Variational inference for Markov jump processes

Guido Sanguinetti

Joint work with Manfred Opper (T.U. Berlin)

The University Of Sheffield.

Talk plan

- Markov jump processes in the life sciences
- Basics of jump processes
- Variational mean field
- Results
- Future extensions?

Markov Jump Processes

- Describe the dynamics of populations of interacting species
- Changes in populations are discrete (birth/death) and happen at random times
- Rates of births and deaths for a species depend on population levels for all species
- Many examples: chemical kinetics, predator-prey models, telecoms, etc.

Relevance to life sciences

- Environmental systems
- Single-cell dynamics
- Low concentration pathogens (see F. Bois talk)
- Usually dynamics is simulated using (variants of) Gillespie's algorithm (SSA)
- SSA does not include observations and requires full specification of system
- Inference is hard: one MCMC approach (Boys et al; "A bloody mess", A. Golightly, PESB2007)

Mathematical notation

- Mathematically, a Markov Jump Process (MJP) is a family of discrete random variables indexed by time **x**(t).
- Markov property:

$$p\left(\mathbf{x}\left(t_N\right)|\mathbf{x}\left(t_{N-1}\right),\ldots,\mathbf{x}\left(t_0\right)\right)=p\left(\mathbf{x}\left(t_N\right)|\mathbf{x}\left(t_{N-1}\right)\right).$$

- Process rates f(**x**'|**x**) defined by

$$\lim_{\delta t\to 0} p\left(\mathbf{x}'\left(t+\delta t\right)|\mathbf{x}\left(t\right)\right)=\delta_{\mathbf{x}'\mathbf{x}}+f\left(\mathbf{x}'|\mathbf{x}\right).$$

- Process rates satisfy $\sum$f(**x**'|**x**)=0 (normalisation)

The University Of Sheffield.

Master equation

- The process rates and the time-dependent probabilities p(**x**) are linked by the *Master equation*

$$\frac{dp_t(\mathbf{x})}{dt} = \sum_{\mathbf{x}' \neq \mathbf{x}} \left[-p_t(\mathbf{x}) f\left(\mathbf{x}'|\mathbf{x}\right) + p_t\left(\mathbf{x}'\right) f\left(\mathbf{x}'|\mathbf{x}\right) \right].$$

- Notice that the Master equation is in fact a system of S^D ODEs, with S the number of states accessible to each species, and D the number of species

Variational Inference

- Deterministic approximate inference technique to approximate a probability distribution p with a simpler one q.
- The measure of proximity between distributions is the Kullback-Leibler (KL) divergence

$$KL(q||p) = -\int dq \log \frac{p}{q}$$

- Notice that you do not need explicit knowledge of p, just of the expectation of log(p) under q

KL divergence between MJPs

- Consider two MJPs p and q, with rates f and g.
- A MJP can be viewed as a probability distribution over trajectories of the system; denote a trajectory as $\mathbf{x}_{0:K}=(\mathbf{x}(t_0),...,\mathbf{x}(t_0+K\delta t))$.
- The KL divergence between these processes is

$$KL\left(q||p\right)=\sum_{\mathbf{x}_{0:K}} q(\mathbf{x}_{0:K})\ln\frac{q(\mathbf{x}_{0:K})}{p(\mathbf{x}_{0:K})}=$$

$$=\sum_{k=0}^{K-1}\sum_{\mathbf{x}_k} q(\mathbf{x}_k)\sum_{\mathbf{x}_{k+1}} q(\mathbf{x}_{k+1}|\mathbf{x}_k)\ln\frac{q(\mathbf{x}_{k+1}|\mathbf{x}_k)}{p(\mathbf{x}_{k+1}|\mathbf{x}_k)}+K_0$$

The University Of Sheffield.

Continuous limit

- By taking the limit $\delta t \to 0$, we can rewrite the KL divergence in terms of process rates.
- Setting the initial KL $K_0=0$, we get

$$KL(q||p) = \int_0^T dt \sum_{\mathbf{x}} q_t(\mathbf{x}) \times$$

$$\sum_{\mathrm{x}':\mathrm{x}' \neq \mathrm{x}} \left\{ g(\mathrm{x}'|\mathrm{x}) \ln \frac{g(\mathrm{x}'|\mathrm{x})}{f(\mathrm{x}'|\mathrm{x})} + f(\mathrm{x}'|\mathrm{x}) - g(\mathrm{x}'|\mathrm{x}) \right\}$$

The University Of Sheffield.

