Parameter Estimation in Heat Shock Response of *E. coli*

Anmol Kagrecha, 15D070024

November 27, 2018

1 Introduction

One of the typical problems in computational modeling of biological systems is to estimate the parameters. Only a subset of parameters can be measured experimentally. The other parameters are estimated from the observed data. As a part of this project, I'll be using the example of heat shock response of *E. coli* to illustrate the parameter estimation problem.

A hybrid extended Kalman filter is applied to estimate the parameters from the observed data. Then, a posteriori indentifiability test is employed to check the reliability of estimates. Finally, these tools are used to discriminate between two different models and determine which model describes the observed data better.

2 Biological System

Exposure to high temperatures causes proteins to unfold from their actual three-dimensional structure which might eventually lead to death of the cell. To mitigate the effects of heat, cells express heat shock proteins whose role is to refold unfolded or misfolded proteins.

The heat shock response in *E. coli* is executed using an intricate architecture of feedback loops. There are three major participants that take part in regulation: proteins called chaperones which help in folding, transcription initiators called σ^{32} which help chaperones to be produced and the unfolded proteins.

At physiological temperatures ($30°C$ to $37°C$), amount of σ^{32} is very low. On increase in temperature, σ^{32} accumulates and causes an increase in the amount of chaperones produced. Amount of σ^{32} then decreases to a steady state value. Due to production of chaperones, number of unfolded proteins also decreases.

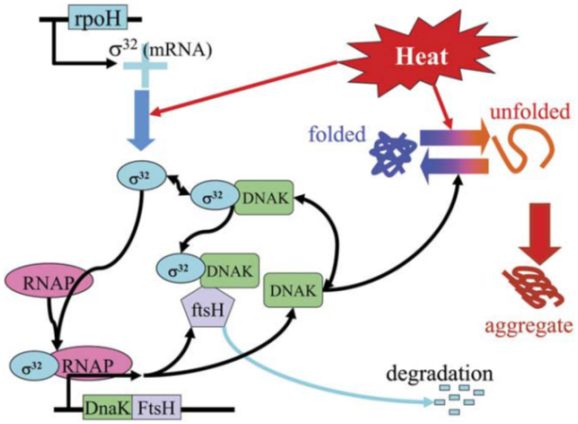


Figure 1: Molecular Implementation

3 The Model

The process can be modelled using the following differential equations -

$$\dot{D_t} = \frac{K_d S_t (1 + K_u U_f)}{1 + K_u U_f + K_s D_t} - \alpha_d D_t$$
$$\dot{S_t} = \eta(T) - \alpha_0 S_t - \frac{\alpha_s K_s D_t S_t}{1 + K_u U_f + K_s D_t}$$
$$\dot{U_f} = K(T)[P_t - U_f] - [K(T) + K_{fold}] D_t$$

where D_t is the number of molecules of chaperones, S_t is the number of molecules of σ^{32}, and U_f is the number of unfolded proteins. The other parameters are constants whose values are given below -

Parameter	**Value**
$\eta(T)$	10 molecule.min^{-1} normally
	60 molecule.min^{-1} in heat shock
$K(T)$	40min^{-1} normally
	80 min^{-1} in heat shock
K_d	3 min^{-1}
α_d	0.015 min^{-1}
α_0	0.03 min^{-1}
α_s	3 min^{-1}
K_s	0.05 molecule^{-1}
K_u	0.024 molecule^{-1}
K_{fold}	6000 min^{-1}
P_t	2 x 10^6 molecules

4 Methods

4.1 Hybrid Extended Kalman Filter

Consider the following system -

$$\dot{x} = f(x, u) + w$$
$$y_k = h_k(x(t_k)) + v_k$$

It is assumed that there is a continuous time process which is to be estimated using discrete time measurements of the output. Assume measurements are available at instants $t_1, ..., t_s$ and $y_1, ..., y_s$ are the corresponding measurements.

The process noise w is assumed to be zero mean Gaussian with covariance matrix Q. The measurement noise v_k is assumed to zero mean Gaussian with covariance matrix $R_k = R$.

