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.

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.

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%]

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%]

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:

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

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%]

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

Continues on next page

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%]

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

.

Continues on next page

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%]

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%]

Continues on next page

(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%].

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.

End of paper