Exam preparation

BIO-369, Randomness and information in biological data

Prof. Anne-Florence Bitbol
EPFL

This exam preparation sheet is an actual exam that was given for this class in a previous year. The exact questions and the exact formatting were retained. The duration of the exam is 3 hours. The only document that is allowed is a two-sided A4 sheet of manually written notes. Calculators are allowed, but no other electronic device is allowed.

In this exam, all 5 problems are fully independent from one another. Moreover, the 3 parts of problem 5 are fully independent from one another too. The indicative number of points allocated to each problem is given in the title of the problem.

For reference, practical points were: Please write down all your answers in the space just below the corresponding question. If you need more space, please attach an extra sheet, and clearly indicate what answers are provided on this extra sheet.

1 Luria-Delbrück experiment – 6.5 points / 43

a) Under the "Lamarckian" hypothesis, mutations giving phage resistance occur in response to exposure of the bacteria to phage. Assume that they occur with a probability μ in each bacteria exposed to phage. What is the probability distribution of the random variable X characterizing whether one bacteria is resistant or not?

b) Now consider that a culture containing N bacteria is exposed to phage. For now, do not assume that μ is small. What is the probability $P(m)$ that m of them become resistant under the "Lamarckian" hypothesis? What is the name of this probability distribution?

c) What does this distribution become if $\mu \ll 1$ and $N \gg 1$ but $N\mu = \lambda$ is of order one? Give the name of this probability distribution and the corresponding expression of $P(m)$ as a function of m and λ. What are the mean and variance of this distribution? Make the calculation explicitly for the mean.

d) Explain what is different for the probability distribution of m under the "Darwinian" hypothesis where phage resistance mutations can be acquired randomly at any time during growth of the bacterial cultures (before exposure to phage).

e) Now assume that the number m of resistant bacteria is counted in many replicate cultures. What method would you suggest to quantitatively compare the two hypotheses and decide which of these two models best describes the data? What quantity would you study?

2 Protein sequences – 4 points / 43

Consider a multiple sequence alignment comprising the sequences of homologous proteins. Each row of the alignment is a sequence and each column is a site. The alignment can be viewed as a matrix with N rows and L columns. Assume that the alignment contains proteins that are very similar to one another, and that we reduce it to binary data, with a 0 if the amino acid present in the sequence at this site corresponds to the consensus one, and a 1 otherwise.

a) How would you construct the covariance matrix of sites? What size would it have?

b) How would you construct the covariance matrix of sequences? What size would it have?

c) If you applied principal component analysis (PCA) to the covariance matrix of sites, and used this to represent the data in two dimensions, what would each marker in the plot represent? What kind of information could PCA give in this case?

d) Same questions for the covariance matrix of sequences.

3 Evolution and SARS-CoV-2 variants – 9.5 points / 43

a) Recall that the Moran process allows to describe the evolution of the composition of a population with fixed size N and with two types of individuals, noted A and B. At each discrete time step, an individual is chosen to divide and an individual is chosen to die. In this framework, how can the number i of individuals of type A change after one discrete time step? Give all the possibilities, and explain how they are obtained (it is not necessary to give the respective probabilities of all these cases).

b) Describe the trajectory of i versus time in the Moran process. How is such a trajectory called? Give a different biological example giving rise to this phenomenon, and explain the differences and the similarities with the Moran process.

c) In the Moran process, how would the trajectory of i versus time be impacted by the respective fitnesses of A and B? After a sufficiently long amount of time, what will happen? Can the two types coexist indefinitely? How are the outcomes impacted by the respective fitnesses of A and B (it is not necessary to give the respective probabilities of all cases)?

d) Now consider Fig. 1, which shows the composition of the population of SARS-CoV-2 viruses infecting patients in the UK versus time. What biological event produces new variants? Is it described in the framework of the Moran process? How does this affect the composition of the population?

e) In Fig. 1, what was roughly the frequency of variant Alpha on 2020-11-21? What was it roughly in February 2021? What do you think this time evolution means about variant Alpha? Compare its trajectory to that of the variants that were present on 2020-08-09.

