Research Notes: Week 5

Subteam 1

June 15-19

The main focus of this week was to understand, clean, and work with the mouse glucose measurements from the Mathews et al. and Li et al. data sets.

1 Working with the data

The major challenges of processing and cleaning the Mathews and Li data sets were sparsity of data and relativity of time of readings.

1.1 Mathews et al data

The Mathews data set consists of 660 female NOD mice. The authors state that they began taking measurements of their mice starting at 8 weeks old every 2 days. However, they only report measurements from the time that they expect the earliest onset of diabetes to occur up until the onset of diabetes. Therefore their time scale is not only relative to the onset of diabetes, but it varies by mouse. The time scale that they officially use in their analysis starts at -24 days before diabetes onset (day 0).

Out of this cohort of 660 mice, 489 became diabetic over the course of the study. Mathews et al determined diabetes as mice who had two successive readings of 250 mg/dL. However, upon further research, we found that two groups of mice in this cohort were from 2 different labs analyzing blood glucose levels in NOD mice: Parker et al (2009) and Xue et al (2015). These two authors report that they classify diabetes as 2 successive readings of $\geq$ 240 mg/dL. Because Mathews et al does not state whether they introduced their own mice measurement in addition to Parker's and Xue's, we decided to use the latter's definition of diabetes for our current data sets.

These diabetic mice were further categorized (by Mathews only) into 354 "progressive" mice and 135 "acute" mice. Progressive mice are defined as having at least one glucose reading of $\geq$ 200 mg/dL prior to diabetes onset. Acute mice are defined as having no glucose measurements above 200 mg/dL prior to diabetes onset.

1.2 Li et al data

The Li data set consisted of 11 NOD mice, all of whom achieved diabetes. The glucose measurements for these mice were taken irregularly, however, we do know that the time reported is measured absolutely from birth (in weeks). Thus, unlike the Mathews data, we can directly compare mice. Likely because of the smaller size of the cohort, the Li mice were not explicitly categorized into acute and progressive mice. However, when plotting the glucose data, we were able to determine such categories visually. We determined that there were 9 progressive and 2 acute mice (green and flattest blue line we determined to be "acute").

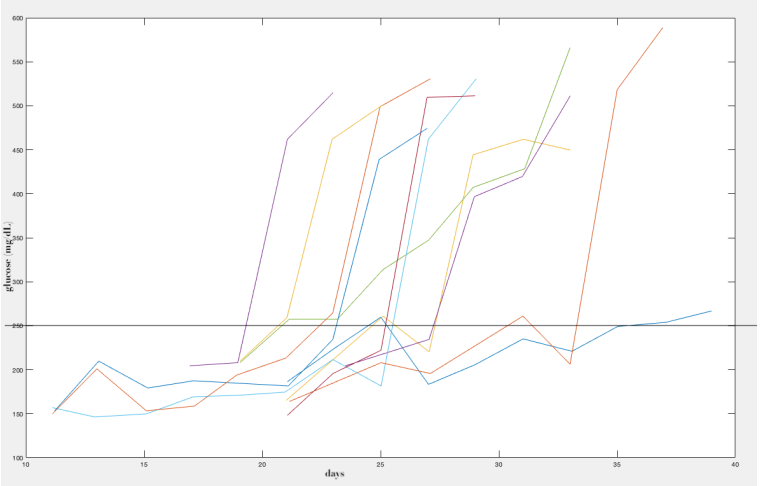


Figure 1: Individual glucose measurements for 11 NOD mice from Li et al.

1.3 Data Alterations

We have two data sets of NOD mice glucose measurements that each have their own strengths and drawbacks. The Mathews cohort is rich in data, however, the time scale is relative and not all mice were measured with the same frequency. The Li et al data has an absolute time scale and no missing data (per mouse), but it is a small data set with only 11 mice.

Goals: Ideally, we would like to utilize the strengths of both data sets while minimizing the drawbacks. We want to work in absolute time and use raw data from both Mathews and Li. Essentially we would like to create one large data set using data from both papers.

Problems: In order to work with all the data, we cannot really get around averaging. But averaging tends to linearize this data and we lose information. In order to be able to average data, we must align each mouse's measurements at a particular time so that an average can be taken uniformly. But what time do we choose? Finally, if we are missing data for mice, do we interpolate? And how will interpolated data skew and affect our final data set?

