IMPERIAL COLLEGE LONDON

MSc COURSE IN COMPUTATIONAL METHODS IN ECOLOGY AND EVOLUTION

EXAM 2

For Internal Students of Imperial College of Science, Technology and Medicine

Exam Date: Tuesday, 27th March 2019, 10:00 – 13:00

Length of Exam: 3 HOURS

Instructions: All sections are weighted equally. It is a three-hour exam, and there are 5 sections, so it is a reasonable guideline to spend about 35 minutes on each section. Most sections allow you to choose between questions, answering ONE. Please read the instructions at the head of each section carefully.

PLEASE PUT ANSWERS TO EACH SECTION IN A SEPARATE EXAM BOOK.

WE REALLY MEAN IT. PLEASE PUT ANSWERS TO EACH SECTION IN A SEPARATE EXAM BOOK. THE REASON FOR THIS IS THEN WE CAN PARALLELIZE MARKING AMONG THE DIFFERENT LECTURERS AND YOU GET THE MARKS BACK SOONER.

Continues on next page

Section 1: Maths

A. Solve ONE of the following exercises [40%]:

(i) Consider the following function:

$$f(x) = \begin{cases} -3\sin x & \text{if} \quad x \leq -\frac{\pi}{2} \\ a\sin x + b & \text{if} \quad \frac{-\pi}{2} < x < \frac{\pi}{2} \\ \cos x & \text{if} \quad x \geq \frac{\pi}{2} \end{cases},$$

and find the values of the constants a and b that make the function constant for all points of $\mathbb{R}$. Hint: Remember that a function $f(x)$ is continuous at x_0 if and only if $\lim_{x\to x_0^+} f(x) = \lim_{x\to x_0^-} f(x) = f(x_0)$.

(ii) Consider the following matrix

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & k \\ 1 & 4 & k^2 \end{pmatrix}$$

and determine the values of the constant k that makes the matrix invertible.

Hint: An $n \times n$ matrix is invertible if its rank is n. And the rank of the matrix is the number of nonzero rows obtained when you transform the matrix into a reduced row echelon form.

Model Answer (Marker – Samraat Pawar (1st), James Rosindell (2nd)):

(i) In any point $x \neq \frac{\pi}{2}, -\frac{\pi}{2}$ the function is continuous, since it is a combination of continuous functions, independent of the values of a and b. Therefore, we have to guarantee continuity at these potentially problematic points. For f being continuous at $-\frac{\pi}{2}$ we require that:

$$\lim_{x\to(-\frac{\pi}{2})^+} f(x) = a\sin\left(-\frac{\pi}{2}\right) + b = \lim_{x\to(-\frac{\pi}{2})^-} f(x) = -3\sin\left(-\frac{\pi}{2}\right),$$

obtaining a first equation, $-a + b = 3$. Now we explore the second point, $\frac{\pi}{2}$:

$$\lim_{x\to(-\frac{\pi}{2})^+} f(x) = \cos\left(\frac{\pi}{2}\right) = \lim_{x\to(-\frac{\pi}{2})^-} f(x) = a\sin\left(\frac{\pi}{2}\right) + b,$$

which leads to a second equation, $0 = a + b$. Therefore, f is continuous $\forall x \quad \in \mathbb{R}$ if and only if:

$$\begin{cases} -a + b = 3 \\ a + b = 0 \end{cases},$$

and, as a consequence, we obtain that $a = -\frac{3}{2}$ and $b = \frac{3}{2}$.

(ii) Following the hint, we transform the matrix into row echelon form. Calling R_i to the row i:

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & k \\ 1 & 4 & k^2 \end{pmatrix} \xrightarrow[R_3 - R_1]{R_2 - R_1} \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & k-1 \\ 0 & 3 & k^2-1 \end{pmatrix} \xrightarrow[R_3 - 3R_2]{} \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & k-1 \\ 0 & 0 & k^2-3k+2 \end{pmatrix}.$$

Since the matrix is invertible if it has rank 3, we need to avoid that the element $(3,3)$ vanishes, obtaining in this way three nonzero rows. The polynomial appearing in this cell can be easily factorized as $k^2 - 3k + 2 = (k-1)(k-2)$, from which we see that, as soon as $k \neq 1, 2$, the matrix will be invertible.

