Computational Biology I: Quantitative Data Analysis
Frühjahrsemester 2019

Exercise 7

Pascal Grobecker, Athos Fiori

How to submit the exercise

Submit your completed `Exercise7.py` file by email to biocomp1-bioz@unibas.ch. The submission deadline is Thursday, 25th April, midnight.

1. Check that your functions are working properly before submitting them!
2. Don't forget to attach your exercise file!
3. Don't change the name of the file or the name of any function within the file!
4. Don't add any additional `import` statements within the file!
5. Please send your solutions from the same email address that you used to registering at the course.

Bacterial persistence

It has been observed for many years that when a culture of bacteria is treated by an antibiotic, the vast majority of cells quickly dies, but a small fraction of the cells somehow survives. These 'persisters' are not mutants that are genetically different from the other cells. If one removes the antibiotic and lets the small number of persister cells grow back into a large culture, and subsequently treats this culture with the antibiotic, then one again finds that the vast majority of cells dies.
This persistence phenomenon is of particular interest in the context of infections; it is thought that these *persisters* form biofilms on tissues which are responsible for many recurrent, chronic infections which fail to respond to even multi-drug treatment. Persistence is a transient phenotype found in any strain of bacteria, and thus there is no single genetic trait responsible for this ability. Instead, it appears that genetically identical cells spontaneously switch back and forth between a 'normal' state in which cells proliferate and are susceptible to the antibiotic, and a persister state in which cells are dormant (not dividing) and are not susceptible to the antibiotic. It is thought that genes which lead to growth arrest may be important in formation of the dormant persister state, and that this is used as a survival strategy on the population level to cope with transient negative changes in their environment. In order to investigate the effect of strain and environment on persister formation, a plating assay can be performed, as illustrated in the figure below.
The number of cells left alive is measured at numerous time points after beginning the treatment with the antibiotic, using counting of colonies. From these measurements, we can estimate so called *killing curves* which show the number of surviving cells as a function of time. When plotting the number of remaining cells as a function of time on a logarithmic scale, then one sees that the kill curve consists of a mixture of two exponentials (showing up as straight lines with different slopes, see Figure 2, left). The first steep slope corresponds to the fast exponential dying of the 'normal' cells, whereas the much slower second exponential corresponds to the slow dying of the persisters.
Given real data, we would like to estimate the death rates of both the normal and persister populations, and the fraction of persisters within our population. In reality inferring these parameters is complicated by several factors, not least of which is the fact that we are estimating the number of live bacteria from a small subpopulation at a few discrete points in time, and not from counting through the whole population continuously over time (see Figure 2, right for examples of experimentally determined killing curves).
We could deal with this by incorporating an error model which accounts for errors in counting coming from the subpopulation sampling, and then integrating over the parameters of this model. However, to simplify the