Posterior processes

- We assume now to have discrete-time, noise-corrupted observations $\mathbf{y}_l$ of the process.
- The noise model is given by $\hat{p}\left(\mathbf{y}_l|\mathbf{x}\left(t_l\right)\right)$.
- The posterior process is still Markovian and is

$$p_{post}(\mathbf{x}_{0:K}) = \frac{1}{Z} p_{prior}(\mathbf{x}_{0:K}) \times \prod_{l=1}^{N} \hat{p}\left(\mathbf{y}_l|\mathbf{x}\left(t_l\right)\right).$$

- The KL divergence with a process q is

$$KL(q\|p_{post}) = \ln Z + KL(q\|p_{prior}) - \sum_{l=1}^{N} E_q\left[\ln p\left(\mathbf{y}_l|\mathbf{x}\left(t_l\right)\right)\right].$$

Mean-field approximation

- The main assumption we will make is that the approximating process q is factorised

$$q_t(\mathbf{x}) = \prod_{i=1}^{D} q_{it}(x_i) \qquad g_t\left(\mathbf{x}'|\mathbf{x}\right) = \sum_{i=1}^{D} \prod_{j \neq i} \delta_{x'_j, x_j} g_{it}\left(x'_i|x_i\right)$$

- The KL-divergence then becomes

$$KL\left(q \| p_{post}\right) = \ln Z - \sum_{l=1}^{N} E_q\left[\ln \hat{p}\left(\mathbf{y}_l|\mathbf{x}\left(t_l\right)\right)\right] +$$

$$\int_0^T dt \sum_i \sum_x q_{it}(x) \sum_{x':x' \neq x} \left\{ g_{it}\left(x'|x\right) \ln \frac{g_{it}\left(x'|x\right)}{\hat{f}_i\left(x'|x\right)} + \tilde{f}_i\left(x'|x\right) - g_{it}\left(x'|x\right) \right\}$$

so that it decomposes as a sum over species i.

The University Of Sheffield.

Constraints

- Each of the factors in the approximating distribution satisfy a 1-dimensional Master equation.
- The complexity is decreased from S^D to DS.
- Using Lagrange multiplier functions $\lambda_i(x,t)$, we obtain the Lagrangian

$$L = KL\left(q||p_{post}\right) +$$

$$-\sum_i \int_0^T dt \sum_x \lambda_i(x,t) \left(\partial_t q_{it}(x) - \sum_{x' \neq x} \left\{ g_{it}\left(x|x'\right) q_{it}\left(x'\right) - g_{it}\left(x'|x\right) q_{it}(x) \right\} \right).$$

with the standard boundary condition $\lambda_i(x,T)=0$.

The University Of Sheffield.

Functional derivatives

- To optimise the Lagrangian, we compute its functional derivatives

$$\frac{\delta L}{\delta q_{it}(x)} = \sum_{x' \neq x} \left[g_{it}\left(x'|x\right) \ln \frac{g_{it}\left(x'|x\right)}{\widehat{f}_i\left(x'|x\right)} - g_{it}\left(x'|x\right) + \widetilde{f}_i\left(x'|x\right) \right] + \partial_t \lambda_i\left(x,t\right) +$$

$$\sum_{x'} g_{it}\left(x'|x\right) \left\{ \lambda_i\left(x',t\right) - \lambda_i\left(x,t\right) \right\} - \sum_l \ln \widehat{p}\left(\mathbf{y}_l|\mathbf{x}\left(t\right)\right) \delta\left(t - t_l\right) = 0 \tag{1}$$

$$\frac{\delta L}{\delta g_{it}\left(x'|x\right)} = q_{it}\left(x\right) \left(\ln \frac{g_{it}\left(x'|x\right)}{\widehat{f}_i\left(x'|x\right)} + \lambda_i\left(x',t\right) - \lambda_i\left(x,t\right) \right) = 0 \tag{2}$$

The University Of Sheffield.

Backward equation

- Inserting equation (2) into (1) and defining $r_i(x,t)=e^{-\lambda_i(x,t)}$ we obtain

$$\frac{dr_i(x,t)}{dt} = \sum_{x' \neq x} \left(\tilde{f}_i\left(x'|x\right) r_i\left(x,t\right) - \hat{f}_i\left(x'|x\right) r_i\left(x',t\right) \right)$$

- For each species, this is a system of S linear differential equations valid at all times except the observation times.

The University Of Sheffield.