Continues on next page

B. Solve ONE of the following exercises [60%]:

(i) Prove that the following differential equation is exact and solve it:

$$(6x^2 - y + 3)dx + (3y^2 - x - 2)dy = 0.$$

(ii) Consider the following integral:

$$\int x\sqrt{x+1}dx$$

and solve it with two different methods: a) by parts, and b) by substitution. Show that, despite of the fact that the solutions seem to be different, they are actually the same, differing only by a constant.

Model Answer (Marker – Samraat Pawar (1st), James Rosindell (2nd)):

(i) This ODE is exact, we can easily check it:

$$\frac{\partial M}{\partial y} = \frac{\partial}{\partial y}(6x^2 - y + 3) = -1 = \frac{\partial N}{\partial x} = \frac{\partial}{\partial x}(3y^2 - x - 2) = -1,$$

and it means that there exist a solution of the ODE, $\Psi(x, y) = C$, such that

$$\begin{aligned} \frac{\partial \Psi}{\partial x} &= M(x, y) \\ \frac{\partial \Psi}{\partial y} &= N(x, y). \end{aligned}$$

This allow us to find the function Ψ integrating either M or N. We choose M:

$$\Psi = \int (6x^2 - y + 3)dx = 2x^3 - xy + 3x + \phi(y),$$

where $\phi(y)$ is a function that we can determine acknowledging that $\partial\Psi/\partial y = N(x, y)$:

$$\frac{\partial \Psi}{\partial y} = \frac{\partial}{\partial y}(2x^3 - xy + 3x + \phi(y)) = -x + \phi'(y) = 3y^2 - x - 2 = N(x, y),$$

which means that

$$\phi'(y) = 3y^2 - 2$$

and, integrating, we get

$$\phi(y) = \int 3y^2 - 2 = y^3 - 2y.$$

Therefore, the final solution is:

$$\Psi(x, y) = 2x^3 - xy + 3x + y^3 - 2y = C,$$

with C a real number.

Continues on next page

(ii) We integrate first by parts using the following choices:

$$\begin{aligned} u &= x, & du &= dx \\ dv &= \sqrt{x+1}dx, & v &= \frac{2}{3}(x+1)^{\frac{3}{2}}, \end{aligned}$$

what yields

$$\begin{aligned} \int x\sqrt{x+1}dx &= \frac{2}{3}x(x+1)^{\frac{3}{2}} - \frac{2}{3}\int (x+1)^{\frac{3}{2}}dx \\ &= \frac{2}{3}x(x+1)^{\frac{3}{2}} - \frac{4}{15}(x+1)^{\frac{5}{2}} + c. \end{aligned}$$

Now we integrate by substitution using the change:

$$u = x+1, \qquad x = u-1, \qquad du = dx$$

that allow us to easily solve the integral:

$$\begin{aligned} \int x\sqrt{x+1}dx &= \int (u-1)u^{\frac{1}{2}}du \\ &= \frac{2}{5}u^{\frac{5}{2}} - \frac{2}{3}u^{\frac{3}{2}} + c \\ &= \frac{2}{5}(x+1)^{\frac{5}{2}} - \frac{2}{3}(x+1)^{\frac{3}{2}} + c. \end{aligned}$$

To see that both solutions are indeed the same, we neglect the constant and substract both solutions:

$$\begin{gathered} \left(\tfrac{2}{3}x(x+1)^{\frac{3}{2}} - \tfrac{4}{15}(x+1)^{\frac{5}{2}}\right) - \left(\tfrac{2}{5}(x+1)^{\frac{5}{2}} - \tfrac{2}{3}(x+1)^{\frac{3}{2}}\right) \\ = x(x+1)^{\frac{3}{2}}\left(\tfrac{2}{3}x - \tfrac{4}{15}(x+1) - \tfrac{2}{5}(x+1) + \tfrac{2}{3}\right) \\ = x(x+1)^{\frac{3}{2}}(0) = 0. \end{gathered}$$

Continues on next page

Section 2: Dynamical Models in Ecology

Please select exactly **one question** and answer it. Please indicate clearly in your answer book which question you are answering.

