Georgia Tech

THE RELATIVE IMPORTANCE OF DATA POINTS IN SYSTEMS BIOLOGY AND PARAMETER ESTIMATION

JENNY JEONG AND PENG QIU

CREATING THE NEXT®

BACKGROUND

- **Mathematical modeling**

  - To understand complex systems (Interactions among components of the system)

  - Ordinary Differential Equations (ODEs)

  - Many unknown parameters such as kinetic rates

- **Estimating model parameters**

  - Nonlinear Equal weight Least square method:

$$Cost = \frac{1}{2}\sum_{i=1}^{n}\left(Y_i^{obs} - Y_i^{pred}(\theta)\right)^2$$

$$\theta^* = \underset{\theta}{\operatorname{argmin}}(Cost)$$

n : The number of data

Y_i^{obs}: i-th experimenatally observed data

$Y_i^{pred}(\theta)$) : i-th data simulated by the model with θ

MOTIVATION

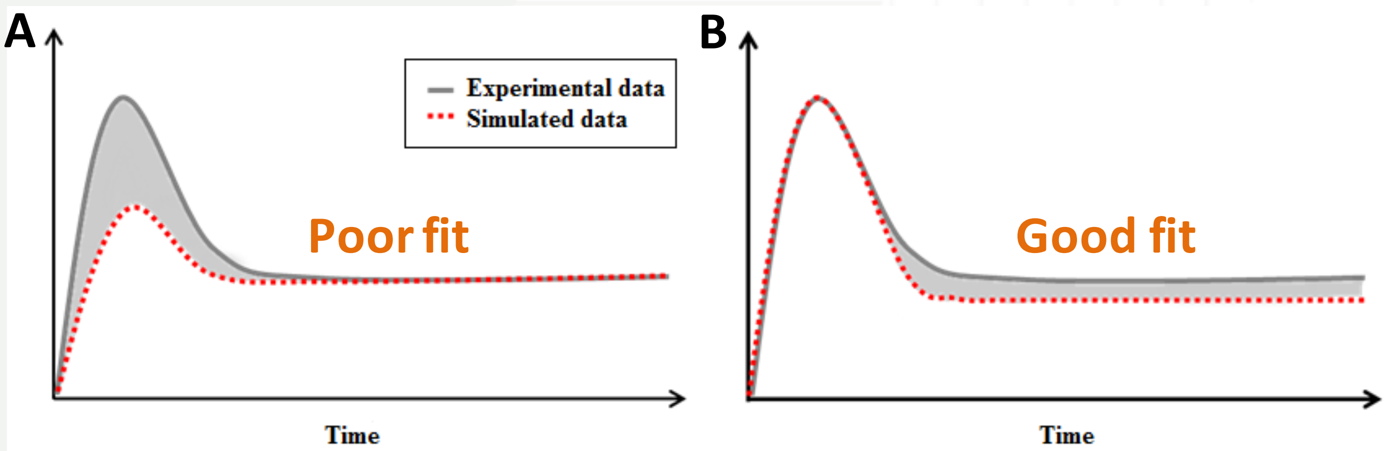
- **Problem 1 :** Limited measureable data → not enough to constrain the parameters
- **Problem 2 :** Limitation of equal weight Least-Sqaure method

  : Treats all data points as equally important.

  → Possible that the equal weight cost function cannot distinguish a poor fit from a good fit (Example)

<u>Example</u> : Assume that we estimated **two sets** of parameters (A and B), which have **the same "Cost" values.**

$$\rightarrow \quad \frac{1}{2}\sum_{i=1}^{n}\left(Y_i^{obs} - Y_i^{pred}(\boldsymbol{\theta}^{\boldsymbol{A}})\right)^2 = \frac{1}{2}\sum_{i=1}^{n}\left(Y_i^{obs} - Y_i^{pred}(\boldsymbol{\theta}^{\boldsymbol{B}})\right)^2$$



WEIGHTED COST FUNCTION

Georgia Tech

- Nonlinear **Weighted** Least square method:

$$Cost = \frac{1}{2}\sum_{i=1}^{n}\left(Y_i^{obs} - Y_i^{pred}(\theta)\right)^2 \boldsymbol{W_i}$$

$$\theta^* = \underset{\theta}{\operatorname{argmin}}(Cost)$$

n : The number of data

Y_i^{obs}: i-th experimenatally observed data

$Y_i^{pred}(\theta)$) : i-th data simulated by the model with θ

$\boldsymbol{W_i}$ **: i-th weight**

- **Intuitive concept of the weights**

1) If we **can** predict one data using the others correctly
   → This data contains **little new information** → **Low** weight