Including observations

- We assume for simplicity that observations for different species are independent

$$\widehat{p}\left(\mathbf{y}_l|\mathbf{x}(t)\right) = \prod_i \widehat{p}_i\left(y_{il}|x_i(t_l)\right) \qquad \forall l$$

- At observations, the Lagrange multiplier has a discontinuity

$$\lim_{t\to t_l^-} r_i\left(x,t\right) = \widehat{p}_i\left(y_{il}|x_i(t_l)\right) \lim_{t\to t_l^+} r_i\left(x,t\right).$$

- Numerically, this can become difficult to solve.
- Easier to solve the ODE for the ratios of r.

Algorithm: approximating the posterior (E-step)

- We can find an approximate posterior as follows:

1. Choose a species i;
2. From an initial guess of $q_i(x)$, compute the averaged rates;
3. Solve the backward equation backward in time from $r_i(x,T)=1$;
4. Solve the Master equation forward in time to update $q_i(x)$;
5. Iterate 2-4 until an optimum is found;
6. Choose another species and follow 2-5;
7. Iterate until convergence.

- This procedure guarantees a decrease in KL

Parameter estimation (M-step)

- Prior parameters can also be estimated by minimising the KL divergence, once an approximating distribution is obtained.
- E-step and M-step can be iterated until convergence is reached.
- Local minima can be a problem, thought must be given to initialisation.

The University Of Sheffield.

