb{E}^\pi\left[\left(\frac{d-N}{N}\right)^2 |\boldsymbol{x}\right].$$

In general terms,

$$\mathbb{E}^\pi\left[\left(\frac{d-N}{N}\right)^2 |\boldsymbol{x}\right] = \mathbb{E}^\pi\left[\left(\frac{d^2}{N^2} - 2d\frac{1}{N} + 1\right)|\boldsymbol{x}\right] = d^2\mathbb{E}^\pi[N^{-2}|\boldsymbol{x}] - 2d\mathbb{E}^\pi[N^{-1}|\boldsymbol{x}] + 1. \tag{13}$$

The minimum argument of a quadratic expression is given By

$$\delta^\pi(\boldsymbol{x}) = \frac{\mathbb{E}^\pi[N^{-1}|\boldsymbol{x}]}{\mathbb{E}^\pi[N^{-2}|\boldsymbol{x}]} = \frac{\sum_{N=x_{\max}}^{\infty} N^{-1} p_i(N|\boldsymbol{x})}{\sum_{N=x_{\max}}^{\infty} N^{-2} p_i(N|\boldsymbol{x})},$$

em que $i = 1, ..., 4$. We call this estimator mean relative squared error (MRSE).

Inferring population sizes – practice

Consider the problem of inferring the population sizes of major herbivores (Carroll and Lombard, 1985). In the first case, one is interested in estimating the number of impala (*Aepyceros melampus*) herds in the Kruger National Park, in northeastern South Africa. In an initial survey collected the following numbers of herds: $\boldsymbol{x}_{\text{impala}} = \{15, 20, 21, 23, 26\}$. Another scientific question is the number of individual waterbuck (*Kobus ellipsiprymnus*) in the same park. The observed numbers of waterbuck in separate sightings were $\boldsymbol{x}_{\text{waterbuck}} = \{53, 57, 66, 67, 72\}$ and may be regarded (for simplicity) as independent and identically distributed.

(a) Impala (b) Waterbuck

Figure 2: Two antelope species whose population sizes we want to estimate.

i) (20 marks) For each data set, sketch the marginal posterior distributions $p_1(N \mid \boldsymbol{x})$, $p_2(N \mid \boldsymbol{x})$ and $p_3(N \mid \boldsymbol{x})$. Moreover, under each posterior, provide (i) the Bayes estimator under quadratic loss and under the loss in (12) and (ii) a 95% credibility interval for N. Discuss the differences and similarities between these distributions and estimates: do the prior modelling choices substantially impact the final inferences? If so, how?

Solution. I use *Stan* platform with *PyStan* interface for *Python programming language*. The problem is that it does not handle unbounded discrete parameters. Therefore we use the approach suggested in the documentation Stan Development Team (2021) considering the population size a continuous variable. I tried to use the package *PyMC3* for this model (this package handles discrete parameters), but given convergence problems, I do not present the results.

As I will present posteriorly, this problem may be related to the sampler Metropolis. For the Geometric and independent approach, it is possible to compare the analytic curve with Stan result (considering N a real variable), and Stan got a well approximate result, while PyMC3 had convergence problems. I used 50 thousand iterations with 5 chains. The parameter adapt delta varied between 0.95 and 0.99.

Choose of the hyperparameters

In table 1, I presented the hyperparameters chosen for the models. For the Raftery approach, we chose a uniform distribution on θ (given we have no information about the populations), and we define the Gamma parameters to have mean and variance 1. For the uninformative model, we can't make prior analysis since the improper choice. For the geometric independent model, we set the hyperparameters to have uniform distribution in $(0, 1)$. Lastly, the parameters in Geometric correlated model have the following interpretation: if we want θ to have mean 0.5, it will have odds 1 and log-odds 0. The same for ν. We choose the variance to be the value such that the prior variance of ν and θ be around the variance of ν and θ in the independent case and a positive correlation of 0.4, since when the population grows (therefore its mean grows, and ν decreases), we expect θ to decrease (the probability to see a specific individual of the population).

Impala data set

In Figures 3, I sketch the marginal posterior distributions for each approach. In table 2, it is shown the Bayes estimator under quadratic loss (the expected value (Robert, 2007, Section 4)), under loss 12 (MRSE), the median point estimate, and 95% highest density interval (HDI).We observe that the shape of the posterior distributions are quite similar, with fast grow between 30-40, and a slow or fast decay, depending on the priors. The more uninformative prior we use, the heavier the tail of the distribution. We observe the mean is not a robust estimative, while the MRSE changes very little. Therefore, the prior modelling is impacted only in the tail. A distribution more uninformative generates larger intervals, what is expected.