Solutions: Our most promising solution right now is to merge the two data sets at a specified time of diabetes onset. Mathews gives us plenty of data pre-onset to diabetes while Li gives us more post-onset to diabetes data. We want to align and average the Li et al data set then merge it with averaged Mathews data (already provided to us). In order to merge the two data sets at a reasonable time of diabetes onset, we determine a distribution of onset times from which to sample.

The steps taken to obtain each of our solutions is outlined below:

1.3.1 Averaging Li et al data

We collaborated with Subteam 2 to work on averaging the data. As we said in Section 1.2, we visually determined that there were 9 progressive and 2 acute NOD mice in the cohort. Because the progressive group has more mice, for now we work with this classification of mouse only. The code below outlines the steps we took to determine, interpolate, and average the data.

```
%% Function to align and average Li et al data
function meandat =onset_time_dist
newdat = zeros(2,1); % collect data
for i = 1:11
    % Read in data files
    i_str = int2str(i);
    file = strcat('dat',i_str,'.csv'); %create file name
    X = readtable(file);
    X = X{:,:};
```

```
    X = flip(X);
    X = X'; %set up table as wide

% FIND SECOND CONSECUTIVE READING OVER 240
  numReadings = length(X);
    for j = 1:numReadings
        currentReading = X(j,2);
        if currentReading >= 240
            if j ~= numReadings
                nextReading = X(j+1,2);
                if nextReading >= 240
                    firstTime = j+1;
                    break;
                end
            end
        end
    end

    topReading = X(2,firstTime); %Reading once reach 2nd 240
    bottomReading = X(2,firstTime - 1); %Reading before 2nd 240
    topTime = X(1,firstTime); %Time when over 2nd 240
    bottomTime = X(1,firstTime - 1); %Time before being 2nd 240
    slope = (topReading - bottomReading) / (topTime - bottomTime); %slope between before and after 2nd 1
    delta_t = (250 - bottomReading) / slope; %Change in time to reach 2nd 240
    X_new = zeros(2, numReadings - firstTime + 2); %Create new X vector
    X_new(:, 2:end) = X(:, firstTime:end); %Shift everything over a spot
    X_new(:, 1) = [bottomTime + delta_t; 250]; %Add initial time point

    % Create interpolated data sets with discrete times
    %INTERPOLATE VALUES TO FILL IN X_NEW
    min_time = round(X_new(1,1)); %minimum time available
    max_time = round(X_new(1,end)); %maximum time available
    vq = interp1(X_new(1,:), X_new(2,:), [min_time:1:max_time], 'pchip', 'extrap');
    %interpolate

    % Exclude acute mice
    if i ~= 1 || i ~=5
        newdat = horzcat(newdat, [min_time:1:max_time;vq]);
    end
end

% Reformat collected data
newdat = sort(newdat');
newdat = newdat(2:end, :);

% Take mean of data
meandat = zeros(2,1);
for j = newdat(1,1):newdat(end,1)
indices = find(newdat(:,1) == j);
result = mean(newdat(indices, 2));
meandat = horzcat(meandat, [j; result]);
end
meandat = meandat(:, 8:end)
end
```

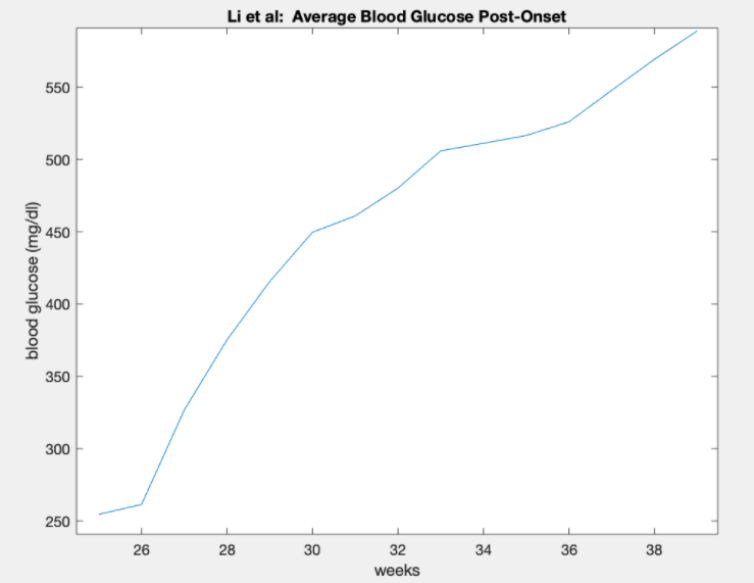


Figure 2: Shifted and averaged glucose data post diabetes onset from Li et al.