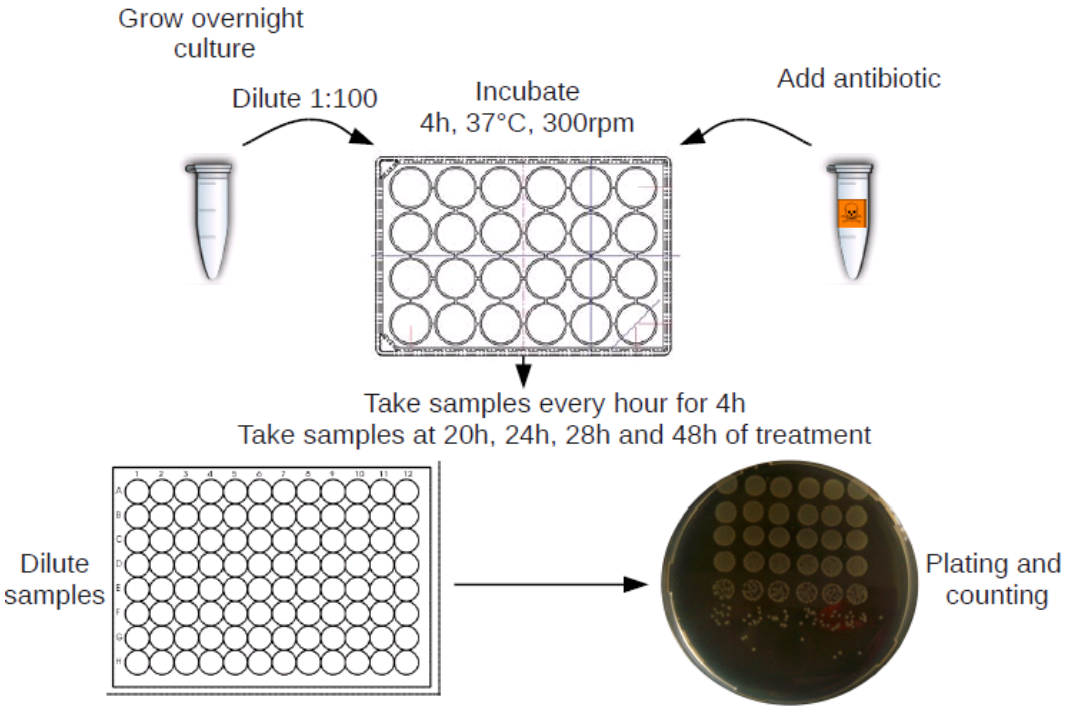

Figure 1: Cultures grown overnight are diluted and incubated in wells together with antibiotic. Samples are periodically taken from these cultures and plated out at different dilutions. Finally, colonies are counted to estimate the number of cells that were still alive at each time point.

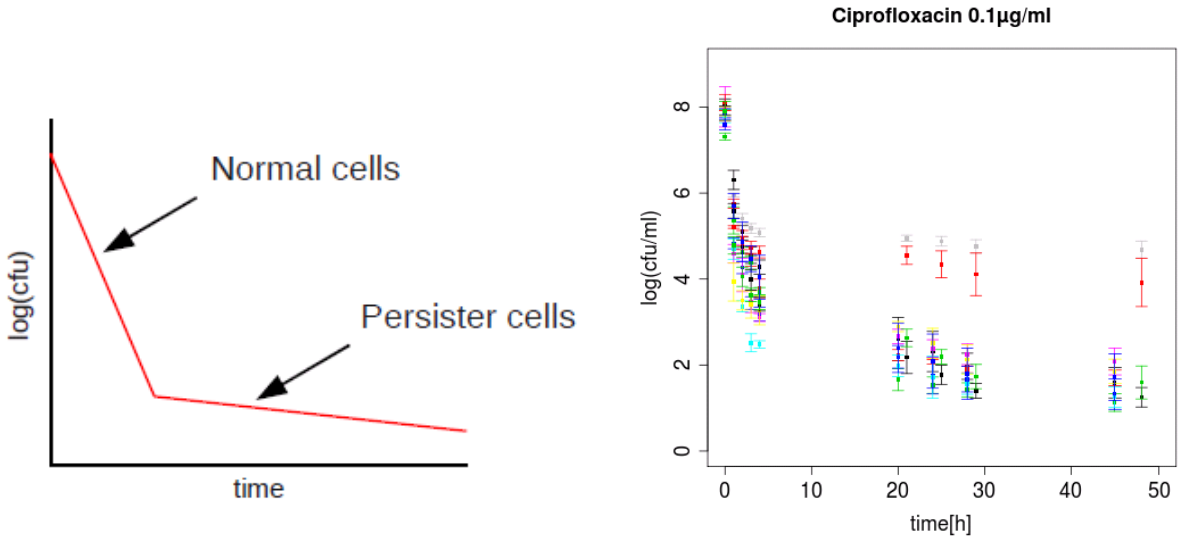

Figure 2: Left: Schematic of killing curve with two distinct population death rates. Right: Killing curves taken from real data.

analysis for this exercise, we will assume here that we have been able to watch every bacteria within our sample and record the time of death of every single one.