Application: Lotka-Volterra

- Transition rates are given by

$$f_{prey}(p+1|p) = \alpha p \quad f_{prey}(p-1|p) = \beta P p$$

$$f_{predator}(P+1|P) = \delta P p \quad f_{predator}(P-1|P) = \gamma P$$

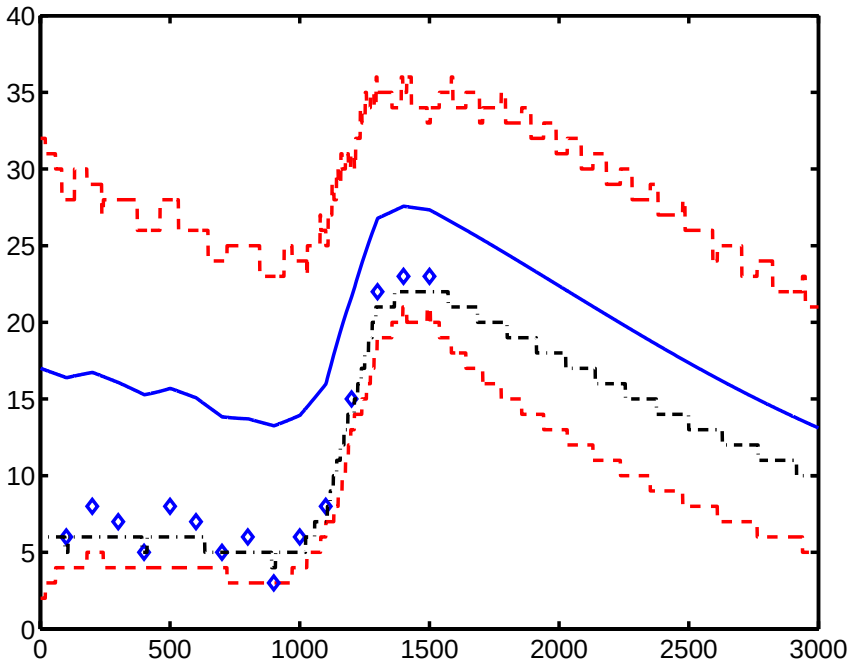
- M-step equations are given by

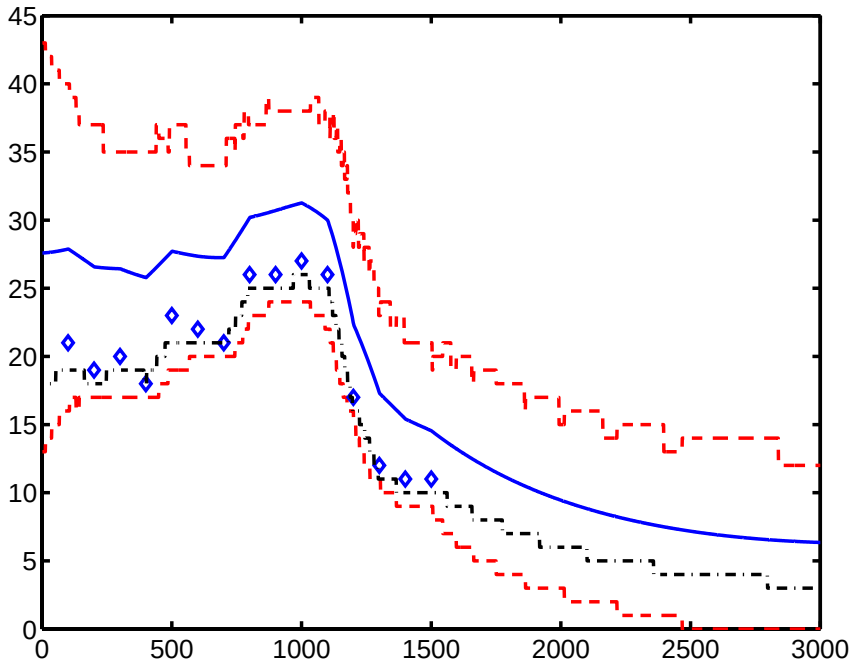
$$\alpha = \frac{\int_0^T \left\langle g_{preyt}\left(x+1|x\right)\right\rangle_{preyt}}{\int_0^T dt \left\langle x\right\rangle_{preyt}}, \quad \beta = \frac{\int_0^T \left\langle g_{preyt}\left(x-1|x\right)\right\rangle_{preyt}}{\int_0^T dt \left\langle x\right\rangle_{preyt}\left\langle y\right\rangle_{predatort}},$$

$$\gamma = \frac{\int_0^T \left\langle g_{predatort}\left(y-1|y\right)\right\rangle_{predatort}}{\int_0^T dt \left\langle y\right\rangle_{predatort}}, \quad \delta = \frac{\int_0^T \left\langle g_{predatort}\left(y+1|y\right)\right\rangle_{predatort}}{\int_0^T dt \left\langle y\right\rangle_{predatort}\left\langle x\right\rangle_{preyt}}.$$

The University Of Sheffield.

Results: Lotka-Volterra





Posterior predator (left) and prey (right) distribution. Diamonds are data, dashed-dotted posterior mode, solid posterior mean.

The noise model is asymmetric and given by

$$\widehat{p}_i\left(y_{il}|x_i\left(t_l\right)\right) \propto \left[\frac{1}{5|y_{il}-x_i(t_l)|}+0.001\right]$$

The University Of Sheffield.

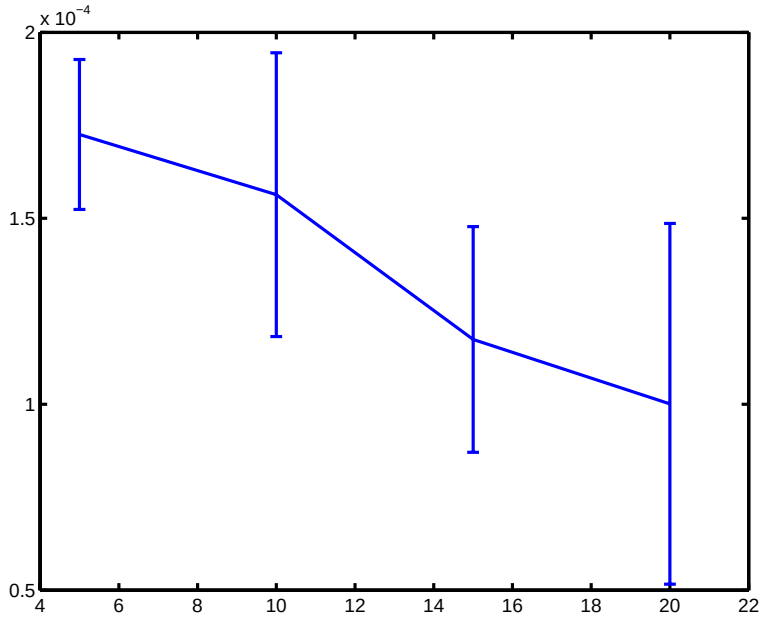
Parameter estimates

- Estimates of the parameters were reasonable

$$\alpha = 5.14 \times 10^{-4} \quad (5 \times 10^{-4}) \qquad \beta = 6.95 \times 10^{-5} \quad (1 \times 10^{-4})$$

$$\gamma = 7.26 \times 10^{-4} \quad (5 \times 10^{-4}) \qquad \delta = 5.77 \times 10^{-5} \quad (1 \times 10^{-4})$$

Estimates appear to converge to the true value when more observations are available.
Results show average of five runs per data-set size.



