Within-host HIV dynamics: estimation of parameters

Level 1 module in "Modelling course in population and evolutionary biology"

(701-1418-00)

Module author: Viktor Müller

Course director: Sebastian Bonhoeffer
Theoretical Biology
Institute of Integrative Biology
ETH Zürich

1 Introduction

An important application of mathematical models is to estimate biological parameters that cannot be measured directly. The basic procedure is to construct a mathematical model that captures the main processes of the biological system and then "fit the model" to real observations. Model fitting works by contrasting observations (empirical data) to the predictions of the model, and finding the set of model parameters for which the difference is the smallest. It is important to keep in mind that parameter estimation depends strongly on the formulation of the model, e.g. models based on alternative hypotheses may yield different estimates from the same data. The "goodness of fit" can then guide our decision to accept or reject a model (and the biological hypothesis behind it) or to choose from alternative hypotheses.

We will illustrate the procedure of parameter estimation on the example of the within-host dynamics of HIV infection[a]. Infection with this virus is typically characterized by a long asymptomatic period that lasts for several years (see Figure 1 for the typical course of the infection). During this time, virus load (the level of the virus in an individual) remains relatively stable, and the infection seems to be "latent". However, the application of the first effective

[a]There are actually two types of the virus: HIV-1 and HIV-2. HIV-1 is responsible for the worldwide epidemic, it has been studied much more intensively and it has originated from a virus of chimpanzees. HIV-2 is largely confined to Africa and has originated from a related virus of sooty mangabeys.

antiviral drugs (introduced around 1995) has revealed a highly dynamic picture of this "latency". Within hours of the start of treatment, the virus load begins to fall rapidly: it falls typically several orders of magnitude within 1-2 weeks. The drugs block the infection of new cells, but do not affect the death of cells that are already infected or the breakdown of virus particles. Therefore, the observed rapid decline reflects the normal decay of infected cells and virus, which is, in the absence of treatment, balanced by equally fast production. The apparent stability of the virus load without treatment thus reflects the almost perfect balance of fast production and decay. The analysis of virus load data from the beginning of treatment allows us to estimate just how fast these processes are.

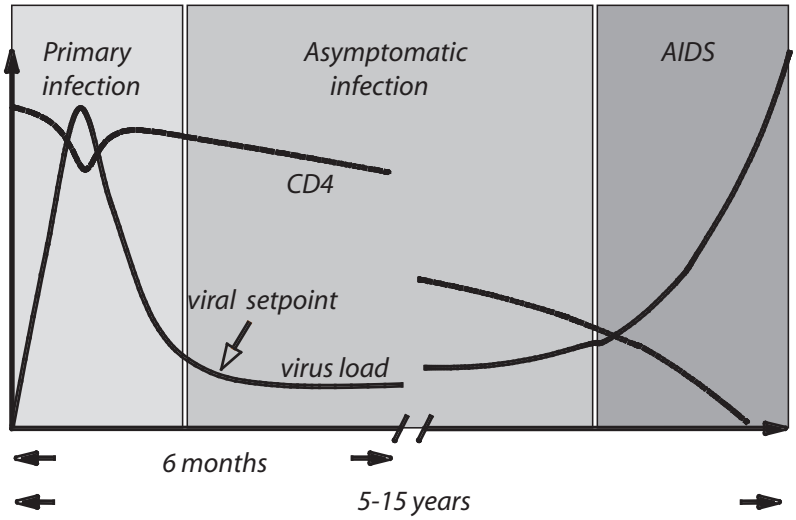

Figure 1: The typical course of HIV-1 infection in an untreated individual. After an individual becomes infected, the virus load rises rapidly to a very high transient peak level, and then settles to a so-called *set-point* level that remains more or less stable for several years. This initial phase of *primary infection* lasts only a few months, and is also characterized by a substantial (though partially transient) decrease in the the CD4+ T cell population, which are the major target cells of HIV and an essential part of the immune system. After primary infection the patient enters the disease-free *asymptomatic phase*. During this stage the virus load remains stable over prolonged periods of time, while the number of CD4 cells continues to decline steadily. Eventually the CD4 cells fall below a critical level (<200 CD4 per μl of blood), which defines the onset of AIDS, and below which the CD4 cells can no longer maintain a functional immune response. At this stage AIDS-defining opportunistic infections typically arise and the virus load may increase sharply.