1. There is the function `loadDeathData` to load the death times of the bacteria given in the data file `data1.dat`. Try to understand how the function is working.

2. The function `plotDeathCurve` allows you to plot a death curve given a vector of death times. The function plots a graph of the remaining fraction of living cells against time. This is very similar to plotting a reverse-cumulative. Try to understand how the function works by looking at its code! Plot the death curves of the three data sets. How do the three data sets differ? (Answers not to be submitted)

3. Assuming the data comes from a mixture of two exponentially distributed signals with different shape parameters λ_i, the likelihood of the data is:

$$P(D|\rho,\lambda_1,\lambda_2) = \prod_i \left(\rho\lambda_1 e^{-\lambda_1 t_i} + (1-\rho)\lambda_2 e^{-\lambda_2 t_i}\right),$$

   where t_i is the time at which cell i died, and ρ is the fraction of persisters. Note also that λ_1 and λ_2 are the slopes of the lines shown in Figure 2, left. Within the function `logLikelihood` implement the log-likelihood function $L(D|\rho,\lambda_1,\lambda_2)$ in terms of a linear combination of the two functions $L_1(t_i|\lambda_1)$ and $L_2(t_i|\lambda_2)$. Make use of the pre-defined exponential PDF function!

4. We are going to use the expectation-maximisation (EM) algorithm in order to infer the unkown parameters ρ, λ_1 and λ_2. As you remember from last exercise, the EM-algorithm is an iterative procedure. In every step of the iteration the parameter values of the previous step are used to calculate the current likelihood (expectation step). We then maximize this likelihood to get updated parameter solutions (maximization step). The iteration continues until the value of the likelihood converges, i.e. there is no significant change in the likelihood between update steps. In order to set up the EM algorithm, you need to construct the EM update equations. Thus, you need to write down the equations for the values λ_i^*, and ρ^* which maximize the likelihood by solving the equations

$$\frac{\partial L(D|\rho,\lambda_1,\lambda_2)}{\partial\rho} \stackrel{!}{=} 0, \quad \frac{\partial L(D|\rho,\lambda_1,\lambda_2)}{\partial\lambda_i} \stackrel{!}{=} 0$$

   As you know from exercise 6, these equations will rely on the probabilities $p_{j,i}$ that the observed death time of bacteria i derives from the distribution $P(t_i|\lambda_j)$, i.e.,

$$p_{1,i} = \frac{\rho L_1(t_i|\lambda_1)}{\rho L_1(t_i|\lambda_1) + (1-\rho)L_2(t_i|\lambda_t)}$$
$$p_{2,i} = \frac{(1-\rho)L_2(t_i|\lambda_1)}{\rho L_1(t_i|\lambda_1) + (1-\rho)L_2(t_i|\lambda_t)}.$$

   Complete the function `EM` which implements the EM algorithm for the data model given above. Make use of the pre-defined exponential PDF function! Note that the function returns for each parameter an array with the value at each step of the iteration.

5. Use the EM algorithm to estimate the unkown parameters ρ, λ_1 and λ_2 from the data. What do you observe for the λ_i values if you use different starting conditions? You can use the function `plotDataAndModel` to visualize your solutions togther with the data. (Answers not to be submitted)

6. In order to determine which cells where persistors we need to calculate the posterior probability that the observed death time is coming from the distribution of persistor death times. Note that the fraction of persistors is allways small, i.e., the persistors distribution corresponds to p_2 if $\rho < 0.5$ and p_1 otherwise. Implement your answer in function `persistors`.