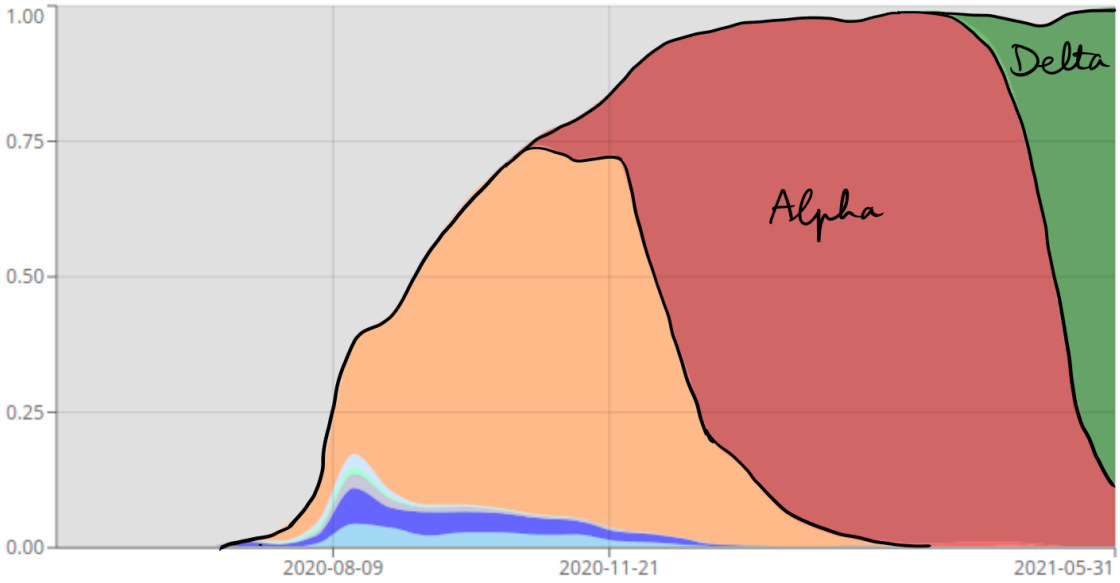

Figure 1: **SARS-CoV-2 variants in the UK.** Composition of the population of SARS-CoV-2 viruses infecting patients in the UK versus time, from sequencing data. Each color represents a different variant strain. Variants Alpha and Delta are labeled by their names. The x axis is time, and the y axis is frequency. *Reproduced and adapted from* `https://covariants.org/`.

f) On 2021-05-17, the frequencies of the detectable SARS-CoV-2 strains in the UK, ranked in decreasing order, were: 0.73; 0.26; 0.01. What variant had frequency 0.73 on 2021-05-17? What variant had frequency 0.26 on 2021-05-17?

g) On 2020-08-17, the frequencies of the detectable SARS-CoV-2 strains in the UK, ranked in decreasing order, were: 0.61; 0.21; 0.07; 0.04; 0.03; 0.02; 0.02. As stated above, on 2021-05-17, they were: 0.73; 0.26; 0.01. Compute the values of the entropy of the population composition on 2020-08-17 and on 2021-05-17. Give both the formula that you used, explaining notations, and the values you obtained. What does entropy mean for a population of viruses? Compare the two values obtained and discuss.

4 Diagnosing the strep throat disease – 8.5 points / 43

Strep throat (streptococcal pharyngitis) is a bacterial infection that may cause a painful and scratchy throat. Initial symptoms resemble those of a regular sore throat from a cold. Treatment by antibiotics is important to reduce complications, which can affect the heart or the kidneys.

Imagine that you are a medical doctor. You examine a patient and you strongly suspect that this patient has strep throat. More precisely, you know that 80% patients with the exact symptoms of your patient actually have strep throat: we will denote this quantity by $P(s)$, the probability that a patient with such symptoms is sick with step throat. On the other hand, given the rise of antibiotic resistance, you are concerned with not prescribing antibiotics in cases where they are useless (in the case of a cold, they would be useless). Therefore, you ask the patient to take a simple test, which is a throat swab.

The throat swab is such that if a patient is sick with strep throat, then in 70% of cases, the test is positive. If a patient is not sick with strep throat, then in 90% of cases, the test is negative.

a) Recall that sensitivity α is the proportion of patients who test positive among sick patients, while specificity β is the proportion of patients who test negative among healthy patients (here "healthy patients" means patients who do not have strep throat). Express the sensitivity α and specificity β in terms of conditional probabilities, and give their values for the throat swab.

b) Suppose that your patient tests positive. Express the probability that this patient is sick with strep throat as a function of α, β and $P(s)$, showing your intermediate calculations. Calculate its numerical value, and comment on it.

