Identifiability Lab

Marisa Eisenberg (marisae@umich.edu)

June 27, 2017

Notes: You might not finish all the sections—that's okay! Feel free to choose what problems you find most interesting. I've also posted solution code here: https://github.com/marisae/param-estimation-SIR, so you can see how things work out for any parts you don't finish.

Part 1: Structural identifiability for the SIR model

We will consider a version of the classical SIR model that you've seen in some of the previous lectures:

$$\dot{S} = \mu N - bSI - \mu S$$
$$\dot{I} = bSI - (\mu + \gamma)I$$
$$\dot{R} = \gamma I - \mu R$$

with measurement equation $y = kI$. The variables S, I, and R represent the number of susceptible, infectious, and recovered individuals, and we take y to indicate that we are measuring a proportion of the infected population (e.g. if not all cases are reported). The parameters μ, b, γ, N, and k represent (respectively) the birth/death rate, transmission parameter, recovery rate, total population size, and the proportion of the infected population which is reported/observed.

Enter the model into the web app COMBOS (http://biocyb1.cs.ucla.edu/combos/)[1], and examine its identifiability. When you enter it into COMBOS, you have to name your state variables using x's, so let $x_1 = S$ and $x_2 = I$.

- Are all the parameters for this model structurally identifiable?
- If any are not, what are the identifiable combinations? Why do you think the combinations have this structure?
- What happens if we re-scale the model to be in terms of fractions of the population instead of individuals? In other words, rescale the model to let $s = S/N, i = I/N$, and $r = R/N$. When you do, you will be able to combine some parameters to let $\beta = bN$ and $\kappa = kN$. Rewrite your model equations in this rescaled and reduced parameter form, and re-run it in COMBOS—how does the identifiability look now?

[1] Meshkat N, Kuo CEZ, DiStefano J III (2014) On Finding and Using Identifiable Parameter Combinations in Nonlinear Dynamic Systems Biology Models and COMBOS: A Novel Web Implementation. PLOS ONE 9(10): e110261. https://doi.org/10.1371/journal.pone.0110261

Part 2: Parameter estimation and uncertainty with the SIR model

Now that we understand a little more of the structural identifiability picture for the SIR model, let's estimate parameters and investigate the uncertainty in the parameter estimates.

We will work with the scaled version of the model, where S, I, and R represent the fraction of the population that is susceptible, infectious, and recovered, respectively. We'll also assume that, since the outbreak we consider is over a short timescale, the population birth-death rate is negligible, i.e. let $\mu = 0$ (since there are probably few births/deaths during this timeframe). The equations are given by:

$$\dot{S} = -\beta SI$$
$$\dot{I} = \beta SI - \gamma I$$
$$\dot{R} = \gamma I$$

with the measurement equation is $y = \kappa I$, where κ is a product of two things: the population size (to convert fraction of the population to number of individuals), and the fraction of cases that are reported (observed). Then y represents the number of observed individuals who are currently infectious. The other parameters The parameters β and γ represent (respectively) the transmission parameter and recovery rate.

1) **Model Simulation**. Simulate the SIR model and plot both the data set (the data set is provided) and the measurement equation $y = \kappa I$. Use the following parameter values: $\beta = 0.4$, $\gamma = 0.25$, $\kappa = 80000$. For initial conditions, let $I(0) = data(0)/\kappa$ (where $data(0)$ is the first data value), $S(0) = 1 - I(0)$, and $R(0) = 0$.

2) **Parameter Estimation**. Next, write code to estimate β, γ, and κ using Poisson maximum likelihood and the dataset provided. Use the parameter values in 1) as starting parameter values, and you can use the initial conditions from 1) as well (note though that they depend on κ, which is a fitted parameter—so while we aren't fitting the initial conditions, they will need to change/update as we fit the parameters!). This means you will need to update your initial conditions inside the cost function, so MATLAB/R uses the updated initial conditions when it tries new parameter values.

Plot the data together with your model using the parameter estimates you found. Be sure to plot the data as circles ('o' in the plot function in MATLAB) and your model simulation as a line so that you can compare your model with the data easily. Based on the 'eyeball test', how well does the model fit the data?

(Extra problem: You might also try adapting the code so that you estimate the initial conditions as unknown parameters! If you do, start by fixing $R(0) = 0$, so that you fit $I(0)$ and let $S(0) = 1 - I(0)$. Then try fitting $R(0)$ as well, and see how this affects the identifiability of your system.)

3) **Identifiability with the Fisher Information Matrix (FIM)**. Generate the output sensitivity matrix for the model, at the time points given by the data set. You may do this either by calculating the sensitivity equations (similar to the previous computer lab session), or by calculating the sensitivities numerically (you can use the provided MATLAB code for calculating the FIM, `MiniFisher.m`).

Use the sensitivity matrix to calculate the simplified form of the FIM, given by $X^T X$, where X is your output sensitivity matrix, and evaluate it at your parameter estimates from 2). What is the rank of the FIM? What does this tell you about the identifiability of your model? Does it match the results from Part 1?

4) **Parameter Uncertainty: Profile Likelihoods**. Now let's examine the structural and practical identifiability of the model parameters and generate confidence intervals using the profile likelihood. Generate profile likelihoods for each of your model parameters (β, γ, and κ). You can play with the range to profile the parameters over, but something like $\pm 25\%$ will likely work well.

For the threshold to use in determining your confidence intervals, we note that $2(NLL(p) - NLL(\hat{p}))$ (where NLL is the negative log likelihood) is approximately χ^2 distributed with degrees of freedom equal to the number of parameters fitted (including the profiled parameter). Then an approximate 95% (for example) confidence interval for p can be made by taking all values of p that lie within the 95th percentile range of the χ^2 distribution for the given degrees of freedom.

In this case, for a 95% confidence interval, we have three total parameters we are estimating (β, γ, and κ), so the χ^2 value for the 95th percentile is 7.8147. Then the confidence interval is any p such that:

$$NLL(p) \leq NLL(\hat{p}) + 7.8147/2$$

In other words, our threshold is $NLL(\hat{p}) + 7.8147/2 = NLL(\hat{p}) + 3.9074$, where $NLL(\hat{p})$ is the cost function value at our parameter estimates from 2).

Plot the threshold on top of your profiles. Are your parameters practically identifiable? What are the 95% confidence intervals for your parameters?

5) **Practical Unidentifiability Issues and Early Epidemic Data**. Lastly, let us consider the case where you are attempting to fit and forecast an ongoing epidemic (i.e. with incomplete data). Truncate your data to only include the first seven data points (i.e. just past the peak), then re-fit the model parameters and generate the profile likelihoods (i.e. re-do 2) and 4) with the truncated data).

- How do your parameter estimates change?
- Does the practical identifiability of the parameters change? How so?
- If any of the parameters were unidentifiable, examine the relationships between parameters that are generated in the profile likelihoods. Can you see any interesting relationships between parameters? What do you think might be going on—why has the identifiability changed?