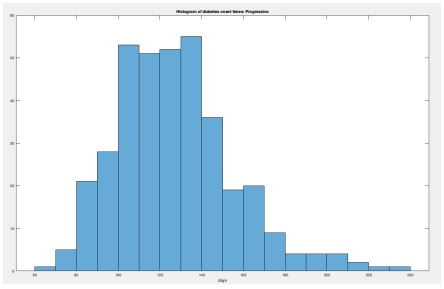
1.3.2 Finding diabetes onset distributions

We found onset distributions for both Li and Mathews data, but since there is more data for Mathews, we believe these distributions to be more reliable. To do this we used code from Professor Shtylla that determines an absolute time span for the raw glucose measurements of each mouse by sampling from a lognormal distribution of time (details of this code can be found *here*). From this time calibrated data set, we determined and collected diabetes onset times (using the threshold definition from 1.1). Our last step was to fit a distribution to these thresholds. We determined a lognormal fit visually. Our code below explicitly runs through these steps to produce diabetes onset time distributions for acute and progressive mice.

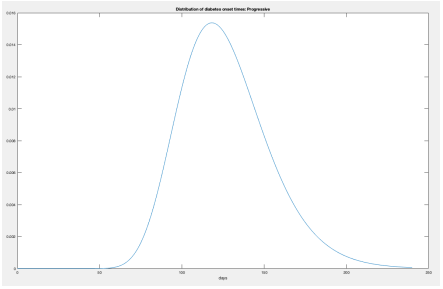
```
%% Function to find onset of diabetes distributions for Mathews et al data
function [onset_distX, onset_distY] = onset_time_dist_mathews

% Load shifted data from Prof. Shtylla
generate_onset_times;

% Define shifted data
X = shifted_progressive;
Y = shifted_acute;

% Determine columns with all NaN (no data)
colstodeleteX = removeCols(X);
colstodeleteY = removeCols(Y);

thresholdX = zeros(1,1); % Initialize onset time vectors
thresholdY = zeros(1,1);

% LOOP FOR PROGRESSIVE ONSET DISTR.
for n = 2:2:length(X(1,:)) % For each mouse...

 % Skip column if all NaN
if ~ismember(n, colstodeleteX)
```

```
  % Find second consecutive reading over 240
  numReadings = 25;
    for j = 1:numReadings
        currentReading = X(j,n);
        % Check next reading if applicable
        if currentReading >= 240 && j ~= 25
            nextReading = X(j+1,n);
            if nextReading >= 240
               firstTime = j+1;
               break;
            end
        end
    end
end

    thresholdX = [thresholdX X(firstTime, n-1)]; % collect onset times
end

% LOOP FOR ACUTE ONSET DISTR.
for n = 2:2:length(Y(1,:)) % For each mouse...

 % Skip column if all NaN
if ~ismember(n, colstodeleteY)

  % Find second consecutive reading over 240
  numReadings = 25;
    for j = 1:numReadings
        currentReading = Y(j,n);
        % Check next reading if applicable
        if currentReading >= 240 && j ~= 25
            nextReading = Y(j+1,n);
            if nextReading >= 240
               firstTime = j+1;
               break;
            end
        end
    end
end

    thresholdY = [thresholdY Y(firstTime, n-1)]; % collect onset times
end

% Remove 0 inserted for initialization
thresholdX(1) = [];
thresholdY(1) = [];

% FIT A DISTRIBUTION TO ONSET TIMES COLLECTED
onset_distX = fitdist(thresholdX', 'lognormal');
onset_distY = fitdist(thresholdY', 'lognormal');

end
```

Figures 3-5 illustrate the histograms used to determine distribution fits as well as the actual distributions obtained by this function.

