Identifiability Practical - NC State Parameter Estimation Tutorial

Marisa Eisenberg (marisae@umich.edu)

July 29, 2016

Questions? If you have questions during the practical, please flag down either Chris Durden or I and we'll be happy to help! We will first do Part 2, Problems 1 and 2 together, to get everyone on the same page for the code. Then Part 1 and the rest of Part 2 you can work on at your own pace. You might not finish all sections—that's okay! Feel free to choose what you find most interesting.

Part 1: Structural Identifiability for the SIR Model. Consider the SIR model:

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

with measurement equation $y = kI$. As discussed in yesterday's session, 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. The parameters μ, β, γ, N, and k represent the birth/death rate, transmission parameter, recovery rate, total population size, and proportion of the infected population which is measured/observed. Are the parameters for this model structurally identifiable? (Show how you determined this.) If not, what are the identifiable combinations? What happens if we re-scale the model to be in terms of fractions of the population instead of individuals?

Part 2: 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 **??**. We will combine this model with modified data from a recent cholera outbreak. The scaled SIWR model is given by the following equations:

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

where

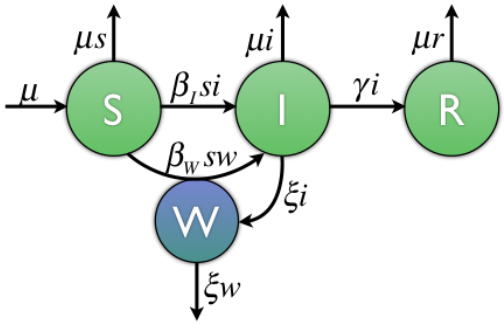

Figure 1: SIWR model of cholera transmission.

- 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
- μ is the birth/death parameter for the population

Since we are considering a short-term outbreak (less than one year), it is reasonable to assume that the effects of births and deaths are negligible, so we set $\mu = 0$. In addition, 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. The Optional Problems section gives a few alternatives for realistic cost functions to try.

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, we can determine them from the data by noticing that if $y = I/k$, then $I(0) = ky(0) \approx kz(0)$, i.e. we can approximate $I(0)$ by the first data point times k (i.e. `data(1)*k` 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. 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.

In addition, change the settings in the optimization function in your main code so that you can see the progress of the optimization algorithm as it goes. This can be done by adding an `optimset` argument to the `fminsearch` command in MATLAB:

```
fminsearch(@(p)siwr_ML(tspan,x0,p,data), params, optimset('Display','iter'));
```

Note: also be sure to set your initial conditions inside the cost function file, since $I(0)$ and $S(0)$ depend on the parameter values (so they will change as you estimate the parameters).

Plot the cholera data together with your model using the parameter estimates you found. Be sure to plot the data as circles ('o' in the plot function) 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? Do you notice any runs or correlated residuals? Are there any potential problems with the model fit? You may want to plot your residuals to see this more clearly.
- 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) 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 yesterday's practical, but now for the SIWR model), or by calculating the sensitivities numerically (see example code provided).

Use the sensitivity matrix to calculate the FIM. What is the rank of the FIM? What does this tell you about the identifiability of your model? Invert your FIM and take a look at the resulting estimate for the covariance matrix and calculate the coefficients of variation (CVs). How certain are your parameters? Would you consider them identifiable?

5) Simulated Data. To explore how noise is affecting the identifiability of your parameters, simulate 20 sets of noisy data assuming a Poisson distribution (or Gaussian, negative binomial, etc.) with your best-fit model trajectory as the mean. Re-estimate the parameters of your model to each of these simulated data sets, and generate scatterplots of your estimated parameter values (do this in pairs, e.g. β_I vs. β_W, β_W vs. ξ, β_I vs. k, etc.). Also plot the true values of your parameters on these same scatterplots in a different color. To do this, Chris has kindly provided a nice function for plotting, `pairsplot_2` (available on the website). How well do the parameter estimates recover the true values? What does this suggest about the effects of noise on model identifiability?

Optional Problems. These are meant to be more open-ended, exploratory problems that you can do later, or if you have extra time. That said, Problems 2π and $2\pi + 1$ are highly recommended to do, as these are useful techniques for a lot of different applications!

2π) Profile Likelihood. Generate profile likelihood plots of your parameters (you can choose the range of values to profile over, but it should include your best-fit parameter values from Problem 2). How does this match up with the results of Problems 2-4? How certain are your parameter values?

$2\pi+1$) Profiled Parameter Relationships. If any parameters appeared practically or structurally unidentifiable in Problem $\pi + 1$, 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?

$2\pi + 2$) Exploring Estimation. Re-run your lab code with an alternative cost function, such as:

- Weighted least squares using Poisson-style noise, $Cost = \sum_i \frac{(data_i - y_i)^2}{data_i}$. This assumes the variance at any given data point is equal to the data.
- Extended/weighted least squares using Poisson noise, $Cost = \sum_i \frac{(data_i - y_i)^2}{y_i}$. This assumes the variance at any given data point i is equal to y_i, the model prediction at that time.
- If you're feeling fancy, you can also try maximum likelihood assuming a Poisson or even negative binomial distribution—you can calculate the cost function for yourself or get it from the slides.

How do the parameter estimates differ from the OLS estimates you did earlier? How does the Fisher Information-based uncertainty differ? Does this change conclusions you might draw about uncertainty/practical identifiability?

$2\pi + 3$) Further Explorations of Practical vs. Structural Identifiability. Try problems 2π and $2\pi + 1$ for both simulated, noise-free data (to test structural identifiability) and for noisy data (either simulated or real). How do your results compare?

$2\pi + 4$) More Fun with Structural Identifiability. Consider the following two compartment model of drug pharmacokinetics:

$$\dot{x}_1 = u(t) + k_{12}x_2 - (k_{21} + k_{01})x_1$$
$$\dot{x}_2 = k_{21}x_1 - (k_{12} + k_{02})x_2$$

where x_1 and x_2 are the masses of drug in the plasma and tissue respectively. The measurement equation is $y = x_1/Vol$, where Vol is the plasma volume. Test the structural identifiability of this model using the differential algebra approach.

Next, let us add an additional parameter to the model, by letting $k_{01}(t) = \frac{V_{max}}{x_1 + K_m}$. How do you think adding this parameter will change the identifiability? Check the identifiability of the new model using the differential algebra approach as well. Is this surprising? What intuition do you have for why this result might be the case?