A. After it was inadvertently introduced in the 1960s, the flightless midge Eretmoptera murphyi has invaded parts of Signy Island, situated 600km from the Antarctic. The midge impacts soil composition, and its actions introduce nitrates in the soil which plants prefer. Signy's ecosystem is nutrient poor and has around 50 moss species and two flowering plant species which all compete for nutrients. A researcher commented on the invasive midge: "It is basically doing the job of an earthworm, but in an ecosystem that has never had earthworms... . Any input of nutrients in an ecosystem that is already adapted to very low nutrient levels and very extreme conditions will have an impact. What those might be and whether they're going to be good or bad, we don't quite know yet."

To assess the impact you are asked to design a simple model for the effects of the midge on Signy's ecosystem.

(i) What processes will you include in your model? What variables will you include in the model? Motivate and justify, and don't forget the model should be simple. [70%]

(ii) Can you hypothesise what the possible outcomes will be ? [30%]

Model Answer (Markers – Samraat Pawar (first), James Rosindell (second)):

(i) An obvious choice is the Lotka-Volterra interaction model. One could argue if that should be in discrete or continuous time, both arguments can be made. The model must include some description of the plants and of the mosses. It should describe the processes of competition within and between groups of species, where the point should be made that a dominant process is competition for nitrogen. For a simple model, the obvious choice is to have an variables for plants and mosses. The effects of midge and nitrogen should to be captured somehow as they mediate competition. They could be included as separate variables, but easier would be to remove those again using a quasi-steady state argument, as we did for apparent competition (making this point would be a better answer)

(ii) The possible outcomes of the L-V model are competitive exclusion, coexistence of alternative stable state. If the observation was made in (a) that the LV model was a good choice, then these are the options. Here, this is quite clearly a case where exclusion or competition will follow, unless there is a second resource that the plants and the mosses compete for.

B. The Gompertz model describes the growth of a population with the differential equation:

$$\frac{dN}{dt} = -rN \ln\left(\frac{N}{k}\right)$$

Here, N is the size of the population, k the carrying capacity, which is assumed to be positive, and r the maximum growth rate. This model describes positive population sizes only.

(i) Calculate all equilibria of the model [35%]

(ii) Calculate the stability of the largest, positive equilibrium [65%]

Model Answer (Markers – Samraat Pawar (first), James Rosindell (second)):

(i) 0 and k

(ii) The model linearised around N* is: $\frac{dN}{dt} = -r \ln N*/k - rk$.

For $N* = k$ the eigenvalue is $-rk$ and the equilibrium is always stable.

Continues on next page

Section 3: Population Genetics & Evolutionary Ecology

Please select exactly **one question** and answer it. Please indicate clearly in your answer book which question you are answering.

A. A green system is being introduced at Silwood to treat its waste water. A mixed tank of volume v will receive waste water at a constant flow rate of f litres per second. The inflow water contains a toxic compound X at a concentration of 0.1 moles per litre. A single bacterial population of density N metabolises that compound into a harmless product.

It is intended that the outflow will drain directly into Silwood stream and from there to Virginia Water and the river Thames. Environmental regulations stipulate that the maximum concentration permitted for releasing water into streams and rivers is 0.01 moles per litre.

A standard chemostat model for this system is:

$$\frac{dN}{dt} = \frac{ckSN}{m+S} - DN$$
$$\frac{dS}{dt} = D(0.1 - S) - \frac{kSN}{m+S}$$

where t is time, S is concentration of compound X, moles per liter, m is the half-saturation constant (moles per litre), N is population density in (cells per litre), k is metabolic rate (moles per cell per second), c is the conversion rate from metabolism to growth in cells/mole, and $D = f/v$ in % (proportion) per second.