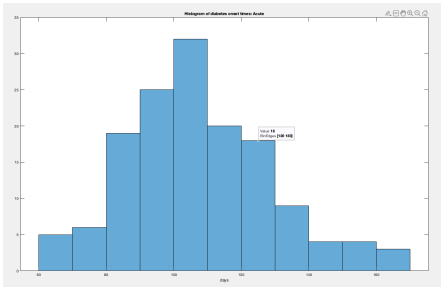


(a) Histogram of diabetes onset times

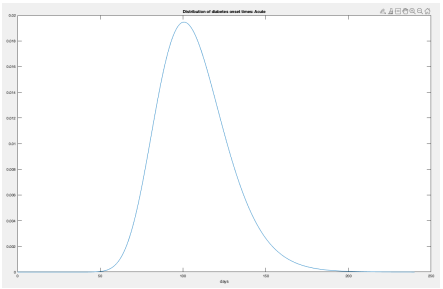


(b) Fitted lognormal distributions for diabetes onset time.

Figure 3: Finding diabetes onset time distributions for progressive mice, Mathew et al.

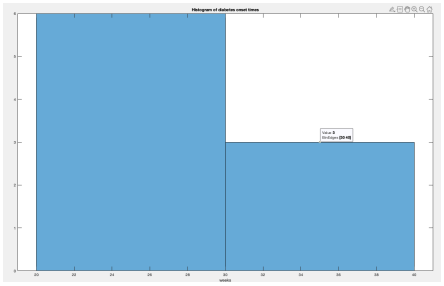


(a) Histogram of diabetes onset times

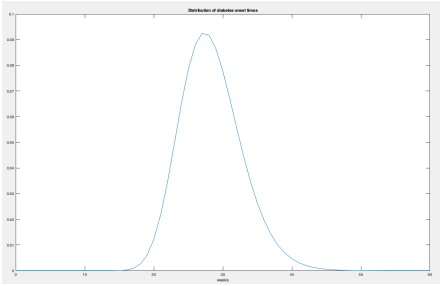


(b) Fitted lognormal distributions for diabetes onset time.

Figure 4: Finding diabetes onset time distributions for acute mice, Mathew et al.



(a) Histogram of diabetes onset times



(b) Fitted lognormal distributions for diabetes onset time.

Figure 5: Finding diabetes onset time distributions for Li et al mice.

Because we fit a distribution to the diabetes onset times, we also obtain mean and standard deviation values (and confidence intervals), Figure 6. We use these to randomly select an onset time around which to construct a disease progression time span for our ODE model.

```
Lognormal distribution
     mu = 4.80969    [4.78775, 4.83162]
  sigma =  0.2134    [0.198979, 0.230092]
```

(a) Mathews et al progressive

```
Lognormal distribution
     mu =  4.65979    [4.62775, 4.69183]
  sigma = 0.195195    [0.175019, 0.22067]
```

(b) Mathews et al acute

```
Lognormal distribution
     mu =  3.33283    [3.21342, 3.45223]
  sigma = 0.155336    [0.104923, 0.297587]
```

(c) Li et al

Figure 6: Mean, standard deviation, and confidence intervals for diabetes onset distributions.

With these onset distributions, we can now determine an absolute time frame for the Mathews et al data.

1.4 Subject to change

We acknowledge that there are inconsistencies within and between each of the data sets. Due to the irregularity of both data sets, we have had to make many assumptions while attempting to consolidate and merge the data sets.

The Mathews data was an accumulation of data from mice from 2009-2014 from a variety of sources. We use Parker et al's and Xue et al's diabetes onset criterion, which differs from Mathews et al's criterion, however, we may in the future need to determine our own definition. Both criteria are somewhat arbitrary.

In order to maintain consistency across the two data sets, we visually classified Li et al mice into categories of progressive and acute mice. However, Mathews has a specific definition of each of these type of mouse that we did not use to determine Li et al mice. We were looking for mice whose glucose readings followed a similar shape (z shaped) so that we could align and average them to create a single data set. Our classification of Li et al mice is not currently consistent with the Mathews data, so we will need to revist and revise these categories.