The following algorithm is employed to find the estimates -

1. Initial estimate of x_0, $\hat{x}_0^+ = \mathbb{E}[x_0]$
2. Initial error covariance $P_0^+ = \mathbb{E}[(x - x_0^+)(x - x_0^+)^T]$
3. The current a priori estimate $\hat{x}_k^-$ is found by integrating the continuous time process in the interval $[t_{k-1}, t_k]$ using the previous a posteriori estimate $\hat{x}_{k-1}^+$.

$$\dot{\hat{x}} = f(\hat{x}, u)$$
$$\hat{x}(t_{k-1}) = \hat{x}_{k-1}^+$$
$$\Rightarrow \hat{x}_k^- = \hat{x}(t_k)$$

4. The current a priori error covariance estimate P_k^- is found by integrating a Lyapunov equation using previous a posteriori error covariance estimate P_{k-1}^+.

$$\begin{aligned}\dot{P} &= A_k P + P A_k^T + Q \\ P(t_{k-1}) &= P_{k-1}^+ \\ \Rightarrow P_k^- &= P(t_k)\end{aligned}$$

where A_k is the Jacobian of f evaluated at previous a posteriori estimate $\hat{x}_{k-1}^+$.

5. The a posteriori estimate is found by adding the a priori estimate and weighted difference between the actual and the predicted output.

$$\begin{aligned}L_k &= P_k^- H_k^T (H_k P_k^- H_k^T + R)^{-1} \\ \hat{x}_k^+ &= \hat{x}_k^- + L_k(y_k - h_k(\hat{x}_k^-)) \\ P_k^+ &= (I - L_k H_k) P_k^- (I - L_k H_k)^T + L_k R_k L_k^T\end{aligned}$$

where H_k is the jacobian of h evaluated at previous a posteriori estimate $\hat{x}_{k-1}^+$.

6. Repeat the steps (3), (4) and (5) for all time instants $t_1, ..., t_s$.

4.2 A-Posteriori Indentifiability Test

The extended Kalman filter doesn't have convergence guarantees. Divergence of algorithm can be easily detected but we need to look out for situations where we get incorrect estimates due to modeling errors.

Suppose we need to estimate parameters θ from noisy measurements as mentioned before. Assume that p different quantities can be measured. Then the parameter estimation problem can be written as -

$$\begin{aligned}\dot{x} &= f(x, \theta, u) + w \\ \dot{\theta} &= 0 \\ x(t_0) &= x_0 \\ \theta(t_0) &= \theta_0 \\ y_k^{(1)} &= h_k^{(1)}(x(t_k)) + v_k^{(1)} \\ &\vdots \\ y_k^{(p)} &= h_k^{(p)}(x(t_k)) + v_k^{(p)}\end{aligned}$$

We assume v is a zero mean Gaussian with covariance matrix R whose diagonal entries are $\sigma_1^2, ..., \sigma_p^2$.

Suppose that by running the HEKF, we find an estimate $\hat{\theta}_0$ of θ_0. Let $x_{\hat{\theta}_0}(t)$ be the solution corresponding to $\hat{\theta}_0$. Assuming that $x_{\hat{\theta}_0}(t)$ is a reasonable estimate of $x(t)$, we can write -

$$\hat{v}_k^{(i)} = y_k^{(i)} - h_k^{(i)}(x_{\hat{\theta}_0}(t_k))$$

We have s samples of p zero mean Gaussian random variables.

4.2.1 Point Estimate of Noise Variance

$$\xi_i = \frac{\sum_{k=1}^{s} (\hat{v}_k^{(i)})^2}{s - 1}$$

ξ_i should be close to σ_i^2.

4.2.2 Interval Estimate of Noise Variance

We can form interval estimates for ${\hat{\sigma_i}}^2$ corresponding to different confidence coefficients $\gamma = 1 - \delta$. Denote $\chi_{s,\delta}$ as the 100δ-th percentile of the χ^2 distribution with s degrees of freedom. Then, ${\hat{\sigma_i}}^2$ lies in the interval

$$\frac{(s-1)\xi_i}{\chi_{s,1-\delta/2}} \leq {\hat{\sigma_i}}^2 \leq \frac{(s-1)\xi_i}{\chi_{s,\delta/2}}$$

with probability γ.

If σ_i^2 doesn't lie in the above interval, we can reject $\hat{\theta}_0$ with $100\gamma\%$ confidence and if σ_i^2 lies in the above interval, we can accept $\hat{\theta}_0$ with $100\gamma\%$ confidence.