2) If we **cannot** predict one data using the others
   → This data contains **a new information** → **High** weight

WEIGHT (UNCERTAINTY)

– ITERATIVE ALGORITHM TO COMPUTE WEIGHTS BY UNCERTAINTY

- **Uncertainty of estimating parameters (θ) , given data (data)**

  - The second derivative of cost function ≈ Hessian(θ^*) = Fisher Information

  : $\boxed{\mathrm{U}\ (\theta\,|\,\mathrm{data}) = \frac{1}{m}\,\mathrm{trace}(I^{-1})}$

  I^{-1}: inverse of Fisher information matrix

  m : The number of parameters

- **Uncertainty of estimating one data point (data1), given the other data points (I_{others})**

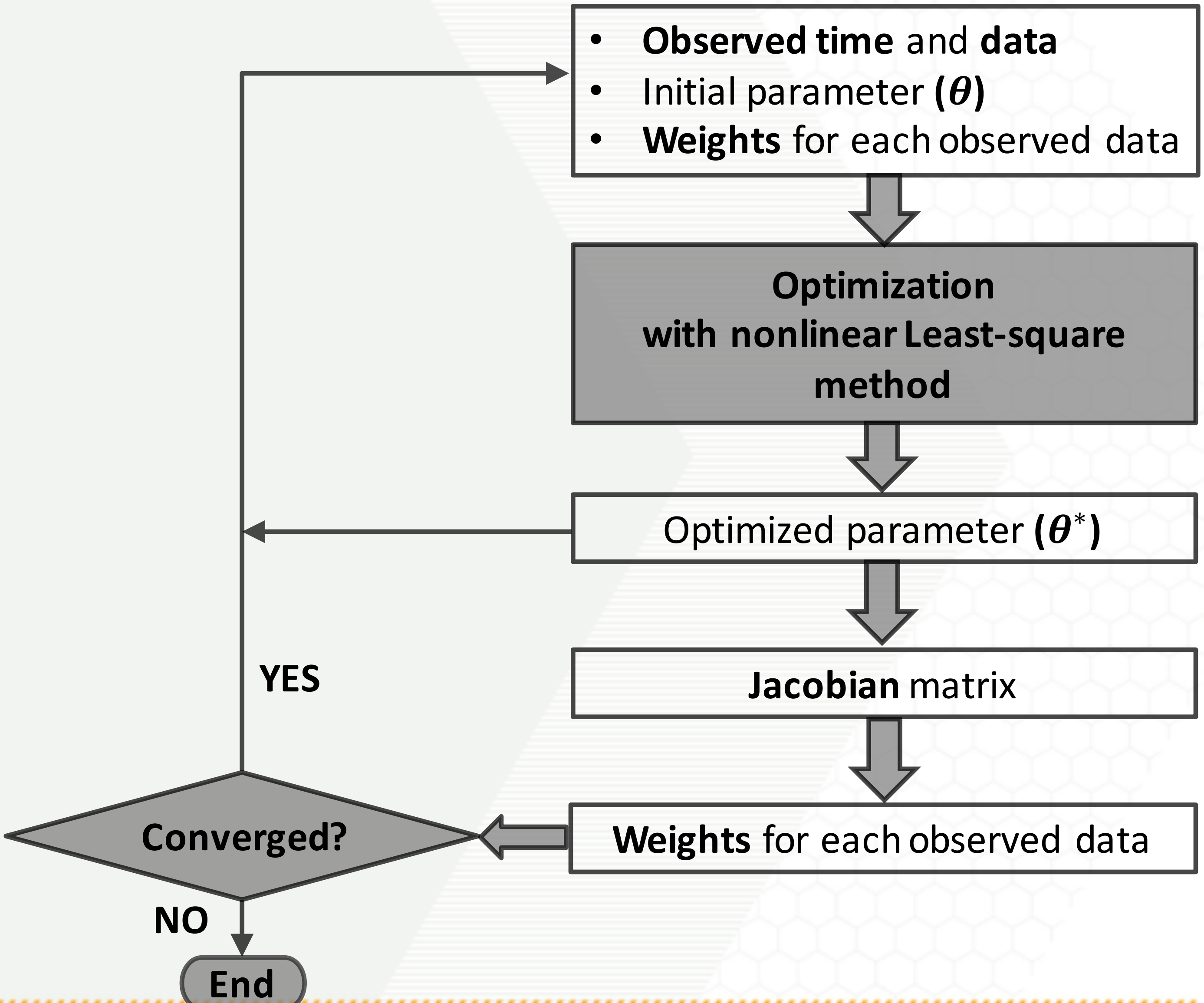
  - The **uncertainty** quantifies how well we can predict the value of a data point, given the other data points.