2 MCMC alterations

We now have all the necessary tools to use both the Mathews and Li data in the DRAM MCMC algorithm. We have:

- Averaged Li et al data for progressive mice
- Averaged Mathews et al data for progressive mice (given)
- Diabetes onset times distributions

To incorporate these new items, we made some alterations to our current algorithm:

1. New parameters
    - Glucose and eFAST sensitive parameters
    - Mean and standard deviation of onset time
        - These become parameters because since we are now able to work with Mathew data in absolute time since birth, we do not want to constrict our time choice. By giving the way we determine a time span of disease progression, we can allow the times model is solved (in the ODE solver) to fit the averaged mouse data more accurately (than a single constant time)
    - Initial conditions
        - Recall that the initial conditions remain constant for the ODE solver were given to us. However, since we do not know for what kind of time frame these initial conditions were initialized for, parameterizing them will hopefully give us a better parameterization of our parameters of interest.

2. Sum of squares (likelihood) function

- In order to allow our time to vary as a parameter, we had to alter our sum of squares function. Previously, our sum of squares took in a set of glucose data and parameters, solved the ODE with those parameters at a constant set time span, and then took the sum of the differences (between model and original data) squared.
- Now, we want to allow our ODE to be solved at a variety of different time spans. To do this, we:
  (a) Solve the ODE for a "large" time range, i.e. enough time for our modeled mouse to go through the full disease progression: slow increase to spiked onset of diabetes
  (b) Using our onset time distribution mean and standard deviation, we randomly select an onset time and construct a time span of x data points (however many original data points we have).
  (c) With this specific time span, we lift the glucose data at those times from the modeled data
  (d) We can then carry on with taking the sum of the square of the differences between modeled and original data to complete the sum of squares function

  The code below executes this process.

```
function ss = T1Dss(params, data)
% Sum-of-squares function

% EXTRACT GLUCOSE DATA
ydata=data(:,2);

% SOLVE ODE MODEL
% Ynowave: no wave wild type (1)
% Ynnowave: NOD no wave (2)
% Ywave: wild type + wave (3)
% Ynwave: NOD + wave (4)
[tmodel ymodel]= T1Dfun_pred(2,data,params); % solved for 40 week time span

% DETERMINE ONSET TIME
mu = params(end-1); % from our onset distribution
sigma = params(end); % from our onset distribution
onset_time=round(exp(mu+sigma^2/2)); % convert to normal from lognormal
onset = ceil(onset_time/7); % obtain integer time points (weeks)

% SELECT GLUCOSE DATA
tmodel=tmodel(1:onset+(length(ydata)-onset));
ymodel=ymodel(1:onset+(length(ydata)-onset));

% SUM OF SQUARES
ss = sum((ymodel-ydata).^2);
end
```

The following figure illustrates the impact of using onset distributions to determine time span of the model. In (a), we can see the original data plotted at a preset, specific time frame. The most important thing to note is the general shape of the data: we see the gradual increase and then the spike of glucose at around 250 mg/dL indicating diabetes onset. In (b), we see 3 different lines. The blue line indicates the ODE solved data using the onset distributions to determine a time span. The yellow line is the same data as in (a). We can see that in the original data's specified time frame, our model does not do a good job of modelling the data. However, the red line indicates the same original data now set at the time frame determined by the onset distributions, here we can see that the original and modeled data are much closer in shape. The original data is at its true shape as can be seen in (a). It may seem odd to adjust the original data to better

match the modelled data, but remember that for the Mathews data, the time is relative (-24 days to onset of disease), so we are merely attempting to fit the original data to an absolute time frame of days since birth.