4.3 Model Selection

Suppose there are many models $\Sigma_1, ..., \Sigma_n$ which can be used to explain the observed data. A model selection algorithm can be used to determine which of the models explain the data best. The algorithm is as follows -

1. Run the HEKF on the models $\Sigma_1, ..., \Sigma_n$ to obtain state estimates $\hat{x}_1^+, ..., \hat{x}_n^+$.
2. Compute the estimates of the measurement noises $\hat{v}_1, ..., \hat{v}_n$.
3. Form point and interval estimates of the variance of each component of $\hat{v}_1, ..., \hat{v}_n$.
4. Discard the models in which the interval estimates do not contain the real variances of v_k.
5. Select the model whose variances match the best with the real variances of v_k.

5 Experiments

To illustrate the use of methods developed above, the following is done -

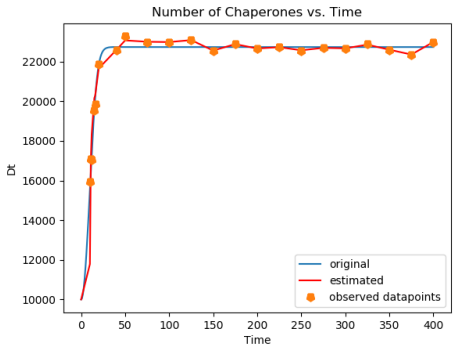
1. Data is generated for four hundred points using the model given in section 3 for the heat shock response.
2. Outputs of D_t and S_t are corrupted by zero mean Gaussian noise of variance $\sigma_1^2 = 1.24 * 10^5$ and $\sigma_2^2 = 737.94$ respectively.
3. A sparse sample of points is provided as the observed data which is similar to what is obtained in the experiments.
4. The sampling is done at seven points between 0 & 40 to collect data for the transient response. Then uniform sampling is done between 50 and 400 at intervals of 25 minutes which gives seventeen points.
5. It is assumed that the parameters α_s and K_d are not available and are needed to be estimated from the observed data.
6. HEKF is applied to the data and estimates of α_s and K_d are obtained.
7. A posteriori identifiability test is applied to see if the model is reasonable.
8. A second model is proposed which doesn't model the actual biological phenomena completely. Model selection algorithm is applied to see which model works better.

5.1 Results of HEKF

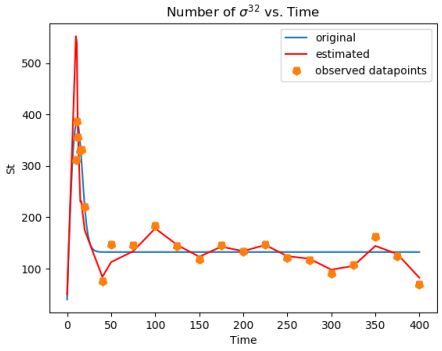
5.2 Parameter Values

The points joined by green lines in Figure 3 are averaged to find the parameter values.

Parameter	**True Value**	**Estimated Value**
K_d	3 min^{-1}	3.0257 min^{-1}
α_s	3 min^{-1}	3.0462 min^{-1}

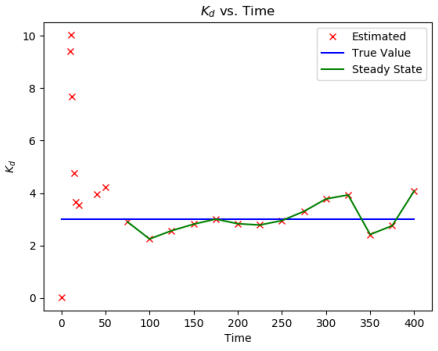


(a) Number of Chaperones

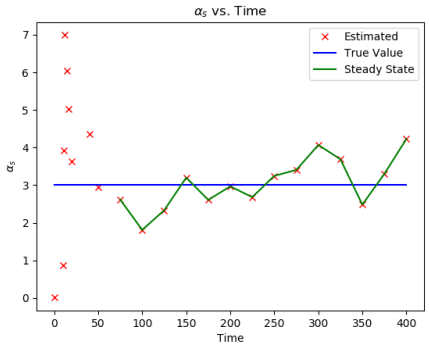


(b) Number of σ^{32}

Figure 2: Variation of Molecules



(a) Variation of K_d



(b) Variation of α_s

Figure 3: Variation of Parameters

5.3 Indentifiability Test

Noise Variance	True Value	Point Estimate	Lower Bound	Upper Bound
σ_1^2	1.24x10^5	0.76x10^5	0.43x10^5	1.46x10^5
σ_2^2	737	961	549	1838

The true value of noise variances lie in the bounds for the required 95% confidence. Hence, we can accept the model with 95% confidence.

