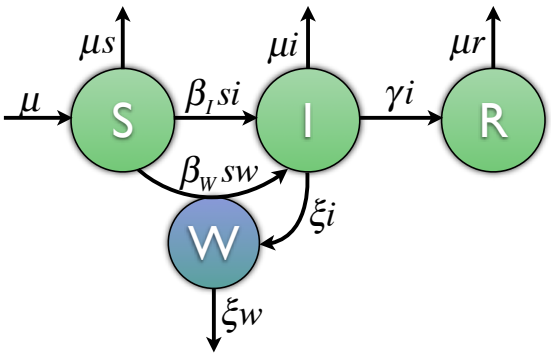
Part 1 – Structural Identifiability for the SIR Model. Consider the SIR model below:

$$\dot{S} = \mu N - \beta SI - \mu S$$
$$\dot{I} = \beta SI - (\mu + \gamma)I$$
$$\dot{R} = \gamma I - \mu R$$
$$y = kI$$

where S, I, and R represent the number of susceptible, infectious, and recovered individuals, and y indicates 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 N is known?

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 (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?

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

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

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 environmental (indirect) 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 (Tuite2011, etc.) (i.e. we don't need to estimate this).

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, use maximum likelihood assuming a Poisson distribution—you can calculate the cost function for yourself or get it from the slides. (Or feel free to use ordinary or weighted least squares if you'd prefer, e.g. assuming the SD of the data is 10% of the data values).

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 $S(0) = 1\text{-I}(0)$, 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 parameter γ will be 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 as follows in MATLAB (see underlined text):

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

Note: 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 an approximate 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) Simulated Data. To explore how noise is affecting the identifiability of your parameters, simulate 20 sets of noisy data assuming a Poisson distribution 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. betaI vs. betaW, betaW vs. xi, gamma vs. k, etc.). Also plot the true values of your parameters on these same scatterplots in a different color. How well do the parameter estimates recover the true values? What does this suggest about the model identifiability?

π + 1) (optional) 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?

π + 2) (optional) Fisher Information Matrix (FIM). Generate the design matrix (output sensitivity matrix) for the model, at the time points given by the data set. Use this to calculate the simplified version of the FIM, given by X^TX, where X is your output sensitivity matrix. 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. How does this compare to your results in part **π + 1)**?

π + 3) (optional) Practical vs. Structural Identifiability. Try parts **4)** and **π + 1)** for both simulated, noise-free data (to test structural identifiability) and for the real data you used for fitting. How do your results compare?