Moment closure and block updating for parameter inference in stochastic biological models

Peter Milner

Supervised by Dr Colin Gillespie and Prof. Darren Wilkinson

Newcastle University

1. Motivation
   - Auto regulatory gene network
2. Method
   - Moment closure
   - Approximate Likelihood
   - Updating missing data
3. Results
   - Data set D_1 - G and I unobserved
   - Data set D_2 - g, i, G and I unobserved
   - Conclusions and future work

Motivation

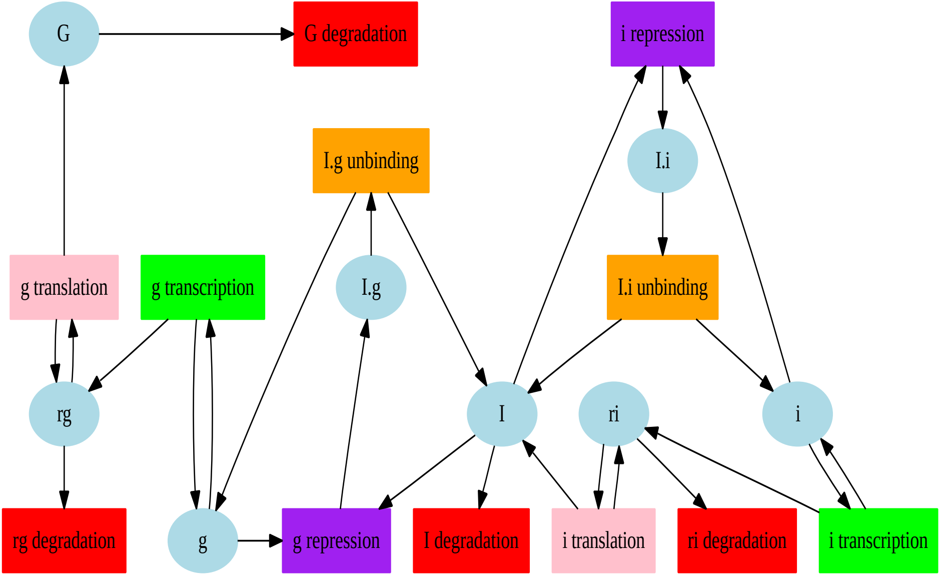
- One of the key problems in systems biology is inferring rate parameters of stochastic kinetic biochemical network models
- If we know:
  1. The description of the system
  2. The initial conditions
  3. The rate parameters

  then we can model the system (stochastically or deterministically)
- Test our understanding of the system/modelling assumptions
- How do we infer these rate parameters initially?

Auto regulatory gene network

Throughout this talk we will use an auto regulatory gene network as an example

- This network has 6 species $Z = (r_g, r_i, g, i, G, I)$
- r_g and r_i are mRNA
- g and i are genes
- G and I are proteins
- Where I regulates the production of itself and G by binding to genes i and g


G
G degradation
i repression
I.g unbinding
I.i
g translation
g transcription
I.g
I.i unbinding
rg
I
ri
i
rg degradation
g
g repression
I degradation
i translation
ri degradation
i transcription

- We can write our model as a list of reactions:

$$
\begin{aligned}
R_1 &: \quad I + i \xrightarrow{c_1} I \cdot i & R_2 &: \quad I \cdot i \xrightarrow{c_2} I + i \\
R_3 &: \quad I + g \xrightarrow{c_3} I \cdot g & R_4 &: \quad I \cdot g \xrightarrow{c_4} I + g \\
R_5 &: \quad i \xrightarrow{c_5} i + r_i & R_6 &: \quad r_i \xrightarrow{c_6} r_i + I \\
R_7 &: \quad g \xrightarrow{c_7} g + r_g & R_8 &: \quad r_g \xrightarrow{c_8} r_g + G \\
R_9 &: \quad r_i \xrightarrow{c_9} \emptyset & R_{10} &: \quad r_g \xrightarrow{c_{10}} \emptyset \\
R_{11} &: \quad I \xrightarrow{c_{11}} \emptyset & R_{12} &: \quad G \xrightarrow{c_{12}} \emptyset
\end{aligned}
$$

- We assume mass action kinetics

- From the chemical master equation we can find a set of ODE's for the moments (see Gillespie, 2009)
- ODE's for the moments usually depend on higher order moments, e.g. for two species X_1, X_2

$$\dot{\mu}_{1,1} = (\mu_{2,0} - \mu_{1,0}^2)c_1 - (\mu_{1,1} - \mu_{1,0}\mu_{0,1})c_1 - \mu_{2,1}c_1 + \cdots$$

where $\mu_{n,m} = E(X_1^n X_2^m)$

- By assuming an underlying distribution we can write higher order moments in terms of lower order moments e.g. $\mu_3 = 3\mu_2\mu_1 - 2\mu_1^3$
- Giving a closed set of ODE's
- We have assumed a underlying Gaussian distribution throughout this talk, other distributions could be used e.g. Poisson, Log-Normal