(i) Include a diagram of your model, showing all your variables and parameters. [10%]

(ii) Derive and interpret the condition(s) needed to meet the legal requirements for the outflow. [30%]

(iii) A preliminary trial of the new system reveals that the concentration in the outflow is too high. What features would you modify to improve the performance? [20%]

(iv) What other factors not considered in your model in part (i) would affect the long-term performance of the system? [20%]

(v) The Silwood sustainable living group want to couple the system with a new source of green protein. A single fish species will grow in the waste tank, feeding on the bacterium population, and fish will be harvested and served in the refectory. Viability of this scheme depends on sustaining a particular rate of biomass production of the fish. Sketch out how you would extend the model in part (i) to include fish feeding on the bacteria and explore whether the dual purpose system (waste water and protein production) would work. [20%]

Model Answer (Marker – Tim Barraclough (1st), Austin Burt (2nd)):

(i) Diagram should show inflow and outflow, and that N and S are state variables within the tank. This is all material from lecture, but applied to an unfamiliar example.

(ii) Steady-state solution, when $dN/dt = 0$ and $dS/dt = 0$

Steady state concentration of X, $\hat{S} = \frac{Dm}{ck-D}$

So $\frac{Dm}{ck-D} < 0.01$ is the condition for meeting environmental regulations

(iii) Need to make the quantity in the inequality smaller, so can decrease m, increase c or k, or decrease D, which could be by reducing the inflow or by increasing the volume = top marks for pragmatically spotting that reducing the inflow would lead to backlog of waste water, so better to change the volume of the tank, which is also easier than changing the bacteria's properties.

Continues on next page

(iv) Flow rate would likely not be constant in reality, mixing would not be perfect - The bacteria might evolve to use different substrates in the water = especially if those were more profitable for growth; or evolve to stick to the sides of the tank. We covered evolution in the classes, not specifically applied to this question so credit for well justified, reasonable suggestions.

(v) This is combining a trophic model with the bacterial model, so something like:

$$\frac{dP}{dt} = abNP - hP,$$

where P is density or biomass of fish, a is feeding rate and b is conversion to fish density/biomass, and h is per capita harvesting rate of fish

Would also need to modify bacterial equation

$$\frac{dN}{dt} = \frac{ckSN}{m+S} - DN - bNP$$

Which in turn would change the expression for steady state of S, replacing D with $D + bP =$ P is still a variable so can't tease out but if assume fish persist then this would reduce effectiveness of waste water system, would need super fast growing bacterium.

Not looking for formal treatment, but creativity for how they would modify model, ability to think about way to implement unfamiliar case, which was the objective of the class practical. Merit – understand how to add in an equation for fish growth; Distinction – some insight into how this change would affect the system and goals.

B. This is a question on gene drive and population suppression. Answer all four parts. Suppose there is a random-mating population with 2 alleles A and B. At generation t, their allele frequencies are $(1 - q_t)$ and q_t respectively, $0 < q_t < 1$.

(i) Express the genotypic frequencies for genotypes AA, AB, and BB, in terms of q_t, under the assumption of Hardy-Weinberg equilibrium. [15%]

Assume the fitness of genotype AA and AB are the same (i.e. both with fitness 1, no heterozygote advantages or disadvantages), but individuals with genotype BB are sterile when they reach adulthood (with fitness 0).

(ii) Show that the frequency of genotype AB after selection is $\frac{2q_t}{1+q_t}$. [25%]

(iii) Further, a gene drive biases the transmission of B such that it is inherited more frequently than by random segregation. Suppose the AB heterozygotes produce B gametes with proportion 1 $(d > 0.5)$, and therefore the allele frequency of B in the next generation is $q_{t+1} = \frac{2q_t d}{1+q_t}$. Show that the equilibrium frequency of allele B is $q' = (2d - 1)$. (Hint: consider solving $q_{t+1} = q_t = q'$) [30%]