1.1 Basic model of treatment

We first set up a model to describe the first few weeks of treatment. The model considers virus particles, V, which disappear (break down or are neutralized) at a rate u, and are produced at a rate k by the infected cells, I. Infected cells die at a rate a, and are produced by the infection

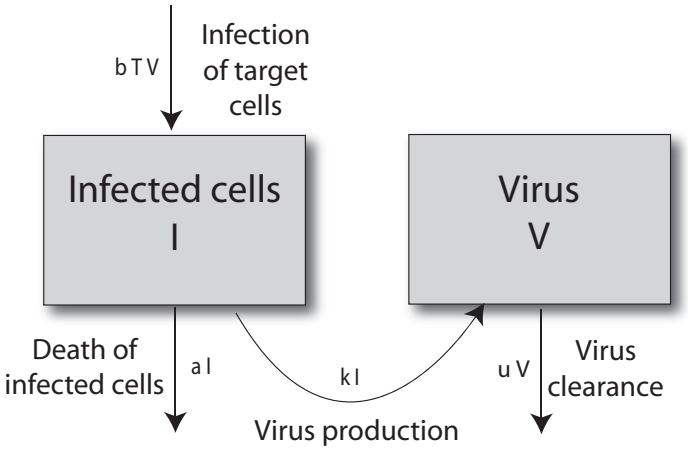

Figure 2: Schematic illustration of the virus dynamics model given by Eqs 1 and 2.

of target cells, T, proportional to the level of virus and to an infectivity rate parameter, b. We assume that the level of target cells does not change in the first few weeks of treatment, we can thus write the following equations:

$$dI/dt = bTV - aI \quad (1)$$
$$dV/dt = kI - uV \quad (2)$$

A schematic illustration for this model is given in Figure 2. As the virus load is relatively stable during the asymptomatic period, we can use the approximation $dI/dt = 0$ and $dV/dt = 0$. From Eq. 1 we then obtain $bTV = aI$, i.e. the death of infected cells is balanced by the infection of new target cells, and from Eq. 2 we obtain that the ratio of virus particles and infected cells is $I = uV/k$ at steady state. Antiviral treatment inhibits the infection of new target cells, which can be modelled by setting the infectivity rate to $b = 0$. This reduces Eq. 1 to simple exponential decay, and a solution for the whole system can be obtained analytically as

$$I(t) = I_0 e^{-at} \qquad \text{and} \qquad V(t) = V_0 \frac{ue^{-at} - ae^{-ut}}{u - a} \quad (3)$$

with $I_0 = uV_0/k$. However, extensions of this model quickly become unsolvable analytically, we will therefore use a simulation approach by numerically integrating the equations. To illustrate the strengths and weaknesses of parameter estimation, let us first use this basic model to generate a simulated data set for treatment, for which we know the "true values" of the parameters. Eqs. 1-2 have been implemented in the R script `treat.R`. Download the script from the web page of the module and run it. Look at the decay of the virus level after the start of therapy.

The simulated data are also written into a file, which we will use in the following as input data for the estimation procedure.

1.2 Parameter estimation

Parameter estimation (model fitting) is performed by finding the set of parameters for which the prediction of the model is closest to the observations. This is done typically by minimizing the sum of squared distances between the measured data and the values predicted by the model (least squares method). Assume we have a set of n measurements $m_1, \ldots, m_n$ at times $t_1, \ldots t_n$. Assume further we have a model M which we want to fit to the data. In general the model will depend on a number of parameters and the time. Then the estimation of the parameters is done by minimizing the following function

$$\mathcal{M} = \sum_{i=1}^{n} (m_i - M(t_i))^2 \tag{4}$$

This means that we are changing the model parameters to minimize the sum of the squared distances between the measured data and the value predicted by the model. The best fit parameters are given by that combination of parameters for which $\mathcal{M}$ is minimal. If the model is linear the best fit parameters can be obtained analytically by performing a *linear regression*. If the model depends nonlinearly then it is more difficult to find the best fit parameters, and it can no longer be guaranteed that the best fit parameters are actually obtained. However, standard statistical software usually has sophisticated routines for fitting of nonlinear models, and `R` can do this for us.

