Identifiability Lab

Marisa Eisenberg (marisae@umich.edu)

July 29, 2019

Notes: This lab is somewhat modular—feel free to choose what problems you find most interesting. You might not finish all the sections—that's okay! Solution code in Matlab, R, and python is available here: https://github.com/epimath/param-estimation-SIR for most of the questions, so you can see how things work out for the 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 the previous lectures:

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

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.

Evaluate the structural identifiability of this model using the differential algebra method, either by hand or with the web app COMBOS (http://biocyb1.cs.ucla.edu/combos/)[1] (or you can use Mathematica/Maple if you have it). If you use 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?
- Reparameterization: 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 be in terms of new variables: $s = S/N, i = I/N$, 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 test the identifiability of β and κ.

[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 and practical identifiability of the 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 (these were denoted s, i, and r in the previous problem). 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:

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

with the measurement equation is $y = \kappa I$, where κ is a product of two things: the population size (N from the previous problem), and the fraction of cases that are reported (k in the previous problem). Thus, $y = \kappa I$ converts the fraction of the population infectious to the observed number of individuals infectious (which is what we measure). 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, we can choose some approximate values from the data by noticing that if $y = \kappa I$, then $I(0) = y(0)/\kappa \approx data(0)/\kappa$, i.e. we can approximate $I(0)$ by the first data point divided by κ (`data(1)/kappa` in MATLAB). Since the data begins early in the epidemic, we can take $R(0) = 0$, and let $S(0) = 1 - I(0)$, since the sum of the fractions of the population in S, I, and R must sum to 1.

2) **Parameter Estimation**. Next, write code to estimate β, γ, and κ using Poisson maximum likelihood and the dataset provided.[2] 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 k, 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.

If you can, it's nice to change the settings in the optimization function so that you can see the progress of the optimization algorithm as it goes. This can be done in MATLAB by adding an `optimset` argument to the `fminsearch` command:

```
fminsearch(@(p)sirCost(times,p,data,x0fcn,yfcn), params, optimset('Display','iter'));
```

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. Just looking by eye, how well would you say the model fits the data?

[2] Note that if you're coding in R, we'll estimate $1/\kappa$ rather than estimating κ itself. This is because κ has a huge potential range, which slows down the optimizer in R, but $1/\kappa$ is usually between 0 and 1.

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 can use the provided code for calculating the FIM (e.g. `MiniFisher.m` in MATLAB).

Use the sensitivity matrix to calculate the simplified form of the FIM, given by $\chi^T\chi$, where χ 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 ±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, with incomplete data. Truncate your data to only include the first seven data points (just past the epidemic peak), then re-fit the model parameters, calculate the FIM, and generate the profile likelihoods (i.e. redo 2 - 4 above). You may need to adjust the percentage range you explore in your profile likelihoods.

- 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 (i.e. plot the profiled parameter vs. the estimated values of the other parameters at each point in the profile—see lecture slides for more info). Can you see any interesting relationships between parameters? What do you think might be going on—what might explain the parameter relationships (i.e. identifiable combinations) that you see?

π) **Extra problems**. 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.

Test the structural identifiability of this model too (e.g. with COMBOS)—you might think it should be the same as Part 1 since our only change was to set $\mu = 0$, but see what happens! How do the results compare to what you would expect? Why?

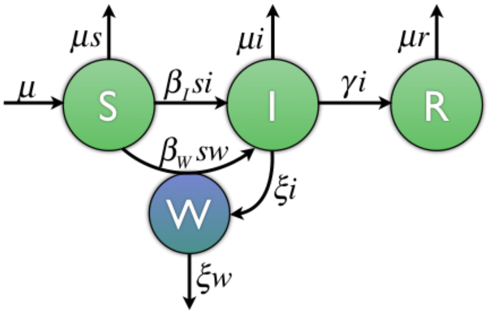

Figure 1: SIWR model of cholera transmission.

Part 3: Modeling Cholera Transmission

Cholera and many waterborne diseases exhibit multiple pathways of infection, which can be modeled (for example) as direct and indirect transmission. A major public health issue for waterborne diseases involves understanding the modes of transmission in order to improve control and prevention strategies (see e.g. Hartley 2006). An important epidemiological question is therefore: given data for an outbreak, can we determine the role and relative importance of direct (human-mediated) vs. environmental/waterborne routes of transmission?