7. We would like to know how reliable our parameter estimates are by obtaining the covariance matrix. As explained in the lecture, we obtain this by looking at the peak around our optimal parameter estimates as a multivariate Gaussian. Then, the covariance matrix is just the inverse of the negative Hessian matrix of second-order partial derivatives of the likelihood function with respect to our parameters $\theta = (\rho, \lambda_1, \lambda_2)$, i.e.,

$$C = (-H)^{-1},$$

where

$$H_{ij} = \frac{\partial^2 L(x|\theta)}{\partial\theta_i \partial\theta_j}.$$

If we define

$$f_i(\rho, \lambda_1, \lambda_2) = \left(\rho\lambda_1 e^{-\lambda_1 t_i} + (1-\rho)\lambda_2 e^{-\lambda_2 t_i}\right)$$

then the first- and second order derivatives of the likelihood function are:

$$\begin{aligned}
\frac{\partial L(D|\rho,\lambda_1,\lambda_2)}{\partial\rho} &= \sum_i \frac{L_1(t_i|\lambda_1) - L_2(t_i|\lambda_2)}{f_i} \\
\frac{\partial L(D|\rho,\lambda_1,\lambda_2)}{\partial\lambda_1} &= \sum_i \frac{\rho(1-\lambda_1 t_i)e^{-\lambda_1 t_i}}{f_i} \\
\frac{\partial L(D|\rho,\lambda_1,\lambda_2)}{\partial\lambda_2} &= \sum_i \frac{(1-\rho)(1-\lambda_2 t_i)e^{-\lambda_2 t_i}}{f_i} \\
\frac{\partial^2 L(D|\rho,\lambda_1,\lambda_2)}{\partial\rho^2} &= -\sum_i \frac{(L_1(t_i|\lambda_1) - L_2(t_i|\lambda_2))^2}{f_i^2} \\
\frac{\partial^2 L(D|\rho,\lambda_1,\lambda_2)}{\partial\rho\partial\lambda_1} &= \sum_i \frac{(1-\lambda_1 t_i)e^{-\lambda_1 t_i}\lambda_2 e^{-\lambda_2 t_i}}{f_i^2} \\
\frac{\partial^2 L(D|\rho,\lambda_1,\lambda_2)}{\partial\rho\partial\lambda_2} &= \sum_i \frac{-(1-\lambda_2 t_i)e^{-\lambda_2 t_i}\lambda_1 e^{-\lambda_1 t_i}}{f_i^2} \\
\frac{\partial^2 L(D|\rho,\lambda_1,\lambda_2)}{\partial\lambda_1^2} &= \sum_i \frac{-\rho^2 e^{-2\lambda_1 t_i} + \rho(1-\rho)\lambda_2 t_i(-2+\lambda_1 t_i)e^{-\lambda_1 t_i}e^{-\lambda_2 t_i}}{f_i^2} \\
\frac{\partial^2 L(D|\rho,\lambda_1,\lambda_2)}{\partial\lambda_2^2} &= \sum_i \frac{-(1-\rho)^2 e^{-2\lambda_2 t_i} + \rho(1-\rho)\lambda_1 t_i(-2+\lambda_2 t_i)e^{-\lambda_1 t_i}e^{-\lambda_2 t_i}}{f_i^2} \\
\frac{\partial^2 L(D|\rho,\lambda_1,\lambda_2)}{\partial\lambda_1\partial\lambda_2} &= -\sum_i \frac{(1-\rho)(1-\lambda_2 t_i)e^{-\lambda_2 t_i}\rho(1-\lambda_1 t_i)e^{-\lambda_1 t_i}}{f_i^2}
\end{aligned}$$

Using the above equations, complete the function `EMcovariance` which constructs the Hessian matrix H and uses it to calculate the covariance matrix C of the likelihood function at the optimum parameter estimates. To perform the matrix inverse, use the function `inv`.