(iv) Let N_t be the population size at generation t. Suppose the population regulates itself according to Beverton-Holt model, which has the following form:

$$N_{t+1} = \frac{R_0 N_t}{1 + N_t/M}$$

where R_0 is the growth rate, and $M(R_0 - 1)$ is the carrying capacity. This model is however inadequate because it does not incorporate the loss of breeding individuals due to the infertile BB homozygotes. We can modify the model by replacing N_t in the numerator with $N_t\left(1 - (2d-1)^2\right)$, that is, the average proportion of breeding individuals, whose genotypes are not BB:

$$N_{t+1} = \frac{R_0 N_t \left(1 - (2d-1)^2\right)}{1 + \frac{N_t}{M}}$$

Show that the equilibrium population size under the modified model is [30%]

$$N' = M\{R_0[4d\,(1-d)] - 1\}$$

.

Model Answer (Marker – Austin Burt (1st), Tin-yu Hui (2nd)):

Continues on next page

Evolutionary genetics

i) frequency of AA $= (1-q_t)^2$

.. AB $= 2q_t(1-q_t)$

.. BB $= q_t^2$

---

ii) frequency of AB after selection

$$= \frac{2q_t(1-q_t)}{(1-q_t)^2 + 2q_t(1-q_t) + 0}$$

$$= \frac{2q_t}{1-q_t + 2q_t} \qquad (\because 0<q_t<1)$$

$$= \frac{2q_t}{1+q_t}$$ #.

---

iii) Given $q_{t+1} = \dfrac{2q_t d}{1+q_t}$

At equilibrium, $q_{t+1} = q_t = q'$

i.e. Solve $q' = \dfrac{2q'd}{1+q'}$

$$1 = \frac{2d}{1+q'} \qquad (\because q' \neq 0)$$

$$1+q' = 2d$$

$$q' = 2d - 1$$ #.

---

iv) At equilibrium, $N_{t+1} = N_t = N'$

i.e. $$N' = \frac{R_0 N'[1-(2d-1)^2]}{1+\frac{N'}{M}}$$

$$1+\frac{N'}{M} = R_0[1-(4d^2-4d+1)] \qquad (\because N'>0)$$

$$M+N' = R_0 M[-4d^2+4d]$$

$$N' = M R_0[4d(1-d)] - M$$

$$N' = M\left\{R_0[4d(1-d)] - 1\right\}$$

Section 4: Maximum Likelihood & GLMs

Please select exactly **one question** and answer it. A calculator may be required.

A. You may use the χ^2 table below for critical values.

Degrees of freedom	$\chi^2_{0.95}$
1	3.84
2	5.99
3	7.81
4	9.49

(i) Let X be a Gamma random variable with two parameters $k > 0$ and $\theta > 0$. The moment generating function of X is $M_X(t) = (1 - \theta t)^{-k}$ for $t < 1/\theta$.

(a) Show that $E(X) = k\theta$ [25%]

(b) Show that $\text{var}(X) = k\theta^2$ [35%]

(ii) Alex, a CMEE student, plotted a log-likelihood function against the parameter of interest p. Describe, as precisely as possible, how Alex can find the maximum likelihood estimate as well as the 95% confidence interval for p. You may include graphs or equations as part of your answer. [20%]

(iii) Please also discuss how Alex can find the 95% confidence interval using approximate normality. You may include graphs or equations as part of your answer. [20%]

Model Answer (Markers – Tin-yu Hui (first), Austin Burt (second)):

