Journal of Ryan Botts: Fall 2008

Ryan Botts

November 12, 2008

1 8 Sept - 16 Sept: Investigating the Dynamics of the System

I am still trying to understand why we can recover the network with fewer data points when using time series data than when we use evenly sampled data (using the known function values). I have looked into finding a translation of the data which would translate the time series data to evenly spaced points. I have not had any luck doing this, so I am going to begin looking into finding orthogonal basis using weighted measures. I am also thinking of testing the fitting on only some of the real time series data to check how spread out we need the data to be. I also want to look into setting up the least squares equations to run each direction simultaneously.

I also modified the code so that we could account for the self-degradation term and did not incorrectly identify this as self-regulation. This seems to make the least squares equations form poorly conditioned matrices, so I might need to look into why this would be.

2 17 Sept - 23 Sept: Paper and Errors

This past week was mostly spent getting the proposal in a little better shape. I have the outline for it and all but 3 sections written, I do however need to edit what is there. I hate writing. I have been keeping some notes as I go, but they are not nearly enough. The moral here is that as you go, each week you should write anything you have learned in a manner that could be used for a paper later on.

During this next week I want to run tests on the data using the known functions with added error. By determining at what magnitude of error we can still recover the network we should be able to figure out maximum allowable time steps used in generating the time series. I also want to generate a collection of time series data where the time step varies depending on the magnitude of the derivative, so that we have fewer points near stable concentrations and more points near unstable concentrations.

3 24 Sept - 30 Sept: A New Method for Identifying Interactions

The beginning of this week I spend working on itemizing the tasks I have in my project. Before writing all of my proposal we want to identify exactly what I see as the goals of this project to make sure that everyone is on the same page. So I spent some time working on this part of the paper.

I spent most of the time this last week implementing a new procedure for identifying interactions. We are considering using:

$$h(x_i) = \int_{[0,1]} \int_{[0,1]} \dots \int_{[0,1]} g_j^2 dx_1 \dots dx_{i-1} dx_{i+1} \dots dx_n, \tag{1}$$

to identify if component i is involved in the regulation of component j. This integral becomes a function of only variable x_i, which allows us to combine all of the summands in the separated representation and avoid many of the complications we had earlier in determining if component i was involved in regulating a specific component. It is still unclear as to what the best way of identifying an interaction from this function is. During this upcoming week I should plot g_j and $h(x_i)$ and see how these terms combine and if we can make any conclusions about promotion or inhibition.

Another method, which might be useful for our purposes is to consider:

$$\begin{aligned} h(x_i) &= \sum_{l=1}^{r} f_i^l(x_i) \frac{1}{N} \sum_{Data} \prod_{j \neq i} f_j^l(x_j) && (2) \\ &\approx \sum_{l=1}^{r} f_i^l(x_i) \prod_{j \neq i} f_{jave}^l(x_j), && (3) \end{aligned}$$

when using rectangles to approximate each of the integrals used in computing the average function values. This method might lend itself to some sort of error analysis using some of the basic integral approximation error formulas. I have outlined what I need to do to implement this technique and need to do it now.

4 1 Oct - 14 Oct: Writing and a Few Test Results

During this time period I have focused on writing. Martin and I met at the beginning and reorganized many of the things that I had written. I have completed these modifications and feel that the proposal is finally taking shape. I only have to finish the last section on what I am currently doing and current results. Some of the new results are not that good, so I am not sure if they should be included.

Using the technique developed last week for identifying interactions and running 100 fits on 243 evenly spaced data points and using rank 2 separable functions with degree 2 polynomial factors resulted in the following interaction matrix:

$$R_1 = \begin{pmatrix} 6.33e+00 & 8.08e-04 & 1.53e-01 & 5.41e-07 & 7.74e-04 \\ 2.75e-02 & 3.35e+00 & 1.57e-01 & 4.79e-07 & 6.90e-04 \\ 2.76e-03 & 8.25e-03 & 1.03e+00 & 8.43e-07 & 8.04e-04 \\ 8.22e-03 & 2.26e-03 & 1.00e-01 & 1.05e+00 & 2.95e-04 \\ 2.12e-01 & 4.38e-04 & 3.38e-02 & 1.42e-04 & 1.12e+00 \end{pmatrix}. \tag{4}$$