The University Of Sheffield.

Application: gene auto-regulation

- Autoregulation is one of the fundamental blocks in gene regulatory networks.
- Protein represses transcription of its own coding gene.
- We use a logical approximation. Process rates are

$$f_{RNA}(x+1|x,y) = \alpha\left(1 - 0.99 \times \Theta\left(y - y_c\right)\right) \qquad f_{RNA}(x-1|x,y) = \beta x$$

$$f_p(y+1|x,y) = \gamma x \qquad f_p(y-1|x,y) = \delta y$$

The critical parameter y_c is the integer threshold for protein count above which repression begins. Θ is the Heaviside step function.

The University Of Sheffield.

Parameter estimation

- Fixed point equations for the parameters are obtained

$$\alpha = \frac{\int_0^T dt \left\langle g_{RNA}\left(x+1|x\right)\right\rangle_{q_{RNA}}}{T\left(1-0.99\frac{1}{T}\int h\left(y_c\right)dt\right)} \qquad \beta = \frac{\int_0^T dt \left\langle g_{RNA}\left(x-1|x\right)\right\rangle_{q_{RNA}}}{\int_0^T dt \left\langle x\right\rangle_{q_{RNA}}}$$

$$\gamma = \frac{\int_0^T dt \left\langle g_p\left(x+1|x\right)\right\rangle_{q_p}}{\int_0^T dt \left\langle x\right\rangle_{q_{RNA}}} \qquad \delta = \frac{\int_0^T dt \left\langle g_p\left(x-1|x\right)\right\rangle_{q_p}}{\int_0^T dt \left\langle y\right\rangle_{q_p}}$$

where $h\left(y_c\right) = \sum_{y \geq y_c} q_p\left(y\right)$ is the posterior probability that the protein levels will be above the critical threshold (function of time).

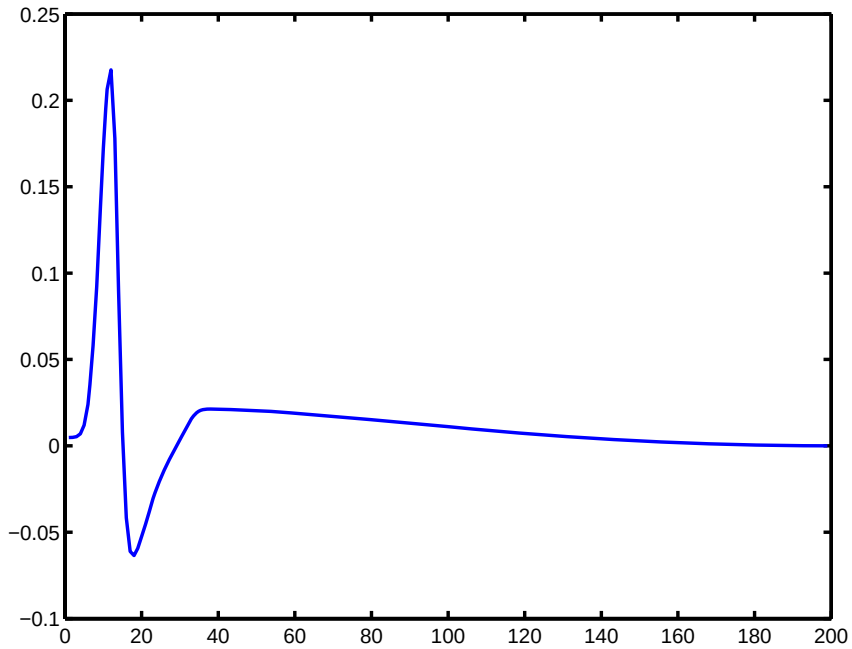
The University Of Sheffield.

