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: Wednesday, 28rd March 2017, 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

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

A. Solve ONE of the following exercises [30%]:

(i) Solve the following integral:

$$I = \int \frac{7x - 6}{x^2 + x - 6} dx$$

(ii) Obtain the Taylor's expansion of the following function at $x_0 = 0$ up to order three:

$$f(x) = \log \sqrt{\frac{1+x}{1-x}}$$

B. Solve ONE of the following sets of exercises [70%]

(i) Consider the following first order differential equation (ODE):

$$a)\ y' = \frac{y^2 - x^2}{xy}.$$

(a) Obtain the general solution $y = f(x, C)$, being C a constant. (40%)

(b) Obtain the value of C for the particular solution $y(e^2) = \sqrt{4}e^2$, where e is the Euler number (i.e. the base of the natural logarithm). (15%)

(c) Explain what the domain of the particular solution $y = f(x)$ you obtained in the previous step is. (15%)

(ii) Consider the following matrix:

$$A = \begin{pmatrix} 1 & -1 \\ 0 & -2 \end{pmatrix}$$

i. Diagonalize the matrix A, and find the matrix P such that $D = P^{-1}AP$, being D diagonal. (40%)

ii. Find the matrix P^{-1}. (15%)

iii. Verify the Cayley-Hamilton theorem, which states that any square matrix is a root of its characteristic polynomial. (15%)

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.

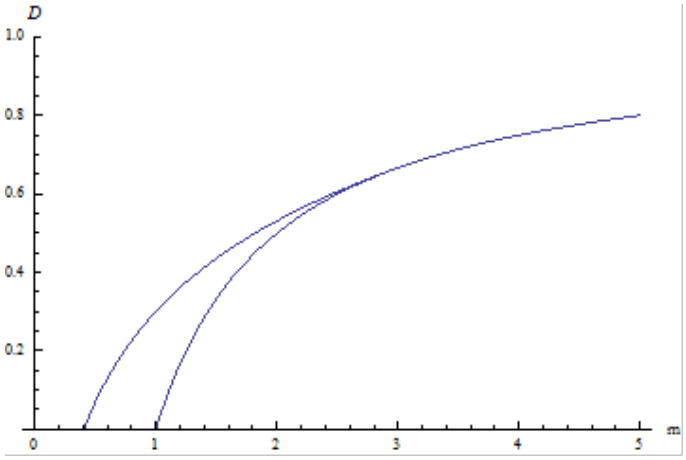
A. A model for a Levins' metapopulation with habitat destruction and in which a rescue effect operates is given by the equation:

$$\frac{dp}{dt} = mp(1 - D - p) - e_0 p e^{-ap}$$

Where p is the fraction of patches that is occupied, e_0 the extinction rate, e is the base of the natural logarithm, m the basic colonisation rate, and D is the fraction of habitat this is destroyed and not available for colonisation or occupation.

(i) Show in a graph how the colonisation and extinction rates per patch depend on p, the fraction of patches that is occupied. Indicate the values that these rates take when crossing the ordinate (i.e. when $p = 0$) [35%]

(ii) Below is a co-dimension 2 bifurcation diagram in parameters m and D for $a = 3$ and $e_0 = 1$. Bifurcations from negative equilibria are not shown. The lines form the boundaries of 3 different regions with different qualitative dynamics. What are the phase portraits in the 3 different regions? Remember that as we have only one variable in this model, the phase plane is one dimensional.[35%]



(iii) Starting from an extant metapopulation in which all habitat is available (i.e. $D = 0$) and assuming, $a = 3$ and $e_0 = 1$, what will you observe if the amount of habitat destroyed is gradually increased? Are these changes reversible?[30%]

B. Before 1997 New Zealand had an immunisation programme against measles in which children were vaccinated with a first dose at age 15 months and a second dose at age 11 years. The programme used the MMR vaccine, which gives life-long protection. The vaccine had an efficacy of 90% (90% of those who received the vaccine were actually protected) and it was assumed that the vaccination scheme had a coverage of 80% (meaning that the vaccine was successfully administered to 80% of the target age class). This would theoretically result in immunisation of 92.16% of the age group over 11 (at first vaccination a fraction of $0.8 \times 0.9 = 0.72$ would be protected, at the second vaccination a further $0.28 \times 0.8 \times 09$). The basic reproductive number, R_0, was estimated to be 12.5. In 1996 mathematical modellers predicted that this immunisation scheme would be insufficient to eliminate measles, and indeed, in 1997 a measles outbreak started.