Using the same technique except including the self-degradation term in the analysis and running 100 fits on 243 evenly spaced data points and using rank 2 separable functions with degree 3 polynomial factors resulted in the following interaction matrix (Should rerun this test):

$$R_2 = \begin{pmatrix} 2.85e+14 & 2.16e+12 & 1.54e+14 & 6.93e+14 & 4.43e+16 \\ 9.60e+16 & 1.45e+16 & 3.88e+14 & 1.49e+15 & 9.61e+16 \\ 3.42e+14 & 1.84e+13 & 1.04e+15 & 1.11e+15 & 7.20e+16 \\ 1.18e+15 & 9.55e+13 & 7.36e+14 & 1.17e+16 & 6.23e+16 \\ 2.95e+14 & 3.30e+12 & 6.17e+12 & 9.49e+14 & 6.25e+16 \end{pmatrix}. \tag{5}$$

Using the time series data with 186 points and the same two respective techniques results in

$$R_1 = \begin{pmatrix} 1.50e+00 & 2.21e+00 & 5.37e+00 & 8.24e-01 & 4.93e-01 \\ 2.25e+00 & 2.16e+00 & 6.99e+00 & 8.09e-01 & 5.13e-01 \\ 1.41e-03 & 3.30e-03 & 1.00e+00 & 3.98e-04 & 2.40e-04 \\ 4.77e-01 & 5.99e-01 & 1.55e+00 & 7.73e-01 & 1.64e-01 \\ 1.03e+00 & 2.28e-01 & 4.04e+00 & 1.63e+00 & 3.65e+00 \end{pmatrix}. \tag{6}$$

$$R_2 = \begin{pmatrix} 4.01e+01 & 1.45e+02 & 9.77e+01 & 1.92e+02 & 7.94e+01 \\ 1.51e+03 & 1.28e+03 & 1.32e+03 & 5.41e+02 & 9.87e+02 \\ 7.76e-04 & 9.59e-04 & 1.00e+00 & 8.37e-04 & 9.34e-04 \\ 1.13e+02 & 4.84e+02 & 1.26e+02 & 5.71e+02 & 3.21e+02 \\ 7.18e+01 & 2.73e+01 & 3.60e+01 & 7.21e+01 & 2.27e+01 \end{pmatrix}. \tag{7}$$

Recall that the true interaction matrix is

$$R = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \end{pmatrix} \tag{8}$$

Using the evenly spaced data we see that R_1 tends to have larger values where there should be interactions, however it does not recover all of them. When the self degradation term is included, then the approximations tend to be much worse.

5 15 Oct. -21 Oct.: Writing and modifying the techniques for identifying interactions

This week I am finishing up writing. Another version is done. I also began working on a presentation of this project for next Tuesday. In addition I ran new tests ignoring the self-degradation and using rank 2 degree 2 sums of separable functions. Using the notation R_i^e to indicate the mean interaction matrix on the evenly spaced data using the ith method for identifying interactions we obtained

$$R_1^e = \begin{pmatrix} 1.20\times10^{-1} & 5.27\times10^{-4} & 1.34\times10^{-1} & 5.02\times10^{-4} & 2.84\times10^{-3} \\ 3.56\times10^{-3} & 4.41\times10^{-2} & 1.37\times10^{-1} & 5.02\times10^{-4} & 2.84\times10^{-3} \\ 3.10\times10^{-3} & 7.46\times10^{-3} & 1.50\times10^{-1} & 5.02\times10^{-4} & 1.98\times10^{-3} \\ 6.45\times10^{-3} & 1.94\times10^{-3} & 2.57\times10^{-1} & 2.23\times10^{-2} & 2.80\times10^{-6} \\ 5.61\times10^{-2} & 6.89\times10^{-5} & 5.60\times10^{-2} & 9.59\times10^{-6} & 6.03\times10^{-2} \end{pmatrix}, \tag{9}$$