Continues on next page

(i) (a) $M_x(t) = (1-\theta t)^{-k}$ for $t < \frac{1}{\theta}$

$$M_x'(t) = \frac{d}{dt}\left((1-\theta t)^{-k}\right)$$

$$= -k(1-\theta t)^{-k-1}(-\theta)$$

$$= \theta k(1-\theta t)^{-k-1}$$

$$E(X) = M_x'(0) = \theta k\left(1-0\right)^{-k-1} = k\theta$$

(b) $M_x''(t) = \frac{d}{dt}\left[\theta k(1-\theta t)^{-k-1}\right]$

$$= \theta k(-k-1)(1-\theta t)^{-k-2}(-\theta)$$

$$= \theta^2 k(k+1)(1-\theta t)^{-k-2}$$

$$E(X^2) = M_x''(0) = \theta^2 k(k+1)((1-0)^{k-2} = \theta^2 k(k+1)$$

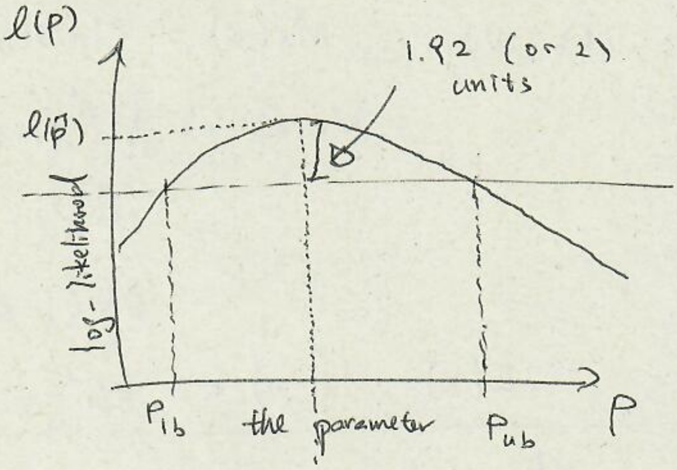
$$Var(x) = E(X^2) - E(x)^2$$

$$= \theta^2 k(k+1) - (\theta k)^2$$

$$= \theta^2 k^2 + \theta^2 k - \theta^2 k^2$$

$$= \theta^2 k$$

(ii)(a) Say, the log-likelihood plot looks like this.



Point estimate: The value of p (say $\hat{p}$) at which the log-likelihood is maximised.

i.e. $l(\hat{p}) \geq l(p)$ for all p

95% C.I.: Draw a horizontal line,

$$\frac{\chi^2_{1,0.95}}{2} = \frac{3.84}{2} = 1.92 \text{ units} \quad (\text{or } 2 \text{ units})$$

below the maximum. The intersections between the horizontal line and the log-likelihood function are the lower- and upper- 95% C.I. for p.

(b) At $\hat{p}$, calculate the second derivative of the log-likelihood function.

The variance of the estimator $\approx \dfrac{-1}{l''(\hat{p})}$

95 C.I. for p based on approximate normality

$$= \hat{p} \pm 1.96 \sqrt{\frac{-1}{l''(\hat{p})}}$$

B. Intra-specific competition for limited resources can lead to fighting between individuals as the rewards are very high. The Mediterranean field cricket (*Gryllus bimaculatus*) is an excellent model to investigate aggressive behaviour between males as they fight over mates and resources. Their behaviour in these encounters are stereotypical and methodological, aiding identification of encounters using an ethogram. You are supervising an undergraduate project to understand the effect of mate presence and male size on the number of aggressive encounters between pairs of male crickets. The measured variables include:

- Number of aggressive encounters in 10 minutes (10 zero's, 40 non-zero recordings)
- Pronotum width in millimetres as a measure of male size (ranges from 3.5 to 7.5mm)
- Treatment: indicating whether a female was present or absent in the pairs fighting

The undergraduates have fitted a Poisson GLM because they mimicked the analyses of recently published research. They have approached you to explain the fundamentals of GLMs and help them with the interpretation of their results.

The result they have obtained is:

```
Call: glm(formula= NoEncounters∼Treatment*ProWidth, family="poisson", data= ↩
    crickets)

                              Estimate  Std.Error  z-value  Pr(>|z|)
(Intercept)                   1.00      0.57       2.10     0.045*
TreatmentWithout              0.20      0.73       4.56     0.001**
ProWidth                      0.60      0.09       7.87     0.000***
TreatmentWithout:ProWidth    -0.50      0.12       6.09     0.004**

Null deviance: 413.78 on 189 degrees of freedom
Residual deviance: 406.51 on 186 degrees of freedom
```

Answer each of the following questions:

(i) Explain the main differences between linear models and generalised linear models and justify why a Poisson family was fitted to their data. [40%]

(ii) Interpret the coefficients of the model output providing approximate estimates of the effect size of pronotum width and presence of females on the number of aggressive encounters between male crickets [30%].