  : $\boxed{\mathrm{U}(\mathrm{data1}\,|\,\mathrm{the\ other\ data}) = \frac{1}{m}\,\mathrm{trace}(I_1 {I_{others}}^{-1})}$ ← **"Weight of data1"**

  → Repeat this calculation for every data points.

ALGORITHM

– ITERATIVE ALGORITHM TO COMPUTE WEIGHTS BY UNCERTAINTY

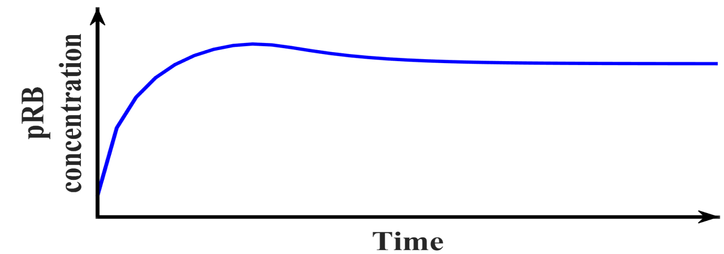


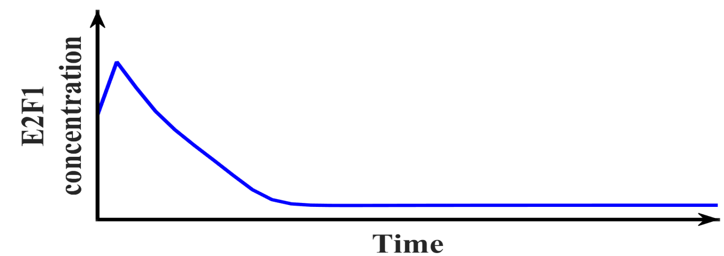
MODEL: G1/S TRANSITION

Georgia Tech

- **The core module of G1/s transition**

- pRB (Retinoblastoma protein) : tumor suppressor
- E2F1 : transcription activator
- pRB inhibits the expression of transcription factor E2F1





- **ODE equations**

$$\frac{d}{dt}[pRB] = K_1 \frac{[E2F1]}{K_{n1}+[E2F1]} \frac{J_{11}}{J_{11}+[pRB]} - \varphi_{pRB}[pRB]$$

$$\frac{d}{dt}[E2F1] = K_p + K_2 \frac{a^2 + [E2F1]^2}{{K_{n2}}^2+[E2F1]^2} \frac{J_{12}}{J_{12}+[pRB]} - \varphi_{E2F1}[E2F1]$$