Waterbuck data set

In Figures 4, I sketch the marginal posterior distributions for each approach. In table 3, it is shown the summary statistics. We observe a similar behavior from the graphics when compared to the impala data set. The uninformative prior generated a much heavier tail, though.

Model	**Parameters**			
Raftery	$\alpha = 1$	$\beta = 1$	$a = 1$	$b = 1$
Geometric independent	$\alpha_1 = 1$	$\beta_1 = 1$	$\alpha_2 = 1$	$\beta_2 = 1$
Geometric correlated	$\eta = [0, 0]$		$\Sigma = \begin{bmatrix} 2.5 & 1.0 \\ 1.0 & 2.5 \end{bmatrix}$	

Table 1: Hyperparameters for the developed models.

In order to see if the real approximation used in Stan is good for the discrete parameter N, we compare the sample generated with the closed-form distributions (up to a constant) geometric

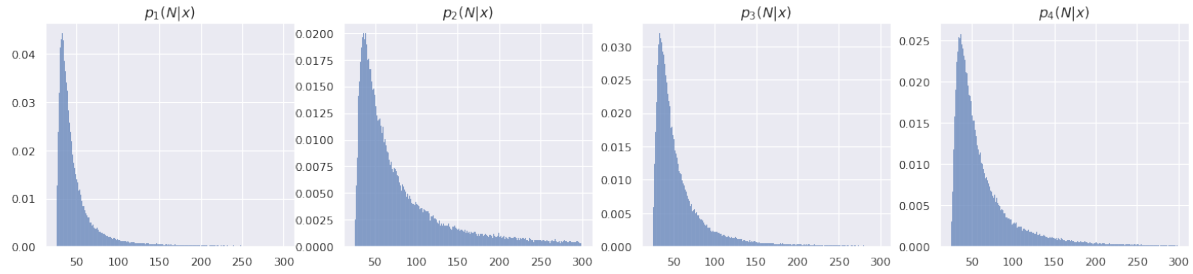

Figure 3: Marginal posterior distributions of N for models Raftery, uninformative, geometric independent, and geometric dependent, respectively, for Waterbuck data set.

Models	Mean	Median	MRSE	HDI
Raftery	88.98	40.65	37.88	[26.00, 124.75]
Uninformative	856.54	67.41	47.37	[26.01, 561.02]
Geometric independent	65.98	46.70	41.39	[26.02, 137.34]
Geometric correlated	72.44	51.82	44.01	[26.04, 161.18]

Table 2: Point estimates and highest density interval for each model in Impala data set.

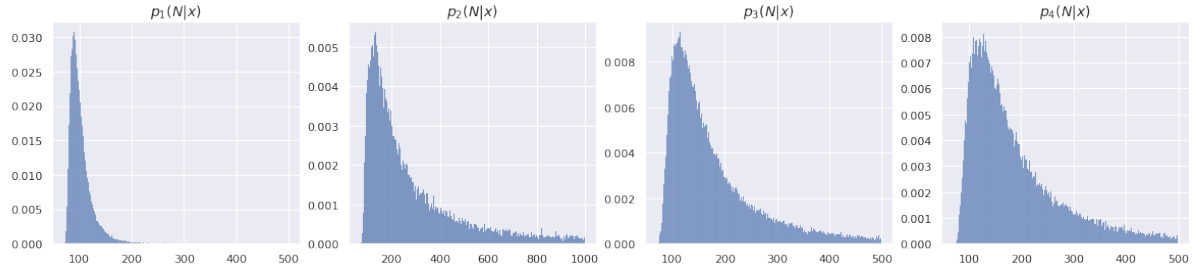

Figure 4: Marginal posterior distributions of N for models Raftery, uninformative, geometric independent, and geometric dependent, respectively, for Waterbuck data set.