8. The covariance matrix is an indicator of whether our parameters are *identifiable*: whether we can be confident in our estimates of a parameter given the data, or whether our data could be equally well explained by another set of parameter values. The correlation matrix is a normalized covariance matrix whose values span the range $-1 \leq \text{corr} \leq 1$ and a value of 0 indicates no (linear) dependency. Complete function `parameterCorrelation` which computes the correlation matrix R given given the covariance matrix C by making use of the formula:

$$R_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}}$$

Additionally, lets define the identifiability of parameter i as:

$$F_i = \sqrt{\frac{1}{N}\sum_j^N 1 - R_{ij}^2}$$

The function also returns the identifiability as defined above for all three parameters $(\rho, \lambda_1, \lambda_2)$.

9. For which data set are the parameters most difficult to identify? What features in the data correspond to larger values in the correlation matrices? What does this correspond to in terms of the death rates of the bacteria? (Answers not to be submitted)

Theoretical questions

---

1. We have performed n independent measurements $D = (x_1, .., x_n)$ for our quantity of interest. We make the assumption that the measurement errors are Gaussian distributed with sandard-deviation σ, which we assume to be a known parameter. Specifically, assuming that the true value of our quantity of interest is μ, the probability to obtain a measurement x is $P(x|\mu, \sigma) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]$. With this, the probability for the entire data-set is

$$P(D|\mu, \sigma) = \prod_{i=1}^{n} P(x_i|\mu, \sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right]$$

   Given the data-set D, and assuming a uniform prior for μ, the posterior distribution $P(\mu|D, \sigma)$ is simply proportional to the likelihood, i.e. $P(\mu|D, \sigma) \propto P(D|\mu, \sigma)$. From the expression above, you should be able to derive that $P(\mu|D, \sigma)$ is itself Gaussian distribution. What are the mean and variance of this posterior distribution for μ?

2. We have sampled n data-points from a Gaussian distribution that has an unknown mean μ and unknown standard-deviation σ. Let $\bar{x}$ denote the mean of the data-points and var(x) the variance of the data-points. We are interested in $P(D|\sigma)$ i.e. we want to get-rid of the μ parameter. Assuming an uniform prior over the mean μ, what's the general form of this distribution?

3. We assume our data $D = (x_1, ..., x_n)$ derive from a mixture of k different Gaussians. For each Gaussian i, with mean μ_i and variance σ_i^2, the probability of obtaining a data-point x is

$$P_i(x|\mu_i, \sigma_i) = \frac{1}{(2\pi\sigma_i^2)^{1/2}} \exp\left[-\frac{(x - \mu_i)^2}{2\sigma_i^2}\right]$$

   We denote ρ_i the fraction of component i in the mixture. If we denote the set of all k means by $\vec{\mu}$, the set of standard-deviations by $\vec{\sigma}$, and the set of fractions by $\vec{\rho}$, then the likelihood is given by

$$P(D|\vec{\mu}, \vec{\sigma}, \vec{\rho}) = \prod_{j=1}^{n} \left(\sum_{i=1}^{k} \frac{\rho_i}{(2\pi\sigma_i^2)^{1/2}} \exp\left[-\frac{(x_j - \mu_i)^2}{2\sigma_i^2}\right] \right)$$

   To find the parameters $\vec{\rho},\ \vec{\sigma}, \vec{\mu}$ that maximise the likelihood, one generally demands that the derivative of the log-likelihood with respect to all parameters equals zero. When one does this, one finds that one can express the conditions for the optimum in terms of consistency equations. In expectation-maximisation these consistency equations are used to iteratively update the parameters. Here we want you to write down these consistency equation for $\rho_i,\ \mu_i$ and σ_i. These consistency equation take on a very simple form if one introduce the notations $P(i|x_i)$ to indicate the probability that data-point x_j derived from component i. That is, we define

$$P(i|x_j) = \frac{P_i(x_j|\mu_i, \sigma_i)\rho_i}{\sum_{h=1}^{k} P_h(x_j|\mu_h, \sigma_h)\rho_h}$$

   write down the consistency equation for $\mu_i,\ \sigma_i$ and ρ_i in terms of the $P(i|x_j)$, x_j and the number of data points n.