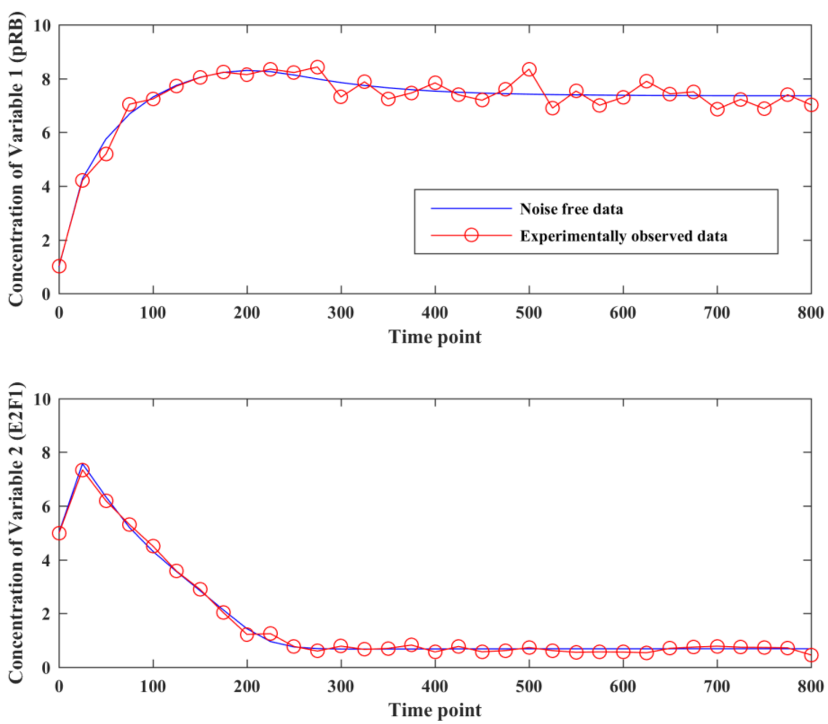
→ **10** unknown model parameters

→ **2** unknown initial values

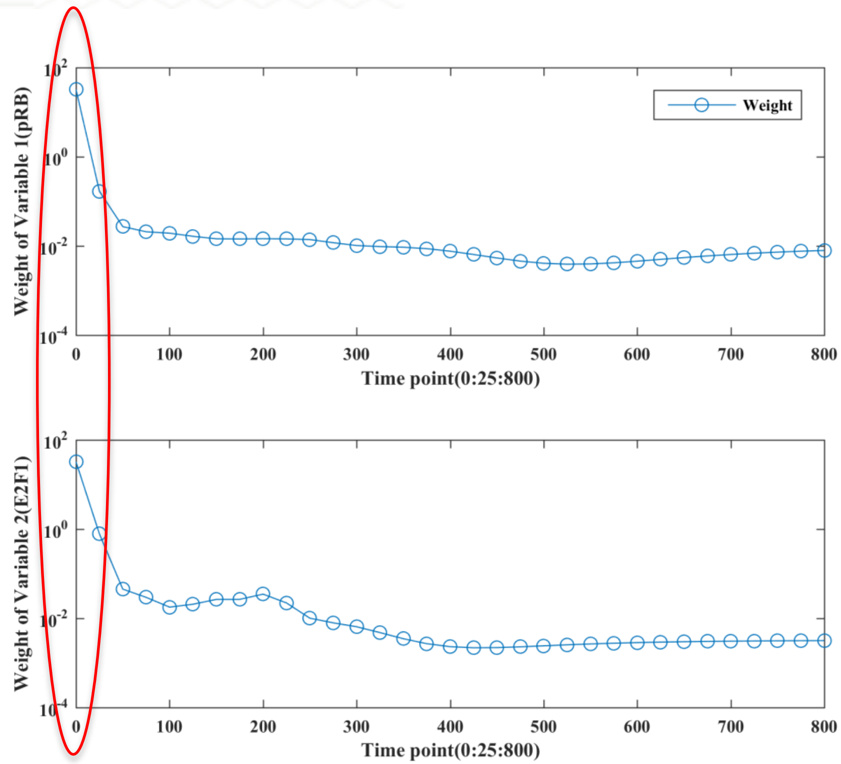
RESULT 1
- EVENLY SPACED TIME POINTS

Georgia Tech