(iii) What does the dispersion parameter test and how is it calculated [20%]?

(iv) Justify whether the students should change their analyses [10%].

Model Answer (Markers – Josh Hodge (first), Julia Schroeder (second)):

(i)
- GLMs can account for constrained response variables, i.e. those that are not continuously distributed. Examples of these include count/Poisson data, binomial (constrained between 0 and 1) and binary (constrained at 0 or 1) data. Linear models can only handle continuously distributed data.
- A log transformation can be applied to a response variable to shoehorn or linearise the measure to be appropriate for linear models. GLMs differ in that they apply a transformation over the linear predictor (the intercept, slope coefficients and errors). This transformation is called a link function, in the Poisson family a log-linear link function is used and in the binomial family a logit or probit link function is applied.
- The Poisson family GLM was fitted to the data for two reasons:
    - The data is count data and therefore is constrained to absolute whole numbers

Continues on next page

  - A log transformation of the number of aggressive encounters wouldn't be appropriate because 20% of the measurements would be removed as log of zero is not mathematically possible and the zero recordings are biologically meaningful in this experiment.

(ii)
- The intercept here refers to the treatment with females.
- There is a significant effect of both pronotum width and the presence of females.
- The effect size of the pronotum width is therefore:
  - When females are present, for every millimetre increase in pronotum width, there is a significant increase in expected log count of aggressive encounters by 0.60.
    * If exponentiated: ..., there is a significant increase in the number of aggressive encounters by a factor of 1.82 or 82%.
  - When females are absent, for every millimetre increase in pronotum width, there is a significant increase in expected log count of aggressive encounters by (0.60-0.50) 0.1.
    * If exponentiated: ..., there is a significant increase in the number of aggressive encounters by a factor of 1.11 or 11%.
- The presence of females therefore increases the number of aggressive encounters by a factor of 6 (if not exponentiated) and 7.5 (if exponentiated).

(iii)
- The dispersion parameter is testing the assumption of Poisson data that the mean and variance are equal. If the variance and mean are not equal the dispersion parameter will indicate over- or underdispersion; the variance is greater than the mean or the variance is lower than the mean respectively.
- The dispersion parameter is calculated by dividing the residual deviance by its associated degrees of freedom.

(iv)
- The dispersion parameter in these results is 2.18 (approximation are acceptable), therefore the students should change their Poisson GLM to a quasi-Poisson GLM to account for this overdispersion.

Continues on next page

Section 5: Bayesian statistics

This section has *one compulsory question* worth 60-100% of the total mark depending on how many assignments (each one worth 10%) you submitted before the deadline. That is, if you submitted all assignments, this section will contribute to 60% of your grade, from 40 to 100%. On the other hand, if you did not submit any assignment, this section will contribute to 100% of your final grade.

This section is divided into five points (i-v), *each one carrying equal weight.*

The time between extinction events of amphibians in South America under current climatic conditions (λ) can be described with an exponential distribution

$$p(x|\lambda) = \lambda e^{-\lambda x}$$

for $x \geq 0$ with $X = \{x_1, x_2, ..., x_n\}$ being a continuous random variable.

Note that $p(x|\lambda) = 0$ for $x < 0$.

The conjugate prior for an exponential distribution is a Gamma distribution

$$p(\lambda|\alpha, \beta) = \frac{\lambda^{\alpha-1} e^{-\lambda/\beta}}{\Gamma(\alpha)\beta^{\alpha}}$$

with $\Gamma(\alpha)$ being the gamma function (a normalising factor).