(i) Explain why you think this immunisation programme would not eliminate the disease [35%]

(ii) Describe in words what happens to number of susceptible individuals over the populations over the years [30%]

Continues on next page

(iii) Recommend changes to the immunisation programme [35%]

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. Answer the following:

(i) Suppose there is an infinite, random-mating population with 2 alleles, A and B, with frequencies p and q, respectively, with the diploid genotypes at Hardy-Weinberg equilibrium. Suppose in AB heterozygotes the B allele is transmitted to a proportion d of gametes (where $d > 0.5$ for gene drive). In addition, BB homozygotes are lethal and produce no gametes. Construct a table to calculate the expected frequencies of allele B in the next generation. [40%]

(ii) Show that the equilibrium frequency of allele B depends on d only. [30%]

(iii) Suppose we have written a genetic drift simulator in `R`. The simulator takes three arguments: p_0, the initial allele frequency of the allele; N, the effective population size; and t, the number of generations you wish to simulate forward in time. Explain how you can estimate the variance of allele frequency due to genetic drift in the next generation, given the current p_0 and N. [30%]

B. An engineering company is designing a bioreactor that breaks down input biomass waste (e.g. wood pulp from a paper mill) and produces ethanol that can be used for fuel. Liquid flows into the reactor containing wood pulp and liquid flows out containing the ethanol at a constant rate. The reactor contains a complex mixture of different bacteria species that together perform the metabolic steps converting cellulose into ethanol. They have asked you to produce a theoretical model to help them design an optimal system.

(i) Sketch out graphically or with equations your approach for modelling the conversion of cellulose into ethanol [35%].

(ii) What key features of the system would you tweak to improve the concentration of ethanol in the outflow [35%]?

(iii) Two species compete for an intermediate substrate (e.g. glucose): a beneficial species that converts it to ethanol and a problem species that converts it to methane, which is undesirable. What options would you explore to control or reduce the effects of the problem species [30%]?

Continues on next page

Section 4: Maximum Likelihood & GLMs

Please select exactly **one question** and answer it. Calculator may be required in some questions. Use the chi-square table below for critical values:

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

A. Answer the following:

(i) Let X be a random variable following exponential distribution with rate parameter $\lambda > 0$. Given the probability density distribution $f_X(x) = \lambda e^{-\lambda x}$ and the moment generating function $M_X(t) = \frac{\lambda}{\lambda - t}$ for some t, show that $E[X] = \frac{1}{\lambda}$and $Var[X] = \frac{1}{\lambda^2}$ [30%]

(ii) After fitting a linear regression model with a slope and an intercept (and also the variance of the residuals), a student suggests to conduct a Likelihood-Ratio test (LRT) to test whether the intercept is significantly different from zero. Please describe carefully the procedures of the LRT, and how a conclusion can be drawn based on the chi-square table provided. [40%]

(iii) Explain, as precise as possible, that how you would construct the 95% confidence interval of your maximum likelihood estimates under approximate normality. You may use appropriate equations, graphics, or R commands to support your answer. Please discuss both univariate (one parameter) and multivariate (multiple parameters) cases. [30%]

B. You have used data on house sparrows to see whether males were consistent in the proportion of offspring they are cuckolded with (extrali-pair offspring, EPO in the brood they care for). So, across a males lifetime, you counted the offspring a male had with their respective social partner (within-pair offspring, WPO), and those that he cares for, but did not sire himself in the same nest (EPO). You then estimated the repeatability of the number of EPO a male finds in its nest, within a year.

To do this you run a Generalized linear mixed model (GLMM). You add no fixed effects, but with male id as a random effect on the intercept. However, because the EPOs are count data, you run it with a logit link function. That also means that we have to calculate the repeatability differently from what we normally do. In Gaussian models, repeatability is the ratio of the variance explained by individual identity over the total phenotypic variance:

Eq 1: $R = \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \sigma^2_\varepsilon}$

were σ^2_α is the variance of the random effect of the MaleID and σ^2_ε is the variance of the residual variance.

However, for the logit-link model, we calculate the link-scale repeatability RL as:

Eq 2: $R_{\text{Link}} = \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \sigma^2_\varepsilon + (\pi^2/3)}$

were π is the constant pi and equals 3.14.