Observed data points



Weights

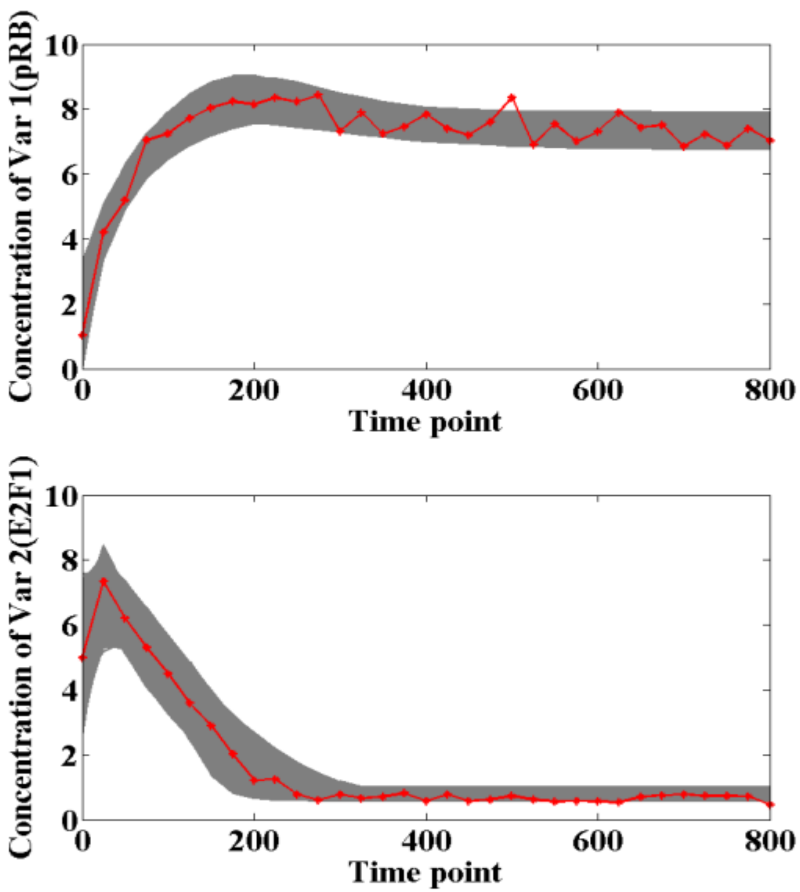


RESULT1

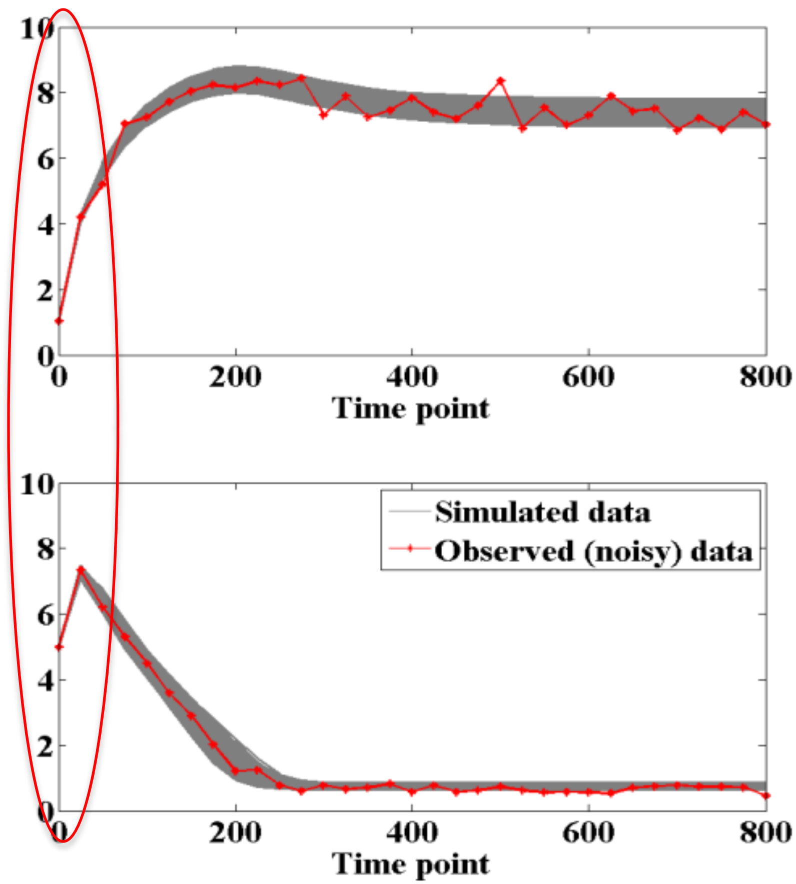
- SAMPLING ALGORITHM

- Collection of acceptable parameter sets near optimal parameter set (Belt)