- From the chemical master equation we can find a set of ODE's for the moments (see Gillespie, 2009)
- ODE's for the moments usually depend on higher order moments, e.g. for two species X_1, X_2

$$\dot{\mu}_{1,1} = (\mu_{2,0} - \mu_{1,0}^2)c_1 - (\mu_{1,1} - \mu_{1,0}\mu_{0,1})c_1 - \mu_{2,1}c_1 + \cdots$$

  where $\mu_{n,m} = E(X_1^n X_2^m)$

- By assuming an underlying distribution we can write higher order moments in terms of lower order moments e.g. $\mu_3 = 3\mu_2\mu_1 - 2\mu_1^3$
- Giving a closed set of ODE's
- We have assumed a underlying Gaussian distribution throughout this talk, other distributions could be used e.g. Poisson, Log-Normal

- Let $\mathbf{x}(t_i)$ be the i^{th} discrete time observation of the process
- We propose

$$\mathbf{x}(t_i)|\mathbf{x}(t_{i-1}) \sim N(\mu, \mathbf{\Sigma}),$$

where μ and $\mathbf{\Sigma}$ are calculated from the moment closure approximation of the process

- Appealing to the Markov property we can approximate the likelihood of the rate parameters (Θ) for a given realisation $\mathbf{x} = \{\mathbf{x}(t_i) : i \in 1, \ldots, N\}$,

$$L(\Theta|\mathbf{x}) = \prod_{i=1}^{N} P[\mathbf{x}(t_i)|\mathbf{x}(t_{i-1})]$$

- We use a Metropolis-Hastings sampler to explore the parameter space (random walk with innovations $w_j \sim N(0, \sigma^2)$)

- Let $\mathbf{x}(t_i)$ be the i^{th} discrete time observation of the process
- We propose

$$\mathbf{x}(t_i)|\mathbf{x}(t_{i-1}) \sim N(\mu, \mathbf{\Sigma}),$$

  where μ and $\mathbf{\Sigma}$ are calculated from the moment closure approximation of the process

- Appealing to the Markov property we can approximate the likelihood of the rate parameters (Θ) for a given realisation $\mathbf{x} = \{\mathbf{x}(t_i) : i \in 1, \ldots, N\}$,

$$L(\Theta|\mathbf{x}) = \prod_{i=1}^{N} P[\mathbf{x}(t_i)|\mathbf{x}(t_{i-1})]$$

- We use a Metropolis-Hastings sampler to explore the parameter space (random walk with innovations $w_j \sim N(0, \sigma^2)$)

Given discrete time observations

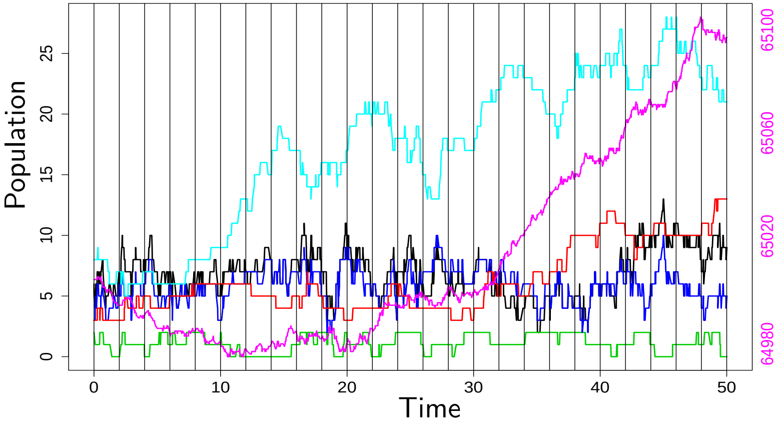


Figure: *A stochastic realisation from the auto-regulatory gene network. With observations on each species;* r_g *(cyan),* r_i *(red), g(blue), i(green), G(pink) and I(black).* $Z(0) = (8, 2, 3, 2, 65000, 6)$