Let us approach the data that we generated by simulation as the scientists approached the empirical data that were collected after the start of the first effective treatment. When plot on a logarithmic scale, it is quite clear that the decline of the virus load follows some sort of exponential decay. Let us first fit a simple exponential decay to the data, i.e. fit the model: $V(t) = V_0 e^{-dt}$. Download and run the R script `estimate.R` to do this. (Note that the script uses a sampling of one data point per day from the simulated data). How can you interpret the estimate obtained for the decay rate d (see footnote[b] for hint)? The script `estimate.R` uses the general nonlinear minimization function `optimize()`. However, on a logarithmic scale, exponential decay behaves as a linear process. Modify the script to minimize the sum of squared distances between the logarithm of the simulated and that of the predicted data. Which estimation is better and why? Then try also linear regression on the log transformed data set. Hint: you can fit a linear model with `linearfit <- lm(log(datapoints) ~ timepoints)` and then extract the intercept and the slope of the regression line with `regcoef <- coef(linearfit)`. Why does linear regression perform better? Try to vary the time points used in the estimation, e.g. take time points from day 2 only.

[b] In a composite decay involving several processes of various rates, the rate of the overall decay is dominated by the slowest of the processes. Look at Eq. 3 and try to explain why this is so.

2 Exercises

2.1 Basic exercises

Eb1. We have observed that fitting a simple exponential decay provides information on the slower of the two decay rates. Now let us fit the model with distinct variables for infected cells and virus (Eq. 3) to the simulated data. To do this, you will need to estimate a and u simultaneously: to do this, you will need the function `optim()`, which can optimize several parameters simultaneously (check out the help page of the function). Sample the data as in the original paper by Perelson *et al.* (1996): take 5 points per day in the first two days, and one point per day in the next five days. Use estimation based on both the raw data and on log transformed data. Notice that the nonlinear minimization function requires an initial guess for the parameters: experiment with providing different values. Note the symmetry of the two decay rates: can model fitting tell you which rate characterizes infected cells and which the virus particles? Vary the number and times of sampling of the data points used in the estimation: how does it affect the precision of the estimates?

Eb2. Calculate also the half-life of free virus and infected cells. The half-life describes the time that an exponentially decaying population requires to fall to half its initial value. The relation between the half-life and the (exponential) decay rate is thus given by $t_{1/2} = ln(2)/a$ for the infected cell population and $t_{1/2} = ln(2)/u$ for the free virus population.

2.2 Advanced/additional exercises

Ea1. Download the original clinical data of patient 104 (see Perelson *et al.*, 1996) from the web page of the module. Perform the estimation procedure on the real data. Can you get the same estimates as in the original paper? Check out the original paper, and try to reproduce the result. Hint: check out the paper by Perelson et al. for the details of their estimation process.

Ea2. Real biological systems are always complex and the effect of the main processes is typically blurred by a number of unknown or at least uncontrolled factors. Such confounding factors add "noise" to any measurement. In this particular case, we can model this by adding normally distributed additive or multiplicative noise to the simulated data. Hint: you can generate normally distributed random numbers with the `rnorm()` function. Generate additive noise by adding random numbers to the data points, and multiplicative noise by multiplying the data points with the random numbers. Repeat the estimation procedure on noisy data. Generate a large number of noisy data series and look at the distribution of the values estimated for the parameters with simple and log transformed fitting. Plot the distributions with the `boxplot()` command. Experiment with the estimation procedure: how many points are required to get reliable estimates for the decay rates? When do you have to measure the virus load to obtain reliable estimates for one or the other decay rate? Given a certain estimation protocol, how many replicates do you need to get a reliable estimation from the mean of the independent estimates? Conclude in the end: what ways are there to improve parameter estimation?

Ea3. How is the estimation affected if the drugs are not 100% effective in blocking new infections? Let ϵ denote the efficacy of treatment, i.e. the fraction of reduction in the infectivity rate. Rephrase Eq. 1 to accommodate this effect, then generate simulated data sets with this model variant. Hint: in the setting of the initial value of bT, use the steady state equation: $bTV = aI$. Repeat the estimation procedure with the standard model (Eq. 3) on the simulated data. How is the estimation affected? Advanced question: can you incorporate <100% efficacy in the estimation, and estimate also efficacy from the data?

Ea4. After the first few weeks of treatment, the decline of the virus load decelerates, which can be explained by the presence of another, long-lived population of infected cells (see Perelson *et al.*, Nature 1997). Extend the model with an equation for such cells, generate new data and repeat the estimation procedure. How is the estimation affected? Repeat the estimation on the simulation of the first week, the first month, or the first year of simulated treatment. Adopt the necessary data from the paper of Perelson *et al.*. Hint: considering the longer time scale, the level of target cells should also be allowed to change. Introduce an equation for the target cells with a constant rate of production: $dT/dt = \lambda - \delta T - bTV$.

Ea5. Try to use the methods that you have learned to estimate the turnover of another virus: HCV. See the papers on the module website for a starting point. Note that this is an advanced exercise which needs additional background research and might not have a "final solution".