c) You decide to be cautious and to ask this patient to perform not one, but several tests. Suppose that two successive tests are performed on a patient, and denote their results by t_1 and t_2. Consider the probability $P(t_1, t_2|s)$ that a patient who is sick with trep throat obtains results t_1 and t_2. Show that it can be expressed as $P(t_1, t_2|s) = P(t_1|s, t_2)P(t_2|s)$.

d) In what follows, we will assume that successive test results (performed on the same patient) can be considered independent from one another (but remember that they depend on whether the patient is sick or healthy). Explain why $P(t_1, t_2|s) = P(t_1|s)P(t_2|s)$.

e) Your patient takes five throat swabs. The successive results are: positive, negative, positive, negative, positive. Express the probability of getting these results if the patient is sick and calculate its value.

f) Similarly, express the probability of getting these results if the patient is healthy and calculate its value.

g) Express and calculate the probability that your patient is sick given these test results. Comment on the result obtained.

h) Now suppose that instead of being the doctor who examined the patient, you are a worker in the laboratory that performs the tests. You know that 50% of the patients who are tested are really sick. What is the probability that the patient considered here, who has taken these five tests and obtained the above results, is sick?

5 Analyzing neuroscience data – 14.5 points total / 43

In Ref. [1], the activity of 40 neurons from the salamander retina was analyzed while visualizing $R = 120$ repetitions of a movie.

5.1 Activity of one neuron – 7.5 points / 43

In this part, we focus on the activity of just one neuron. In Fig. 2, which focuses on one neuron, each marker represents a spike of activity of this neuron. The x axis represents time after the beginning of the movie and the y axis the index of the repetition of the movie. The right panel is a zoom over an area of the left panel corresponding to the times 0.2 to 0.7 seconds after the beginning of the movie and to the first 6 repetitions of the movie.

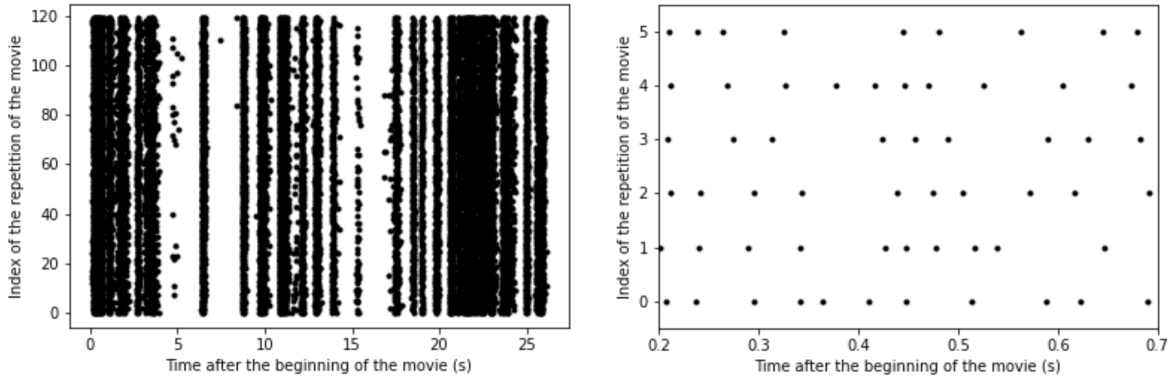

Figure 2: **Activity of one neuron from the salamander retina versus time in the movie.** Each marker represents a spike of activity of the neuron. The left panel shows the complete data for this neuron. The right panel is a zoom over an area of the left panel corresponding to the times 0.2 to 0.7 seconds after the beginning of the movie and to the first 6 repetitions of the movie. *Data from Ref. [1].*

a) In order to discretize this data, we need to choose a time bin $\Delta\tau$ such that we will evaluate neuron activity for each time bin during the movie. For each repetition of the movie, we assign to a given time bin the number of times the neuron spiked during this particular time bin in this particular repetition of the movie. For instance, if during the first repetition of the movie, the first neuron fired 0 times in the first time bin and 1 time in the second time bin, then the discretized data will be "01..." for this neuron in the first repetition of the movie. Based on Fig. 2, and assuming that activity is never more frequent than in the right panel, is the discretized data going to be binary (i.e., consisting only of 0 and 1) if we choose $\Delta\tau = 100$ ms? What about $\Delta\tau = 5$ ms?

b) In what follows, we will assume that $\Delta\tau$ has been chosen such that the discretized data is binary. This means that spike counts are 0 or 1 in all cases. Let us focus on one particular time bin, and denote by p the probability that the neuron is active in this time bin. Express the entropy of the activity of one neuron in this time bin. What is the maximum possible value that this entropy can take? When is it obtained? Same questions for the minimum possible value.