Identifiability of critical parameter

- The critical parameter y_c is found by optimising (by search) the free energy

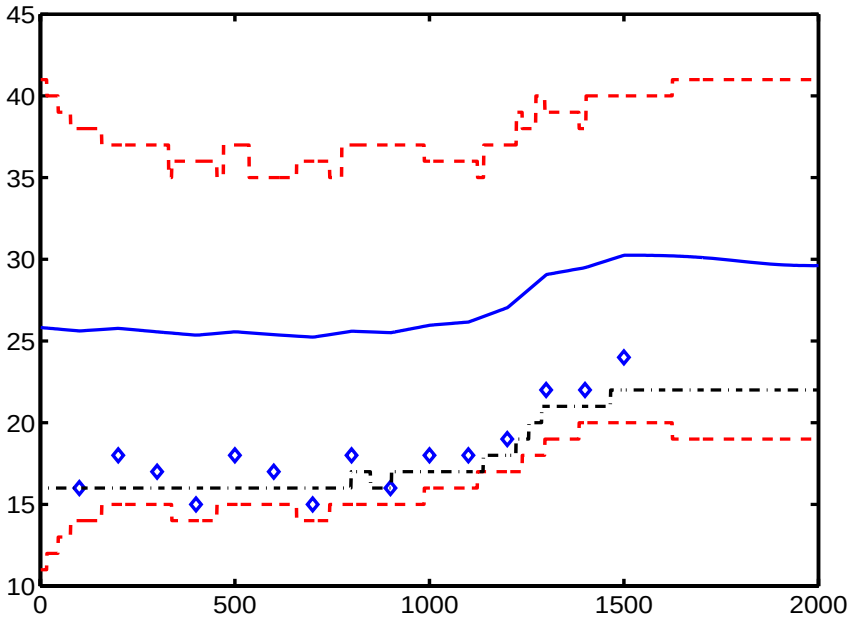
$$\mathcal{L}_{y_c} = const + \left\{ 2\int_0^T dt\bar{g}h\left(y_c\right) + \log\left[1 - 0.99\frac{1}{T}\int_0^T h\left(y_c\right)dt\right]\int_0^T dt\bar{g} \right\}$$

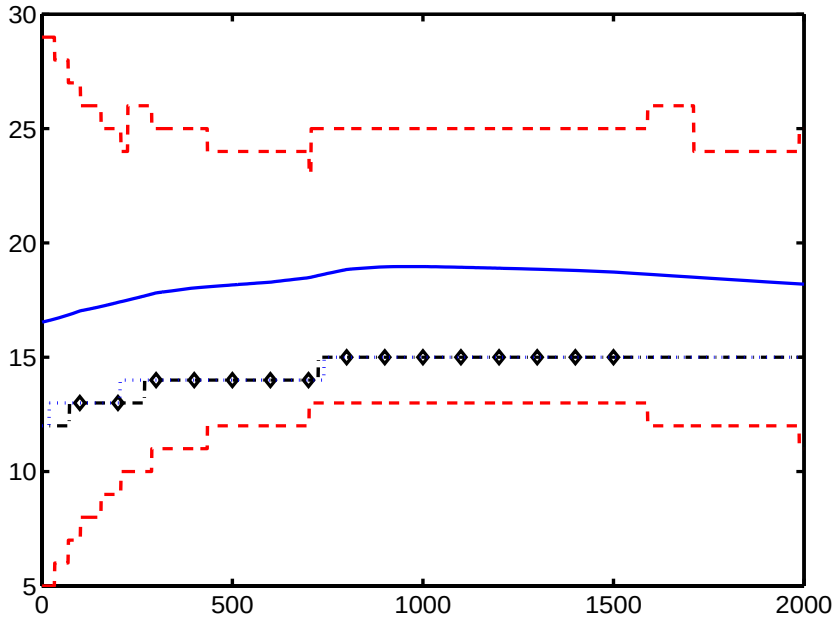
Clearly, if the protein levels in the data never exceed the threshold, y_c is not identifiable. If it does, we get free energies with a well defined minimum.



The University Of Sheffield.

Autoregulatory network: results





$\alpha = 3.4 \times 10^{-3} \quad (2 \times 10^{-3})$

$\beta = 4.4 \times 10^{-5} \quad (6 \times 10^{-5})$

$\gamma = 1.9 \times 10^{-4} \quad (5 \times 10^{-4})$

$\delta = 6.7 \times 10^{-5} \quad (7 \times 10^{-5})$

The University Of Sheffield.

Conclusions

- Efficient framework for posterior inference and parameter estimation in MJPs.
- Oustrips MCMC (Boys *et al.*) by orders of magnitude.
- Numerics are tricky and could be improved.
- Issues swept under the carpet: regularisation, cut-off on number of states.
- Readily extends to missing data.

Future work

- More complex, realistic networks.
- Missing data.
- Hybrid systems: one species has large numbers, so approximate as deterministic or diffusion.
- Hybrid systems: particles also have a spatial dimension, and diffuse in the environment.