$$R_2^e = \begin{pmatrix} 1.18\times10^{+0} & 1.49\times10^{-4} & 1.09\times10^{+0} & 1.93\times10^{-7} & 1.63\times10^{-4} \\ 1.84\times10^{-3} & 2.37\times10^{+0} & 1.05\times10^{+0} & 1.95\times10^{-7} & 1.61\times10^{-4} \\ 4.77\times10^{-5} & 1.93\times10^{-3} & 8.67\times10^{-1} & 8.69\times10^{-7} & 2.45\times10^{-5} \\ 7.80\times10^{-5} & 5.53\times10^{-4} & 3.24\times10^{-1} & 1.06\times10^{+0} & 3.20\times10^{-6} \\ 3.17\times10^{-3} & 5.79\times10^{-5} & 2.52\times10^{-1} & 9.66\times10^{-6} & 1.11\times10^{+0} \end{pmatrix}, \tag{10}$$

and

$$R_3^e = \begin{pmatrix} 1.24\times10^{-1} & 1.42\times10^{-4} & 3.16\times10^{-2} & 1.48\times10^{-4} & 3.92\times10^{-4} \\ 7.00\times10^{-4} & 1.43\times10^{-1} & 4.12\times10^{-2} & 1.46\times10^{-4} & 3.86\times10^{-4} \\ 3.01\times10^{-3} & 8.60\times10^{-3} & 8.74\times10^{-1} & 7.55\times10^{-5} & 1.40\times10^{-3} \\ 3.90\times10^{-3} & 6.55\times10^{-4} & 1.94\times10^{-1} & 4.65\times10^{-1} & 2.03\times10^{-6} \\ 3.98\times10^{-2} & 5.77\times10^{-5} & 3.97\times10^{-2} & 4.84\times10^{-6} & 6.17\times10^{-1} \end{pmatrix}. \tag{11}$$

As mentioned before the larger the value of $r_{i,j}$ the larger the interaction. The largest values of $r_{i,j}$ in row i correspond to the strongest interactions with component i. The first technique identifies very few correct interactions, this may be attributed to the fact that this is a rank 2 model and this technique cannot account for interactions between terms in the sum. The second technique correctly identifies all of the terms along the diagonal and all of the terms in the third column, but not many others. The third technique identifies the interactions along the diagonal, the interactions in the third column and a few others. We are still working on the results for the fourth technique. We would expect the second and third techniques to achieve nearly the same results as the third is a discrete approximation of the second. In the case of evenly spaced data we would expect this to be a good approximation and the results to be similar, however when using on time series data, we would expect the third method to do much better.

We next sampled the regulation functions using the time series data generated by our system. We obtained this data using the initial concentrations of 1 and then knockout data where each component is systematically removed and the concentrations set to 0, i.e. using the starting concentrations of $[1,1,1,1,1]$, $[0,1,1,1,1]$, $[1,0,1,1,1]$, and so on. We started from each of these six states and ran our time series over 10 time steps of size 0.1, generated using a Runge-Kutta fourth order approximation, for a total of 66 data points. Again using a rank 3 sum of separable cubic polynomial factors, and 100 fits we found the following mean interaction matrices

$$R_1^t = \begin{pmatrix} 1.06\times10^{+0} & 1.40\times10^{+0} & 1.03\times10^{+0} & 1.08\times10^{+0} & 6.75\times10^{-1} \\ 9.05\times10^{-1} & 9.45\times10^{-1} & 1.04\times10^{+0} & 1.15\times10^{+0} & 6.09\times10^{-1} \\ 6.89\times10^{-1} & 9.54\times10^{-1} & 1.26\times10^{+0} & 1.16\times10^{+0} & 6.62\times10^{-1} \\ 8.86\times10^{-1} & 8.85\times10^{-1} & 1.15\times10^{+0} & 1.19\times10^{+0} & 7.28\times10^{-1} \\ 8.51\times10^{-1} & 1.52\times10^{+0} & 1.08\times10^{+0} & 1.17\times10^{+0} & 9.75\times10^{-1} \end{pmatrix}, \tag{12}$$