This may be a bit hopeful so we consider $D_1 = \{r_g, r_i, g, i\}$

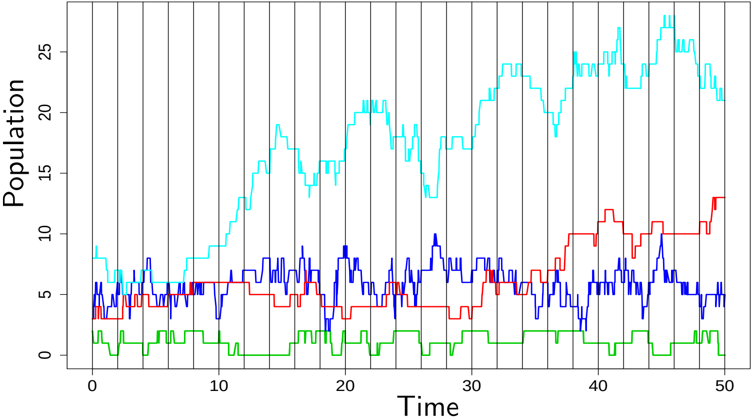


Figure: *A stochastic realisation from the auto-regulatory gene network. With observations on each species; r_g(cyan), r_i(red), g(blue) and i(green).*

This may still be a bit hopeful so we consider $D_2 = \{r_g, r_i\}$

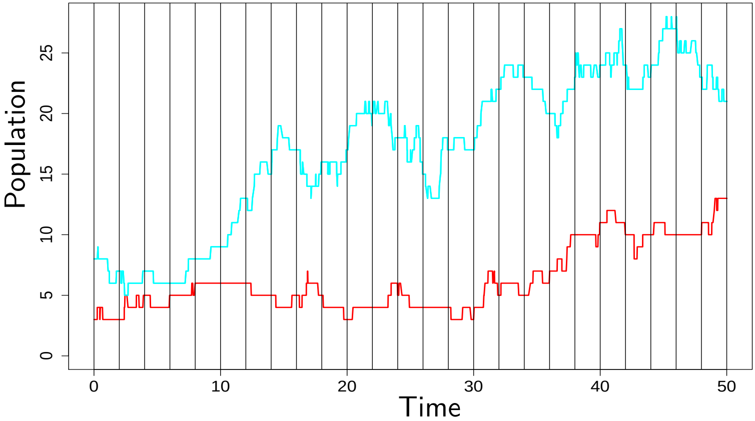


Figure: *A stochastic realisation from the auto-regulatory gene network. With observations on each species; r_g (cyan) and r_i (red).*

Bridge updating

- How to update the unobserved species?
- We want to be able to update our missing data conditioned on all the data we can
- We do this using a block updating scheme (following Durham & Gallant (2002))

- Suppose we have data $Z(t) = (X(t), Y(t))^T$, where $X(t)$ is known
- Our goal is to sample $Y(t_{i+1})$ conditioned on $Z(t_j)$, $Z(t_M)$ and $X(t_{i+1})$, where $t_j < t_{i+1} < t_M$
- Such a sample can be approximated by a skeleton bridge $Y(t_{i+1})$ for $i = j, j+1, \ldots, M-2$
- Constructing such a bridge is non trivial so a Metropolis Hastings step is used

- Suppose we have data $Z(t) = (X(t), Y(t))^T$, where $X(t)$ is known
- Our goal is to sample $Y(t_{i+1})$ conditioned on $Z(t_j)$, $Z(t_M)$ and $X(t_{i+1})$, where $t_j < t_{i+1} < t_M$
- Such a sample can be approximated by a skeleton bridge $Y(t_{i+1})$ for $i = j, j+1, \ldots, M-2$
- Constructing such a bridge is non trivial so a Metropolis Hastings step is used

- We can construct a proposal distribution for Y^{i+1}

$$q(Y^{i+1}|X^{i+1}, Z^i, Z^M, \theta) \sim N\left\{\mu^*, \frac{M-i-1}{M-i}\Sigma^*\right\},$$

where,

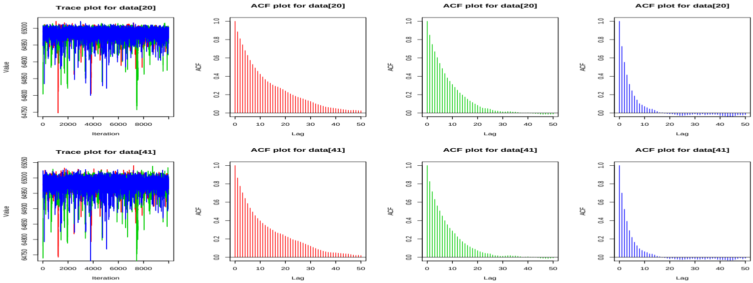
$$\mu^* = \mu_y + \Sigma_{yx}(\Sigma_{xx})^{-1}(X^{i+1} - \mu_x)$$
$$\Sigma^* = [\Sigma_{yy} - \Sigma_{yx}(\Sigma_{xx})^{-1}\Sigma_{xy}]$$

and,

$$\mu_x = X^i + \frac{X^M - X^i}{M-i}, \quad \mu_y = Y^i + \frac{Y^M - Y^i}{M-i}$$