Models	Mean	Median	MRSE	HDI
Raftery	104.98	97.19	96.68	[74.74, 143.35]
Uninformative	1111.69	235.74	155.66	[75.0, 1920.81]
Geometric independent	221.70	154.46	132.79	[75.08, 470.09]
Geometric correlated	220.72	166.26	140.25	[78.86, 476.56]

Table 3: Point estimates and highest density interval for each model in Waterbuck data set.

independent and uninformative. To calculate the constant, we sum the values until we reach the zero from *NumPy*. From figure 5, we note that the approximation is good, even though for the uninformative model, both suffer from numerical issues. Because of this we follow with this approach.

j) (25 marks) Let $\bar{x} = K^{-1}\sum_{k=1}^{K} x_k$ and $s^2 = K^{-1}\sum_{k=1}^{K}(x_k - \bar{x})^2$. For this problem, a sample is said to be *stable* if $\bar{x}/s^2 \geq (\sqrt{2}+1)/\sqrt{2}$ and *unstable* otherwise. Devise a simple method of moments estimator (MME) for N. Then, using a Monte Carlo simulation, compare the MME to the three Bayes estimators under quadratic loss in terms of relative mean squared error. How do the Bayes estimators compare to MME in terms of the stability of the generated samples? *Hint*: You may want to follow the simulation setup of Carroll and Lombard (1985).

Solution. I first observe that both data sets are unstable by the above definition. The mean over variance from impala data is 1.6, while waterbuck's is 1.3. Both smaller than 1.71. Since the binomial distribution has two unknown parameters, I need to calculate two moments to obtain the MME. I calculate the sample mean $\hat{\mu}$ and the sample second moment $\hat{\sigma}^2 - \hat{\mu}^2$, where $\hat{\sigma}^2$ is the

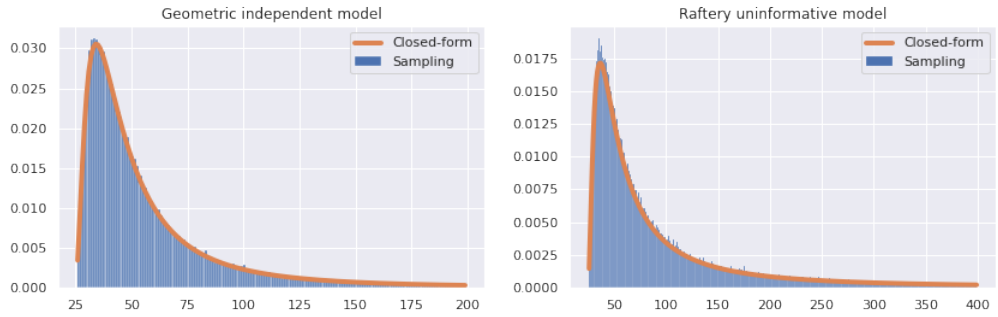

Figure 5: Approximation analysis of Stan's assumption that N is a positive real number.

sample variance. Therefore, I build the following system of equations

$$\hat{\mu} = N\theta,$$
$$\hat{\sigma}^2 = N\theta(1-\theta),$$

that is

$$\hat{\sigma}^2 = \hat{\mu}(1-\theta) \implies \hat{\theta} = 1 - \frac{\hat{\sigma}^2}{\hat{\mu}}$$

and

$$\hat{N} = \frac{\hat{\mu}}{\hat{\theta}} = \frac{\hat{\mu}^2}{\hat{\mu} - \hat{\sigma}^2}.$$

For instance, the N_{MME} calculated were 56.54 for the impala data set, and 271.84 for the waterbuck.

The used setup to compare the estimators and to analyse the stability is from (Carroll and Lombard, 1985, Section 3). First, I generate eight samples with different parameters, and calculate the four Bayes estimates, besides the moment estimator. For each sample, I also generate a perturbed sample adding one to the largest value. I keep the same hyperparameters defined before except the beta hyperparameters of θ, which I increase to 2, because the non-informative prior was giving low N estimates. The correlation in the last model was increased to 0.8, given the figure 1.

Table 4 presents the results. The first thing we observe is that small values of θ induces smaller estimates of N, therefore it is important to put stronger priors of θ. This can be done studying this parameter in controlled spaces, where it is possible to know the population size. The correlated model increases the estimates, without the need of improper priors. THe moment estimator is undefined in some cases anf more unstable, but not always.

After that, $3 \leq K \leq 22$, $0 < \theta < 1$, and $1 \leq N \leq 100$ was randomly and uniformly chosen. There were 1500 (around 12 hours of computation) generated cases with these parameters, and we separated them between stable, when $\hat{\mu} \geq (1+1/\sqrt{2})\hat{\sigma}^2$ and unstable otherwise. We calculated the Bayes estimates and the moment estimate for each to obtain a Monte Carlo loss estimation. Several samples had convergence problems (less than 1%), specially in the uninformative case. I could not spend more time trying to solve this, because it was already too computationally expensive. Few simulations had serious convergence problems (around 50%), and were discarded. I also observe that more time calibrating the priors is very important, and it caused serial problems. The results can be found in table 5. The uninformative is the best estimator in this case. This is caused because the others estimate down the N value. The geometric and correlated case was the opposite, and it may be caused by the prior strong correlation. The moment estimator appeared to be very unstable. I conclude one need to specify improved hyperparameters for the priors. All the codes can be found in the Jupyter notebook at Github repository Moschen (2021).