Equal weight cost function



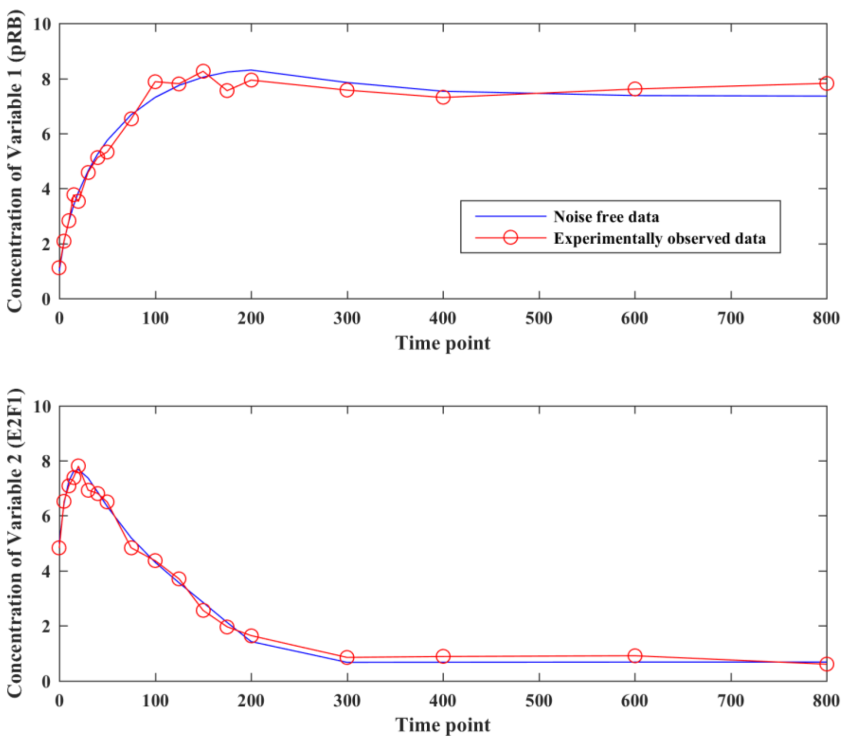
Weighted cost function



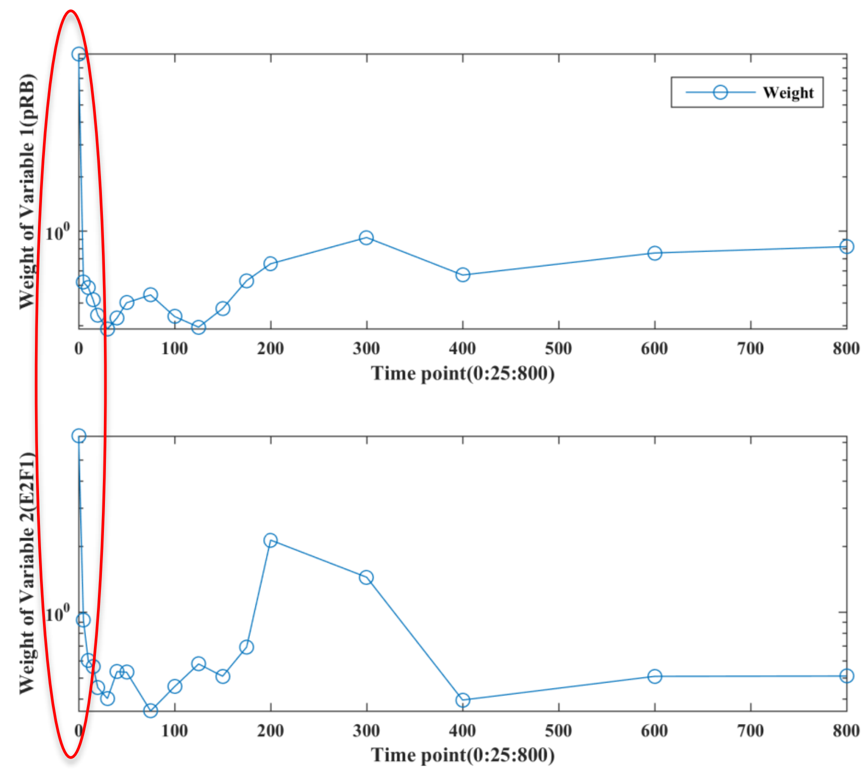
RESULT 2
- UNEVENLY SPACED TIME POINTS

Georgia Tech

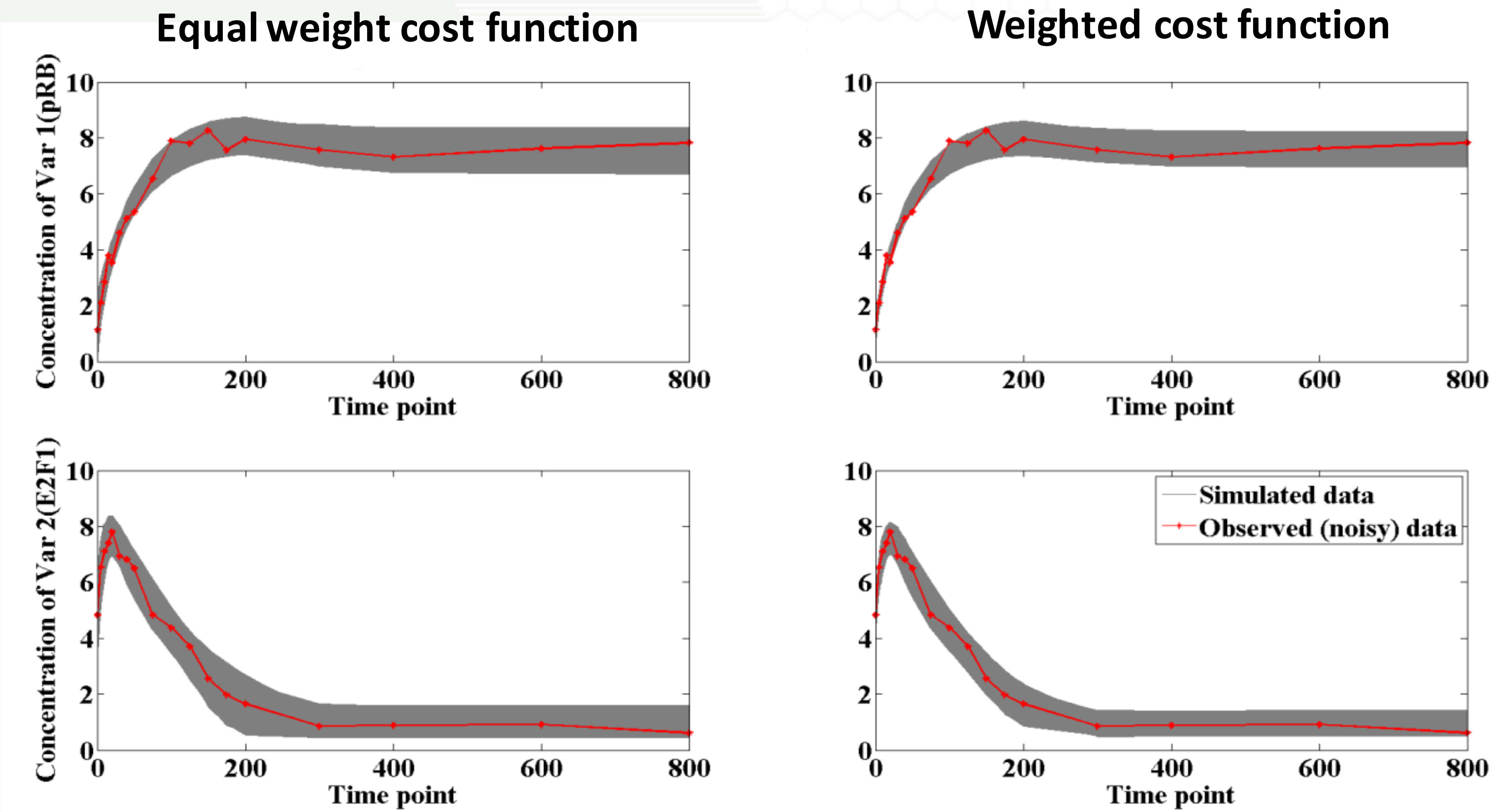
Observed data points



Weights



RESULT 2
- SAMPLING ALGORITHM



DISCUSSION

- **Our weighted least square cost function**

1) Reduce the redundancy in the experimental data
2) Lead to parameter estimation that are not biased toward the redundant measurements in the data.
3) Helpful in strategically choosing measurement time points that avoid redundancy in a real experiment.
4) Prove that what biologists do when they design experiments is reasonable in a mathematical aspect.

ACKNOWLEDGEMENT

- Professor Peng Qiu, and our group members
- National Science Foundation (NSF)
- Georgia Tech college of Engineering
- Georgia Tech SGA