To examine this question, we will use the SIWR model developed by Tien and Earn (2010), shown in Figure 1. We will combine this model with modified data from a recent cholera outbreak. The scaled SIWR model is given by the following equations:

$$\begin{aligned}
\dot{S} &= -\beta_I SI - \beta_W SW \\
\dot{I} &= \beta_I SI + \beta_W SW - \gamma I \\
\dot{W} &= \xi(I - W) \\
\dot{R} &= \gamma I
\end{aligned}$$

where

- S, I, and R are the fractions of the population who are susceptible, infectious, and recovered
- W is a scaled version of the concentration of bacteria in the water
- β_I and β_W are the transmission parameters for direct (human-human) and indirect (environmental) cholera transmission
- ξ is the pathogen decay rate in the water
- γ is the recovery rate

The recovery time for cholera is reasonably well known, so we can fix $\gamma = 0.25$ based on previous work (Tuite 2011, etc.) (i.e. we don't need to estimate this). The SIWR model has previously shown to be structurally identifiable using the differential algebra approach (Eisenberg 2013).

Data & Measurement Equation: Data from a recent outbreak in Angola is given on the course website. To connect the model with the data, we will use the following measurement equation: $y = I/k$, where $1/k$ is a combination of the reporting rate, the asymptomatic rate, and the total population size.

Estimation: For fitting, we'll use ordinary least squares (OLS) for now, i.e. $Cost = \sum_i (data_i - y_i)^2$. However, if you want to, feel free to try another cost function also! They can give quite different answers, both for the parameter estimates and for their uncertainty/practical identifiability, so it can be interesting to see.

1) SIWR Model Simulation. Write code to simulate the SIWR model and plot both the data set provided and the measurement equation $y = I/k$ (i.e. plot both the data and y in one graph vs. time). Use the following parameter values: $\beta_I = \beta_W = 0.75$, $\xi = 0.01$, $k = 1/89193$. For initial conditions, similarly to Part 2, we will use $I(0) = kz(0)$ (i.e. k times the first data value), and take $R(0) = 0$, and let $S(0) = 1 - I(0)$. Lastly, let $W(0) = 0$.

2) Parameter Estimation. Write code to estimate the model parameters β_I, β_W, ξ, and k using the data set provided. The parameters μ and γ will remain fixed (not fit). Use the parameter values from 1) as starting values and the initial conditions from 1) as well.

Plot the cholera data together with your model using the parameter estimates you found. Be sure to plot the data as circles (in MATLAB, use 'o' in the plot function) and your model simulation as a line so that you can compare your model with the data easily.

- How well does the model fit the data? Do you notice any runs or correlated residuals? Are there any potential problems with the model fit?
- Based on your estimated parameters, which transmission pathway would you say is more important/contributes more to this outbreak?

3) Practical Identifiability Issues. Unfortunately, it turns out that the waterborne transmission pathway parameters, β_W and ξ, are often practically unidentifiable when noisy data is considered (Eisenberg 2013). To examine this in a simple way, try simulating your model twice, first with the estimated parameters you found in 2), and then again where you take β_W to be 5/6 the value in 2) and ξ to be 6/5 the value in 2).

Plot both versions of the models together, along with the data. How different are the two fits to the data? What does this tell you about the identifiability of these two parameters? How does that affect the certainty of our estimates of the relative contributions of the two transmission pathways?

4) Profile Likelihoods. Generate profile likelihood plots of your parameters (you may want to adjust the range you profile over). *Note:* because we're using least squares, this will change our threshold value for the profile likelihoods! You can use the lecture slides to recalculate the threshold for the 95% confidence intervals, and we can go through it if you have any questions.

How does this match up with the results of Problems 3? What can you conclude about your model identifiability? If any parameters are unidentifiable, examine the relationships between these parameters and the other parameters, by plotting the profiled parameter vs the estimated values of the other parameters at each point in the profile (see lecture slides for more info). Can you distinguish any potential identifiable combinations?