The repeatability on the original scale for the logit-link model needs to take into account overdispersion, and that is captured by including the fixed effects parameter estimates. The equation to estimate the repeatability on the original scale is:

Eq 3: $R_{\text{Original}} = \frac{\sigma^2_\alpha P^2/(1+\exp(\beta_0))^2}{(\sigma^2_\alpha + \sigma^2_\varepsilon)P^2/(1+\exp(\beta_0))^2 + P(1-P)}$

with Eq 4: $P = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}$

The result from this is $R_{\text{Original}} = 0.10$ (95% CI = 0.08–0.12)

Continues on next page

You get the following output from using `MCMCglmm()`:

```
model<-MCMCglmm(EPO~1, random=~MaleID, data=d, family="poisson",
nitt=500*1000, thin=10*10, burnin=3000*10, verbose=FALSE)

Iterations = 30001:499901
Thinning interval  = 100
Sample size  = 4700
DIC: 4370.051
G-structure:  ~PID

post.mean l-95% CI u-95% CI eff.samp
MaleID     2.10    1.616     3.18     2862

R-structure:  ~units

post.mean l-95% CI u-95% CI eff.samp
units    0.60   0.3673   0.8779     2356

Location effects: cbind(EPO, WPO) ~ 1

post.mean l-95% CI u-95% CI eff.samp  pMCMC

(Intercept)    -2.00   -2.428   -1.887     3678 <2e-04 ***
```

(i) Verbally describe what each of the components of the output are [25%].

(ii) Roughly estimate the link-scale repeatability of a male being cuckolded, and write down how you do that. Report both the link scale, and the original scale repeatability as you would in a paper, and interpret the original scale repeatability biologically.[40%]

(iii) Verbal justify why we use the logit link function, and why this means we have to estimate two different repeatabilities. [35%]

Continues on next page

Section 5: Bayesian statistics

This section has **one compulsory question** worth 70% of the total mark. The remaining 30% will be assessed based upon your submission of the practical given to you previously in class.

During your latest field trip in Costa Rica you observed how brightly coloured poison dart frogs (part of the *Dendrobatidae* family) were. In fact, the brightness of their skin colouration is correlated with their toxicity. To investigate the prevalence of toxic frogs in the area under study, you collected n samples of poison dart frogs and observed that k of them have bright skin colour (and thus are toxic). We want to estimate the **population** frequency of the red colour phenotype, $f \in [0, 1]$.

(i) Assuming a generic likelihood function $p(k|f, n)$, where k is our observed data, and prior distribution $p(f)$, write the expression for the posterior distribution of f, $p(f|n, k)$. Please indicate the interval for the integration over f explicitly. [10%]

If we define the likelihood function as a Binomial distribution:

$$p(k|f, n) = \binom{n}{k} f^k (1 - f)^{n-k} \tag{1}$$

and the prior function as a Beta distribution $B(\alpha, \beta)$:

$$p(f) = \frac{1}{B(\alpha, \beta)} f^{\alpha-1} (1 - f)^{\beta-1} \tag{2}$$

then the posterior distribution of f is a Beta distribution with parameters $\alpha' = k+\alpha$ and $\beta' = n-k+\beta$.

(ii) What is the frequentist estimate of f? What is the maximum likelihood estimate of f? What is the maximum *a posteriori* mode of f using the noninformative conjugate prior $p(f) \sim B(\alpha = 1, \beta = 1)$? [15%]

(iii) Assuming we collected 100 samples and 35 of them have bright skin colour, please complete the `R` code below (fill in the `'???'`s) in order to generate both the exact and approximated posterior distribution of f using the informative prior $p(f) \sim B(\alpha = 2, \beta = 1)$. [20%]

```
# we evaluate our parameter f over a grid of 100 values for the whole range [0,1]
f <- seq(0, 1, ???)

# suppose we collected 100 samples and 35 of them have bright skin colour
k <- 35
n <- 100
# alpha and beta are the parameter values for the posterior Beta distribution
alpha <- ???
beta <- ???

# we now evaluate the density function to obtain the EXACT posterior distribution
y <- ???(???, shape1=alpha, shape2=beta)

# we now use Monte Carlo sampling to obtain the APPROXIMATED posterior distribution
    . Make a reasonable choice for the number of random samples.
y_sampled <- ???
y_sampled_distribution <- ???
```

(iv) If we use a Normal distribution as prior information, such as $p(f) = N(\mu, \sigma^2)$, we cannot derive a closed form and cannot sample directly from the posterior distribution. We can use a rejection sampling algorithm for *indirect* sampling of the posterior distribution. This algorithm requires the identification of an envelope function $g(f)$ and a constant $M > 0$ such that $p(k|f, n)p(f) < Mg(f)$. Identify both a suitable envelope function $g(f)$ and a value for M assuming that we know that the maximum density value for the posterior distribution is K. Describe what happens to the algorithm and/or the approximation if we choose $M >> K$ or $M << K$. [25%]

End of paper