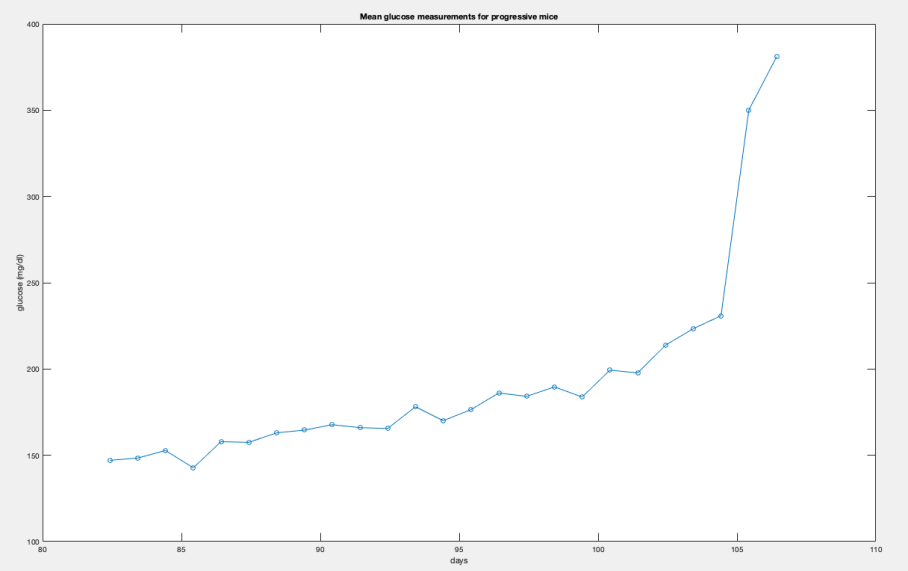


Figure 7: Original data and time span, Mathews et al.

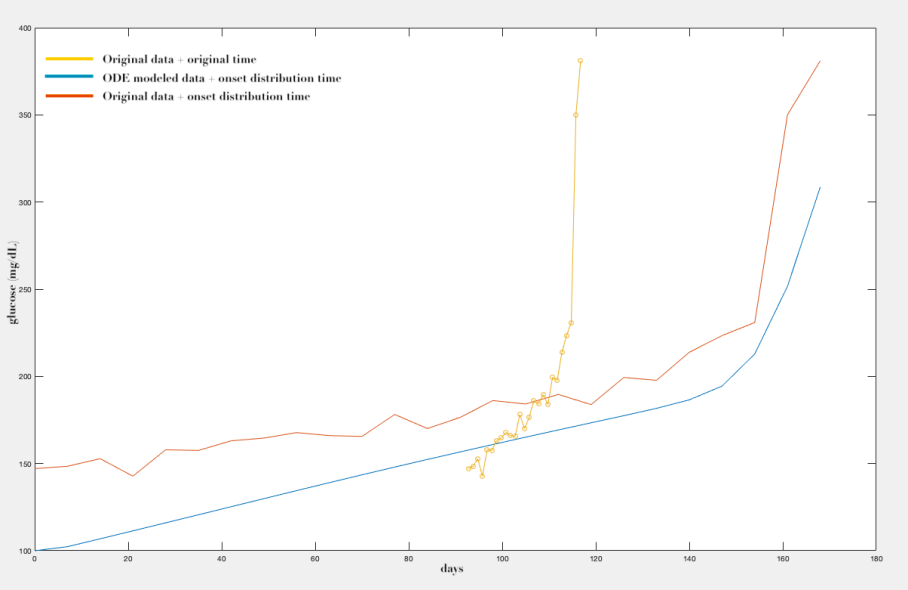


Figure 8: Comparison of model with original data with different time spans, Mathews et al.

2.1 Problems and Improvements

Unfortunately, our MCMC algorithm does not currently work under these newly described conditions. We know that when the algorithm is sampling parameter sets, it is rejecting everything, thus we have no actual MCMC chain. It is difficult to determine exactly where the issue is occurring because the majority of the work of the algorithm occurs in a function that we did not write (mcmcrun from Marko Laine's mcmcstat). However, by doing a little stepping through, it seems that our parameters are being sampled outside of their upper and lower bounds (that we set manually in our parameter structure) and thus the chain cannot

consider them as possible candidates. The cause of this could be from introducing the time and initial condition parameters, as these are parameters whose upper and lower bounds we are less confident about.

For now the goal is to continue debugging our code to get a functioning MCMC algorithm that will consider our new parameters.

3 PSO alterations

As a metric for comparison to DRAM MCMC, the Particle Swarm Optimization (PSO) routine already in place for the T1D system was updated to reflect the changes in data handling. Alterations in the algorithm were made in regard to:

1. New parameter subsets
2. Newly compiled dataset

3.1 New parameter subsets

eFAST sensitivity analysis by A. Do identified parameters EG_0, α_B, μ_e, μ_r, D_{ss}, and Q_{panc} to be the most sensitive parameters in the model. Simply put, a parameter's *sensitivity* refers to its relative influence over the model outcome. Sensitivity analysis asks "if the value of this parameter were changed slightly, how much would the model behavior be impacted?". For the aforementioned parameters, the model behavior would be impacted drastically by minimal variation in value. Given this consideration, it is of great importance to determine accurate values for these sensitive parameters to ensure the predictive validity of the model.