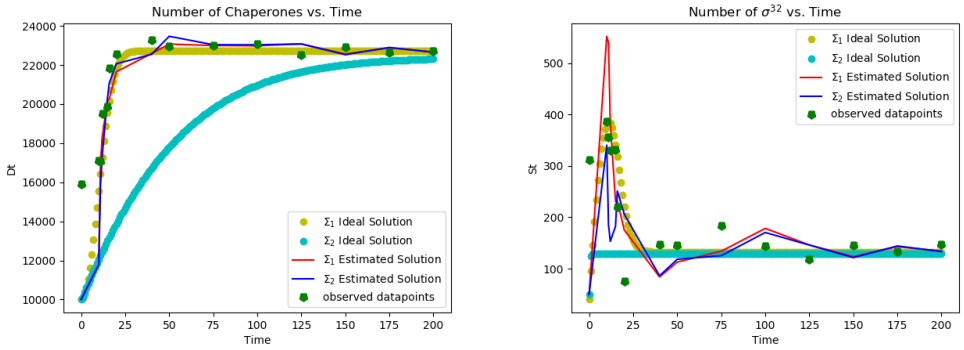
5.4 Model Selection

Consider another model with the following equations -

$$\dot{D}_t = \frac{K_d S_t(1 + K_u U_f)}{1 + K_u U_f + K_s D_t} - \alpha_d D_t$$
$$\dot{S}_t = \eta'(T) - \alpha_0 S_t - \alpha_s S_t$$
$$\dot{U}_f = K(T)[P_t - U_f] - [K(T) + K_{fold}]D_t$$

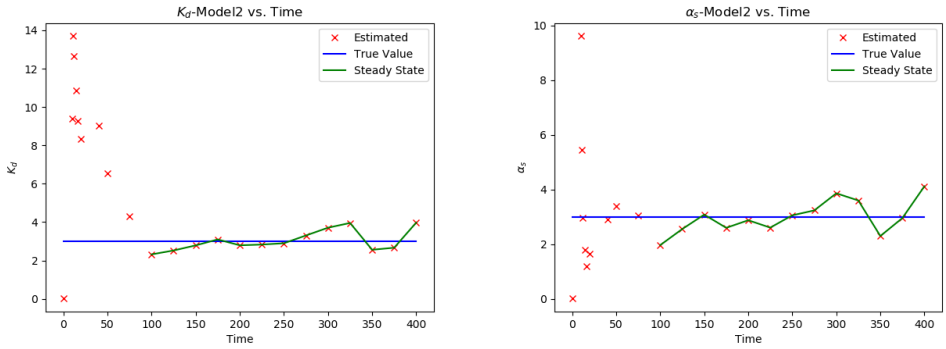
One of the feedback for the number of σ^{32} molecules loops has not been accounted for. Instead the value of $\eta(T)$ for the heat shock response has been changed to account for the increase in steady state value of σ^{32} molecules. The new value of η is $\eta' = 390$ during heat shock response.

On applying the HEKF, we get the following -



(a) Number of Chaperones (b) Number of σ^{32}

Figure 4: Variation of Molecules in Different Models



(a) Variation of K_d (b) Variation of α_s

Figure 5: Variation of Parameters in Model 2

Parameter	**True Value**	**Estimated Value**
K_d	3 min^{-1}	3.031 min^{-1}
α_s	3 min^{-1}	2.983 min^{-1}

On applying the identifiability test, we obtain -

Noise Variance	True Value	Point Estimate	Lower Bound	Upper Bound
σ_1^2	1.24x10^5	1.33x10^7	7.78x10^6	2.61x10^7
σ_2^2	737	13793	7875	26375

Value of the variances don't lie in the intervals needed for 95% confidence. We reject Σ_2 with 95% confidence and accept Σ_1 with 95% confidence.

6 Conclusion

It is demonstrated how methods like HEKF can be used for parameter estimation in computational biology problems. Secondly, a method of a posteriori identification was illustrated. Finally, HEKF and the identifiability test were used to discriminate between two competing models and choose the better one.

7 References

1. G Lillacci, Khammash M (2010), Parameter Estimation and Model Selection in Computational Biology
2. El-Samad H, Prajna S, Papachristodoulou A, Doyle J, Khammash M (2006) Advanced methods and algorithms for biological networks analysis
3. Simon D (2006), Optimal State Estimation

8 Code

An associated jupyter notebook can be found in this github repository.