$$R_2^t = \begin{pmatrix} 1.09\times10^{+4} & 1.82\times10^{+4} & 2.10\times10^{+4} & 1.61\times10^{+4} & 1.27\times10^{+4} \\ 5.86\times10^{+4} & 5.90\times10^{+4} & 7.55\times10^{+4} & 2.53\times10^{+5} & 5.48\times10^{+4} \\ 4.31\times10^{+0} & 1.16\times10^{+1} & 6.22\times10^{+0} & 1.67\times10^{+1} & 6.31\times10^{+0} \\ 1.92\times10^{+4} & 5.04\times10^{+4} & 2.75\times10^{+4} & 6.29\times10^{+4} & 1.57\times10^{+4} \\ 9.24\times10^{+4} & 6.34\times10^{+4} & 2.90\times10^{+4} & 2.15\times10^{+5} & 4.57\times10^{+4} \end{pmatrix} \tag{13}$$

and

$$R_3^t = \begin{pmatrix} 8.34\times10^{-1} & 8.54\times10^{-2} & 1.20\times10^{-1} & 4.75\times10^{-2} & 2.31\times10^{-2} \\ 8.62\times10^{-2} & 8.21\times10^{-1} & 2.49\times10^{-1} & 1.40\times10^{-1} & 4.05\times10^{-2} \\ 7.33\times10^{-3} & 1.28\times10^{-2} & 1.17\times10^{+0} & 1.74\times10^{-2} & 9.14\times10^{-3} \\ 7.23\times10^{-2} & 9.22\times10^{-2} & 2.62\times10^{-1} & 8.96\times10^{-1} & 5.23\times10^{-2} \\ 2.79\times10^{-2} & 1.10\times10^{-1} & 1.62\times10^{-1} & 7.96\times10^{-2} & 9.31\times10^{-1} \end{pmatrix}. \tag{14}$$

The mean MSE over all of these fits was 3.14×10^{-2}, which is curiously larger than the error obtained when using this same model on nearly 20 times as many evenly spaced data points. The data was not fit very well, which might account for why these results identified fewer correct connections.

Here we see that the first technique does not identify any interactions correctly, this technique does depend on values of the function which are not supported by the data, which may explain this. The second technique does not clearly identify many interactions, which, again may be due to the fact that the technique relies on integrals over regions which are not supported by the data, or might be that there are simply too few data points. The third technique identifies all interactions along the main diagonal, the interactions in the third column and a few other interactions. The results in this case are nearly as good as with the evenly space data, but with only 66 data points as opposed to 1024. It is possible that the fourth technique may perform better than this as it is entirely dependent on the data.

I also tested using a rank 3 with cubic factors. This did not yield any useful results. I also began working on a fourth technique for identifying interactions. This would use

$$r_{i,j} = \max_{x_j^k k=1,\dots n} \left(F(x_j^k)\right) - \min_{x_j^k k=1,\dots n} \left(F(x_j^k)\right), \tag{15}$$

where

$$F_{i,j}(x) = \frac{1}{N}\sum_{k=1}^{n}\left(\sum_{l=1}^{r} s_l f_j^l(x)\prod_{i\neq j} f_i^l(x_i^k)\right)^2 \tag{16}$$

The advantage to this technique would be that it relies entirely on values of our regression function at known data points. This removes the possibility that in taking that maximum or minimum values we are use values which are not supported by the data.

6 22 Oct. -28 Oct.: Presentation

This week I prepared to give a presentation of my proposal and finished up some of the last changes on the paper. I finished analyzing the results using the techniques we discussed. During this next week I am going to improve the code for the discrete method of identifying interactions.

7 29 Oct. -4 Nov.: Paper and Writing Efficient Code

This week I finally finished a draft of the proposal that is ready to give to possible committee members. I should be getting input back on these during this next week.

We hypothesize will perform best on the time series data as it is only dependent upon values of the regression function where we have data support, however we have no results prior to this as the original code for identifying these interactions was too inefficient. This week we developed a way to speed this process up and implemented it in the code. We compute all of the values of the regression function over all of the data points, and store them so that we do not need to compute them every time. We may also take a mean over the data points before adding the functions together. We identified that the maximum and minimum of the original function occur near 0 and 1 so we added those to our evenly spaced data. With these changes we obtained the following results.