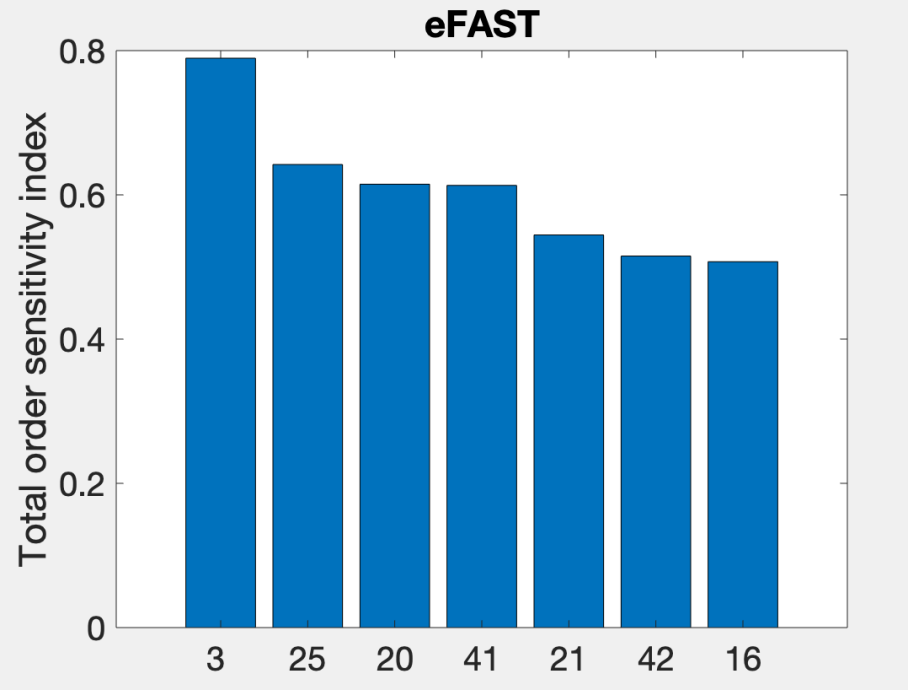


Figure 9: Results of eFAST Sensitivity Analysis by A. Do

To tune these values, the PSO algorithm was edited to fix all model parameters except for those of the parameters identified above. These were then optimized to fit the Li et al. averaged data centered at the mean of its onset-time distribution. Unfortunately, upon inspection, the fitted model did not perform well; neither in its shape nor in its values. This leads us to believe that our interpretation of the eFAST results is likely incomplete, and parameter dependencies have not been appropriately considered.

Additionally, a second subset of parameters was defined. Because the only observable available is glucose, it seemed reasonable to fit just parameters directly involved in the calculation of glucose within the ODE system. The parameters were SI, R_0, and EG_0. There is notable overlap between this subset and the sensitive subset. When run on the same data described above, the performance was poor. The fitting produced a higher sum-of-squares result than the sensitive parameter subset, suggesting a worse fit.

The two subsets were also run together. This produced the best outcome of the three tested subsets, however all were inferior to allowing every parameter in the system to vary ± 10 percent. This makes sense: allowing for variation in more parameters allows for more massaging of the fit, leading to a better fit.

3.2 New dataset

Like in MCMC, the dataset was changed, as was the relationship of the ODE with the dataset. This took form in changes to the sum-of-squares objective function. The objective function was altered to allow the onset time to vary according to the distribution extracted from Mathews et al. This distibution is shaped by many more datapoints, and hence was considered a more reliable choice than the distribution produced from the Li et al. data. The hyperparameters of the log-normal distribution were treated as parameters of the T1D system and were allowed to vary within their 95 percent confidence intervals. With each evaluation of the objective function, an onset time would be sampled from the proposed distribution. The model output at this time would then be selected for comparison against the observed data points. This allowed a sort of sliding window for the selection of model predictions, and a fit that conserved more information about individual mouse glucose than the previous approach of taking a simple average.