c) For one particular time bin, the neuron is active in 110 movie repetitions out of 120. Compute the entropy of neuron activity in this time bin. Same question in a time bin where the neuron is active in 10 movie repetitions out of 120.

d) What is a small value of the entropy in a given time bin telling us about neuron activity? What about a large one? What is the heterogeneity of the values of entropy in different time bins but for the same neuron telling us about neuron activity?

e) The activity of the neuron in one particular time bin is modeled by a binary random variable with probability p of being active and probability $1 - p$ of being inactive. Write down the probability that the neuron is active in this time bin in m repetitions of the movie out of N total repetitions. What is the name of this probability distribution?

f) What is the likelihood that the neuron is active in this time bin in m repetitions of the movie out of N total repetitions (meaning N measurements of the activity of this neuron) given the model above and the value p? Find the maximum-likelihood estimate of p as a function of m and N.

5.2 Activities of two neurons – 2.5 points / 43

We now focus on two neurons, and consider their activities in all time bins and all repetitions of the movie. There are B time bins and R repetitions of the movie. The activity of neuron 1 in a given time bin of a given repetition of the movie is denoted by a binary random variable X that can take values 0 and 1, and the activity of neuron 2 in this time bin of this repetition of the movie is denoted by a binary random variable Y that can also take values 0 and 1.

a) Write down the expression of the mutual information $I(X;Y)$ between these two random variables as a function of probabilities.

b) What is the minimum possible value of $I(X;Y)$? In what case would it be reached?

c) What does the value of $I(X;Y)$ inform us about? In particular, what would a large value of $I(X;Y)$ mean for the two neurons considered?

5.3 Maximum entropy model of neuron activity – 4.5 points / 43

In this section, we choose to make the following mapping: a value 1 is associated to a spiking neuron and a value -1 (instead of 0 before) to an inactive neuron. Hence, the activity of the neuron in a given time bin of a given repetition of the movie is represented by a random variable that can take values in $\{-1, 1\}$.

a) Let us focus on the activity of one neuron, and let us consider the associated random variable X. Imagine that you have measured this mean neuron activity over time experimentally and found a value m. What is the maximum-entropy probability distribution $P(x)$ of X consistent with this measurement? Explain the derivation. What constraints should you employ to find the values of the parameters involved in $P(x)$? It is not necessary to solve the equations corresponding to these constraints.

b) Now we focus on the activities of two neurons, and on the associated random variables X and Y. Imagine that you have measured the mean neuron activities over time experimentally and found a value m for the first neuron and n for the second one. Imagine that you have also measured the mean of the product XY. What is the joint maximum-entropy probability distribution $P(x, y)$ consistent with these measurements? It is not necessary to write down the derivation.

c) For N neurons indexed by i ranging from 1 to N, to which the random variables $X_1, X_2, \ldots, X_N$ are associated, the maximum entropy distribution consistent with the measurements of the mean activity of each neuron and of the mean products of the activities of all pairs of neurons reads

$$P(x_1, x_2, \ldots, x_N) = \frac{1}{Z} \exp\left[-\sum_{i=1}^{N} h_i x_i - \sum_{i=1}^{N} \sum_{j<i} J_{ij} x_i x_j \right] . \tag{1}$$

How would you compute the values of Z, h_i and J_{ij}? Do you think this calculation is easy or difficult in practice?

d) In Fig. 3, we show the inferred values of J_{ij} for an ensemble of 10 neurons studied in Ref. [1]. What do you think a large absolute value of J_{ij} mean for a pair of neurons i, j? What about a small one?

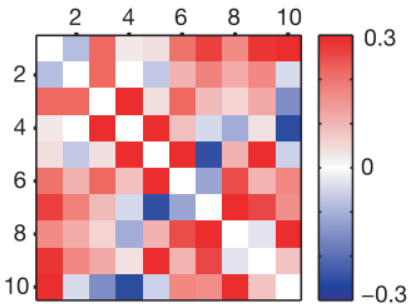

Figure 3: **Parameters J_{ij} from maximum entropy inference.** The inferred values of J_{ij} for a set of 10 neurons from the salamander retina are shown in color in a matrix, where each element (i, j) of the matrix contains the corresponding J_{ij}. The colorscale is shown on the right. *Reproduced from Ref. [1].*

References

[1] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–1012, Apr 2006.