$$R_1^e = \begin{pmatrix} \underline{1.06 \times 10^{0}} & 3.66 \times 10^{-15} & \underline{2.21 \times 10^{0}} & 1.09 \times 10^{-27} & 1.49 \times 10^{-14} \\ \underline{2.93 \times 10^{-1}} & \underline{5.14 \times 10^{-1}} & \underline{3.94 \times 10^{0}} & 6.60 \times 10^{-28} & 8.54 \times 10^{-15} \\ \underline{1.91 \times 10^{-5}} & 1.00 \times 10^{-4} & \underline{9.98 \times 10^{-1}} & 6.86 \times 10^{-6} & \underline{1.26 \times 10^{-2}} \\ 1.45 \times 10^{-16} & 3.96 \times 10^{-4} & \underline{2.25 \times 10^{0}} & \underline{7.28 \times 10^{-1}} & 3.88 \times 10^{-15} \\ \underline{2.38 \times 10^{0}} & 4.29 \times 10^{-6} & \underline{2.45 \times 10^{-10}} & 1.02 \times 10^{-10} & \underline{1.61 \times 10^{0}} \end{pmatrix}, \tag{17}$$

$$R_2^e = \begin{pmatrix} \underline{3.44 \times 10^{-1}} & 2.99 - \times 10^{15} & \underline{1.73 \times 10^{0}} & 2.54 \times 10^{-14} & 9.32 \times 10^{-15} \\ \underline{6.26 \times 10^{-2}} & \underline{6.05 \times 10^{-1}} & \underline{3.61 \times 10^{-1}} & 6.43 - \times 10^{15} & 2.33 \times 10^{-15} \\ \underline{3.28 \times 10^{-2}} & 4.18 \times 10^{-5} & \underline{9.96 \times 10^{-1}} & 7.94 \times 10^{-6} & \underline{3.28 \times 10^{-2}} \\ 1.95 \times 10^{-12} & 2.09 \times 10^{-4} & \underline{7.23 \times 10^{-1}} & \underline{9.11 \times 10^{-1}} & 9.76 \times 10^{-15} \\ \underline{4.47 \times 10^{-2}} & 1.15 \times 10^{-8} & \underline{4.47 \times 10^{-2}} & 1.24 \times 10^{-12} & \underline{3.41 \times 10^{-1}} \end{pmatrix} \tag{18}$$

and

$$R_3^e = \begin{pmatrix} \underline{3.16 \times 10^{-1}} & 2.77 \times 10^{-15} & \underline{1.73 \times 10^{0}} & 2.54 \times 10^{-14} & 8.99 \times 10^{-15} \\ \underline{6.20 \times 10^{-2}} & \underline{5.94 \times 10^{-1}} & \underline{2.46 \times 10^{-1}} & 6.43 \times 10^{-15} & 2.27 \times 10^{-15} \\ \underline{3.16 \times 10^{-2}} & 4.18 \times 10^{-5} & \underline{8.86 \times 10^{-1}} & 7.94 \times 10^{-6} & \underline{3.16 \times 10^{-2}} \\ 1.95 \times 10^{-12} & 1.93 \times 10^{-4} & \underline{5.28 \times 10^{-1}} & \underline{9.03 \times 10^{-1}} & 9.76 \times 10^{-15} \\ \underline{4.09 \times 10^{-2}} & 1.11 \times 10^{-8} & \underline{4.09 \times 10^{-2}} & 1.24 \times 10^{-12} & \underline{3.32 \times 10^{-1}} \end{pmatrix} \tag{19}$$

We see that using this much data we are able to perfectly identify the interactions in the original system.

8 5 Nov. -11 Nov.: Feedback on paper and some results

This week I gave out several copies of the proposal to start getting feedback on it. So far I have heard good things from Dr. Just and Dr. Lin. As I suspected, background research on this project needs more work, so I have begun to work on that. I am still waiting on runs of the code to finish running to obtain results from the fully discrete method of identifying interactions.

I did finish one very basic result on the magnitude of the errors when using the continuous method, the result also holds in the mixed and continuous case, but in both cases assumptions must be made on the error in the approximation over the entire interval. This result relies on a small max error over the entire interval, and as we are using least squares fitting we do not currently have any way to make this error small. I am attempting to work out a result on the error in the interaction matrix based on the error found using the L^2 norm. It doesn't currently seem that promising.

I also finished writing up an analysis of the efficiency of the discrete method of identifying interactions. It was very helpful to break things down into smaller bits of computations and to then put all of those together.