Note that $\alpha > 0$ and $\beta > 0$ and that the expected value is $\alpha\beta$ and the variance is $\alpha\beta^2$

(i) Show that the posterior distribution $p(\lambda|x)$ is a Gamma distribution $G(\alpha', \beta')$ with $\alpha' = \alpha + 1$, assuming we have a single observation x. Please note that $\beta' = \beta + x$.

(ii) Assume that, based on past observations, you expect a time between extictions of 3.5 *a priori* but with a large uncertainty associated to it. Choose suitable values for hyper-parameters α and β to fit this prior belief and calculate the posterior mean with $x = 2.5$.

(iii) Assume that you calculate a Bayes factor for testing $M1 = \{\lambda \geq t\}$ vs. $M2 = \{\lambda < t\}$ with $t > 0$ being a threshold on whether or not to activate a conservation strategy. You obtain a value of 150. Discuss the support for $p(\lambda|x) \geq t$ and $p(\lambda|\alpha, \beta) \geq t$ in light of the definition and interpretation of Bayes factors.

Assuming that the 95% highest density posterior interval for λ is $[0.29 - 28.69]$, what can we say about the probability that the time between extictions is larger than 28.69?

(iv) Assume that your prior information is now described by a Normal distribution $p(\lambda|\mu, \sigma^2)$, that is you lack a conjugate prior. Describe an algorithm (or write a pseudo code) for obtaining samples for the posterior distribution $p(\lambda|x)$. Be as precise and formal as possible and highlight any pros and cons of the chosen algorithm.

(v) Answer either point (v-a) or (v-b).

(v-a) Describe the rationale behind the sequential Monte Carlo (SMC) MCMC algorithm to estimate parameters and perform model selection. What are the main advantages (and disadvantages, if any) over a standard MCMC? What are the additional parameters of the algorithm?

(v-b) Describe the main features of representing probabilistic relationships between random variable with a Bayes network.

Model Answer (Marker – Matteo Fumagalli (1st), Tin-Yu Hui (2nd)):

(i)

Here we need to use Bayes formula. By dropping all terms that do not contain λ, we realise that the posterior is a Gamma distribution and, after some algebra, we find that $\alpha' = \alpha + 1$.

Continues on next page

$$
\begin{aligned}
p(\lambda|x) &\approx p(x|\lambda)p(\lambda|\alpha,\beta) \\
p(\lambda|x) &\approx \lambda e^{-\lambda x}\lambda^{\alpha-1}e^{-\lambda/\beta} \\
p(\lambda|x) &\approx \lambda^{\alpha-1+1}e^{\lambda x-\lambda/\beta}
\end{aligned}
$$

for the latter we evince $p(\lambda|x) = G(\alpha', \beta')$ with $\alpha' = \alpha + 1$.

We went through a similar (but different) case in class. Here I am testing how students recall Bayes equation and whether they are able to apply it.

(ii)

Given the expected value and variance of a Gamma distribution, a suitable prior might be $G(\alpha = 1, \beta = 3.5)$, which produces the desired expected value with a large variance.

With this choice the posterior mean is the expected value of $G(1 + 1, 3.5 + 2.5)$ which is 12.

Here I am testing how students can translate some prior information into a suitable probability distribution and whether they know the meaning of posterior mean and expected value.

(iii)

This bayes factor provides very strong support for M1, and it represents the shift in odds when we move from the prior towards the mean.

Given the HPD, we can say that the posterior is lower than 0.05.

Here I am testing whether students have the concept of bayesian model testing clear and the difference with p-values. I expect the exact wording for the interpretation.

(iv)

A suitable algorithm is MCMC with prototype as seen in class. I expect that students discuss the choice of proposal density (symmetric or not) and the regression step. I also expect some comments on thinning, conditions for Metropolis-Hastings (for instance), and convergence.

Here I am testing whether they are familiar with the process behing generating samples using iterative methods.

(v-a)

This question relates to one additional reading on SMC and I expect students to recognize the adaptive set of tolerances and the need to monitor convergence especially for model selection.

(v-b)

This question related to one additional reading on Bayes networks and I expect students to mention and discuss nodes, links between nodes, set of conditional probabilities, chain rule.

End of paper