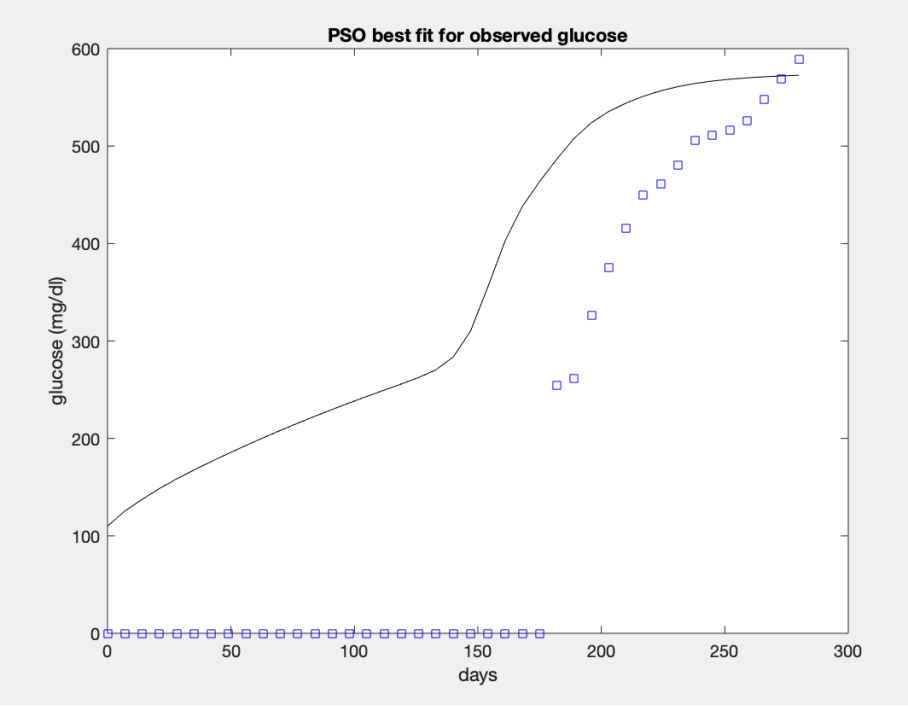


Figure 10: Fitting Li et al. average with PSO

By implementing this sliding window mechanism, the fit to the data was significantly improved. Before implementation, the sum of squares objective function would return values $\sim 6,000 - 11,000$ no matter the parameter subset being optimized. After implementation, the sum-of-squares values were reduced to $\sim 1,000 - 7,000$. The lower sum-of-squares values show that when the onset time is treated as variable, the fit of the model to the data becomes better. Despite this, however, the visualized fit is notably 'off'. In Figure 10, the shape of the optimized function resembles that of the data closely, but is shifted along the x-axis. Upon further inspection, the PSO algorithm consistently drove the mean of the onset time distribution to its

maximum allowable value. This suggests that the true value of the mean that optimizes the fit is beyond the allowable range. Essentially, the fit pictured in Figure 10 is shifted as far right as allowed, but would naturally tend to move even further in the same direction. This is attributed to the differences in the Mathews et al. onset time distribution and the Li et al. onset time distribution. To fix this, we propose different strategies of combining the datasets, specifically shifting the Li et al. data according to the Mathews distibution, or using the mean value from Li and the standard deviation from Mathews to maintain shape, but change the values within the distribution. Additionally, we concede that the fit of the pre-onset part of the solution is unknown, as no data from that time period was 'observed'. We propose combining the pre-onset data from Mathews with the data considered above (and smoothing the meeting point if necessary) to better inform the fit of the model before diabetes onset.

4 Improvements

Immediately, we seek to debug the DRAM MCMC, figuring out why all candidate parameters are being rejected and fixing this issue. Once we have a working MCMC routine, we intend on implementing several improvements (most of which have been discussed previously in this report):

1. Reclassify Li et al. mice according to Mathews's definitions of 'Progressive'/'Acute'; recalculate Li averaged data based on this
2. Fuse the two data sets at the time of onset for a full depiction of diabetes
3. Examine our use of interpolation, determine its necessity, and possibly find a way to use the original times associated with the data.

5 Moving forward 'Big Picture'

Using MCMC for parameter estimation is a powerful technique overall, but the quality of its fits relies on the information known about the system and the quality of data. Because of the level of manipulation involved in creating our data, we think a long-term goal could involve the use of Kalman filtering to determine informative priors for the parameters of the system. This would involve compiling distributions of best-fit parameters for a series of individual mice with UKFs, and using these priors to fit the population as a whole with DRAM MCMC. The use of informative vs. uniform priors is known to result in better fits, and we think that this intersection between the work of the two subteams could be very interesting.