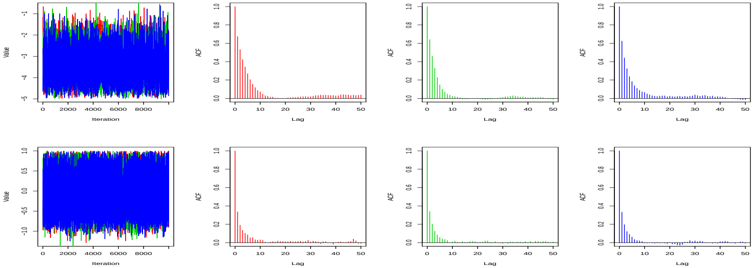
- We can sample $q(\cdot|\cdot)$ for $i = j, \ldots, M-2$ to construct a skeleton bridge

- We will now apply our block updating method to the two data sets
  1. D_1: We have 50 observations on $X(t) = (r_g, r_i, g, i)$ and impute $Y(t) = (G, I)$
  2. D_2: We have 50 observations on $X(t) = (r_g, r_i)$ and impute $Y(t) = (g, i, G, I)$
- In each data set we have limited the number of genes (i) to 2 and the steady state value for $G \approx 70000$.
- We would like to know:
  1. Which block size is best for updating the missing data
  2. How much we can find out about our rate parameters and unobserved species



	$M = 1$	$M = 4$	$M = 8$
G[20]	344	571	1417
G[41]	382	623	1547
Mean (all)	465	747	1656

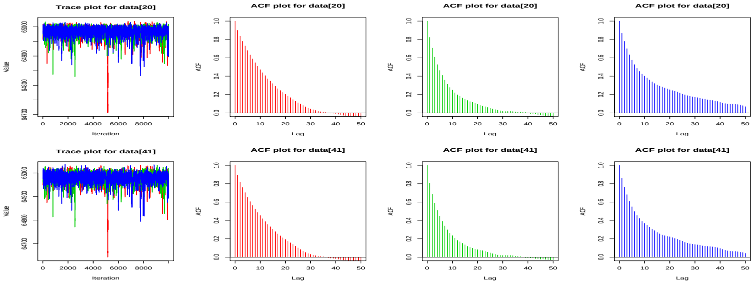
Table: *Effective sample sizes for different blocks (G)*



	$M = 1$	$M = 4$	$M = 8$	True	Mean	sd
c1	1305	1936	1830	0.08	0.052	0.043
c4	3125	3618	3624	0.9	0.96	0.50

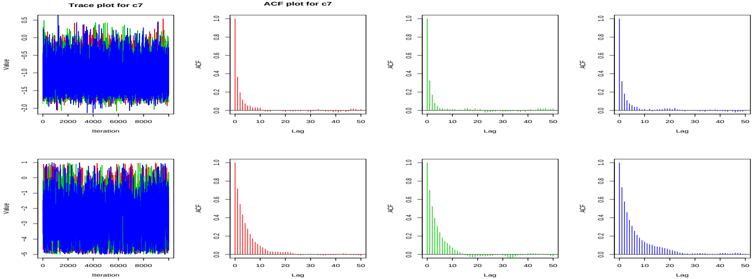
Table: *Effective sample sizes of the parameters c_1 and c_4, for different blocks.*

Data set D_2 - g, i, G and I unobserved



	$M = 1$	$M = 4$	$M = 8$
G[20]	411	731	402
G[41]	413	759	445
Mean (all)	460	912	495

Table: *Effective sample sizes for different blocks updating G.*



	$M = 1$	$M = 4$	$M = 8$		True	Mean	sd
c_7	3458	4410	3765		0.35	0.37	0.15
c_{11}	1266	1431	1146		0.05	0.15	0.24

Table: *Effective sample sizes of c_7 and c_{11} for different blocks.*

Conclusions and future work

- The most efficient choice of block length is model specific, a block length of 4-8 gave the best results in testing
- Conditioning on the observed data leads to more efficient updating of the unobserved data
- Develop a model for *Bacillus subtilis* sporulation and apply our method

References

- Durham, G. B. & Gallant, R. A. (2002), Numerical techniques for maximum likelihood estimation of continuous time diffusion processes. *Journal of Business and Economic Statistics* **20**, 279-316.
- Gillespie, C. S. (2009), Moment closure approximations for mass-action models. *IET Systems Biology* **3**, 52-58.
- Wilkinson, D. J. (2006). Stochastic Modelling for Systems Biology.*Chapman & Hall/CRC*