	Parameters			Estimators				
Sample	**N**	**p**	**K**	**Raft**	**Uninf**	**Geom id**	**Geom corr**	**Moment**
1	75	0.32	5	45.32	58.99	50.88	68.52	78.87
Perturbed	75	0.32	5	47.39	63.81	53.48	72.18	105.63
2	34	0.57	4	29.05	31.57	31.07	40.95	24.18
Perturbed	34	0.57	4	31.11	35.04	33.35	43.81	26.16
3	37	0.17	20	22.28	27.47	20.92	27.21	< 0
Perturbed	37	0.17	20	24.28	30.60	22.61	29.68	< 0
4	48	0.06	15	9.09	10.62	8.31	10.49	< 0
Perturbed	48	0.06	15	10.64	12.44	9.71	12.44	< 0
5	40	0.17	12	16.37	18.98	15.86	19.51	31.39
Perturbed	40	0.17	12	18.34	21.13	17.50	21.70	61.12
6	74	0.68	12	59.98	62.13	63.35	75.51	58.26
Perturbed	74	0.68	12	61.33	64.08	65.02	77.28	59.17
7	55	0.48	20	42.51	47.14	45.57	52.93	42.78
Perturbed	55	0.48	20	44.01	49.64	47.53	55.49	44.48
8	60	0.24	15	29.56	35.31	30.65	37.88	37.33
Perturbed	60	0.24	15	31.46	37.97	32.72	40.81	42.84

Table 4: Simulations generated by the parameters N, θ, and K, and the estimators calculated by each model, besides the moment estimator. For each sample, the first line indicates the estimates for it, while the second line the estimates fot the perturbed sample, adding one to the largest value.

Estimates	**Stable cases**	**Unstable cases**
Raftery	329.65	1071.89
Uninformative	297.37	953.31
Geometric independent	305.31	1050.04
Geometric correlated	580.64	979.06
Moment	336.27	$> 1e10$

Table 5: Mean quadratic loss estimated by Monte Carlo for each estimator.

Appendix

A Binomial and Poisson

Suppose $X|N \sim \text{Binomial}(N, p)$ and $N \sim \text{Poisson}(\mu)$. We shall derive the distribution of X. By the Law of total probability,

$$\begin{aligned}
\Pr(X = k) &= \sum_{n=0}^{\infty} \Pr(X = k|N = n) \Pr(N = n) \quad [\Pr(X = k|N = n) = 0 \text{ if } k > n] \\
&= \sum_{n=k}^{\infty} \binom{n}{k} p^k (1-p)^{n-k} \frac{e^{-\mu}\mu^n}{n!} \\
&= \frac{e^{-\mu}p^k}{k!} \sum_{n=k}^{\infty} \frac{(1-p)^{n-k}}{(n-k)!} \mu^{n-k}\mu^k \\
&= \frac{e^{-\mu}(p\mu)^k}{k!} \sum_{n=k}^{\infty} \frac{((1-p)\mu)^{n-k}}{(n-k)!} \qquad (14) \\
&= \frac{e^{-\mu}(p\mu)^k}{k!} \sum_{m=0}^{\infty} \frac{((1-p)\mu)^m}{m!} \qquad [m = n-k] \\
&= \frac{e^{-\mu}(p\mu)^k}{k!} e^{(1-p)\mu} \\
&= \frac{e^{-p\mu}(p\mu)^k}{k!},
\end{aligned}$$

what implies $X \sim \text{Poisson}(p \cdot \mu)$.

Bibliography

Carroll, R. J. and Lombard, F. (1985). A note on N estimators for the binomial distribution. *Journal of the American Statistical Association*, 80(390):423–426.

Moschen, L. M. (2021). Exam a1. Available at https://github.com/lucasmoschen/phd-bayesian-statistics/blob/main/notes/exam/A1.pdf.

Raftery, A. E. (1988). Inference for the binomial n parameter: A hierarchical bayes approach. *Biometrika*, 75(2):223–228.

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

Stan Development Team (2021). Mark-recapture models. Available at https://mc-stan.org/docs/2_27/stan-users-guide/mark-recapture-models.html.

Xie, Y. and Carlin, B. P. (2006). Measures of bayesian learning and identifiability in hierarchical models. *Journal of Statistical Planning and Inference*, 136(10):3458–3477.