Bayesian inference for partially observed Markov processes, with application to systems biology

Darren Wilkinson
http://tinyurl.com/darrenjw
School of Mathematics & Statistics, Newcastle University, UK

Bayes–250
Informatics Forum
Edinburgh, UK
5th–7th September, 2011

Systems biology modelling

- Uses accurate high-resolution time-course data on a relatively small number of bio-molecules to parametrise carefully constructed mechanistic dynamic models of a process of interest based on current biological understanding
- Traditionally, models were typically deterministic, based on a system of ODEs known as the Reaction Rate Equations (RREs)
- It is now increasingly accepted that biochemical network dynamics at the single-cell level are intrinsically stochastic
- The theory of stochastic chemical kinetics provides a solid foundation for describing network dynamics using a Markov jump process

Stochastic Chemical Kinetics

Stochastic molecular approach:

- Statistical mechanical arguments lead to a Markov jump process in continuous time whose instantaneous reaction rates are directly proportional to the number of molecules of each reacting species
- Such dynamics can be simulated (exactly) on a computer using standard discrete-event simulation techniques
- Standard implementation of this strategy is known as the "Gillespie algorithm" (just discrete event simulation), but there are several exact and approximate variants of this basic approach

Lotka-Volterra system

Trivial (familiar) example from population dynamics (in reality, the "reactions" will be elementary biochemical reactions taking place inside a cell)

Reactions

$$X \longrightarrow 2X \qquad \text{(prey reproduction)}$$
$$X + Y \longrightarrow 2Y \qquad \text{(prey-predator interaction)}$$
$$Y \longrightarrow \emptyset \qquad \text{(predator death)}$$

- X – Prey, Y – Predator
- We can re-write this using matrix notation

Forming the matrix representation

The L-V system in tabular form

	Rate Law	LHS		RHS		Net-effect	
	$h(\cdot, c)$	X	Y	X	Y	X	Y
R_1	$c_1 x$	1	0	2	0	1	0
R_2	$c_2 xy$	1	1	0	2	-1	1
R_3	$c_3 y$	0	1	0	0	0	-1

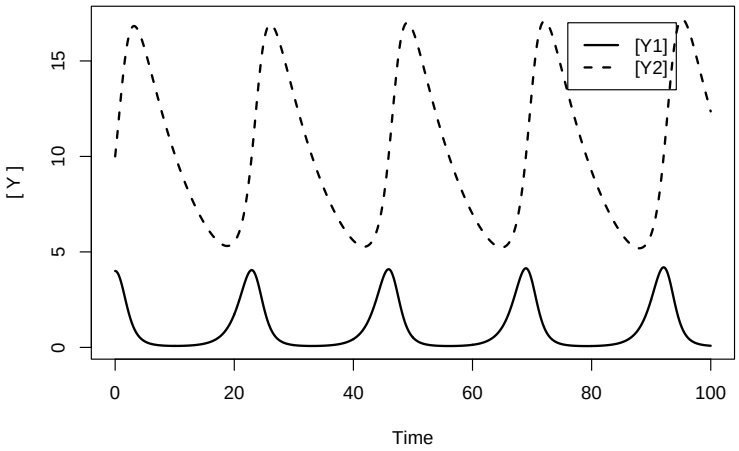
Call the 3×2 net-effect (or reaction) matrix N. The matrix $S = N'$ is the stoichiometry matrix of the system. Typically both are sparse. The SVD of S (or N) is of interest for structural analysis of the system dynamics...

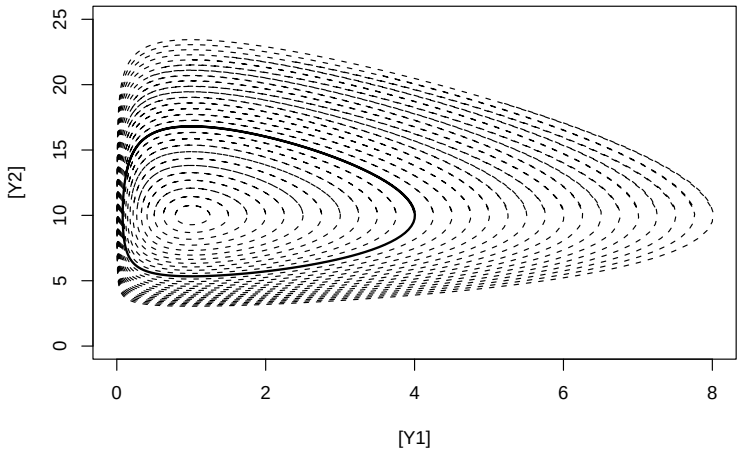
Stochastic chemical kinetics

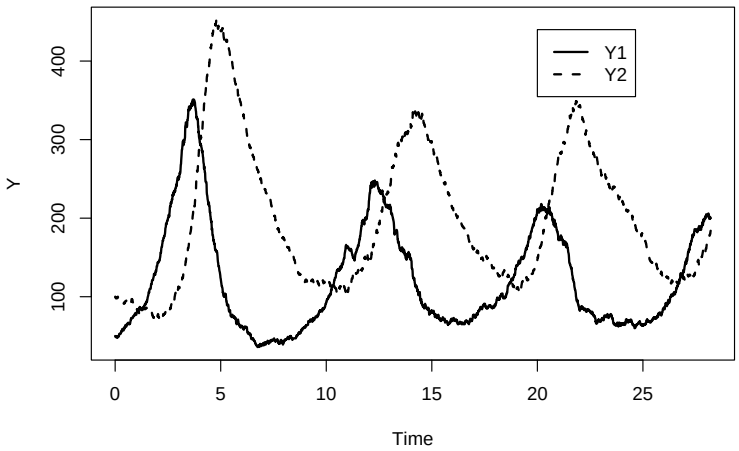
- u species: $\mathcal{X}_1, \ldots, \mathcal{X}_u$, and v reactions: $\mathcal{R}_1, \ldots, \mathcal{R}_v$
- $\mathcal{R}_i: \; p_{i1}\mathcal{X}_1 + \cdots + p_{iu}\mathcal{X}_u \longrightarrow q_{i1}\mathcal{X}_1 + \cdots + q_{iu}\mathcal{X}_u, \; i = 1, \ldots, v$
- In matrix form: $P\mathcal{X} \longrightarrow Q\mathcal{X}$ (P and Q are sparse)
- $S = (Q - P)'$ is the stoichiometry matrix of the system
- X_{jt}: # molecules of $\mathcal{X}_j$ at time t. $X_t = (X_{1t}, \ldots, X_{ut})'$
- Reaction $\mathcal{R}_i$ has hazard (or rate law, or propensity) $h_i(X_t, c_i)$, where c_i is a rate parameter, $c = (c_1, \ldots, c_v)'$, $h(X_t, c) = (h_1(X_t, c_1), \ldots, h_v(X_t, c_v))'$ and the system evolves as a Markov jump process
- For mass-action stochastic kinetics,

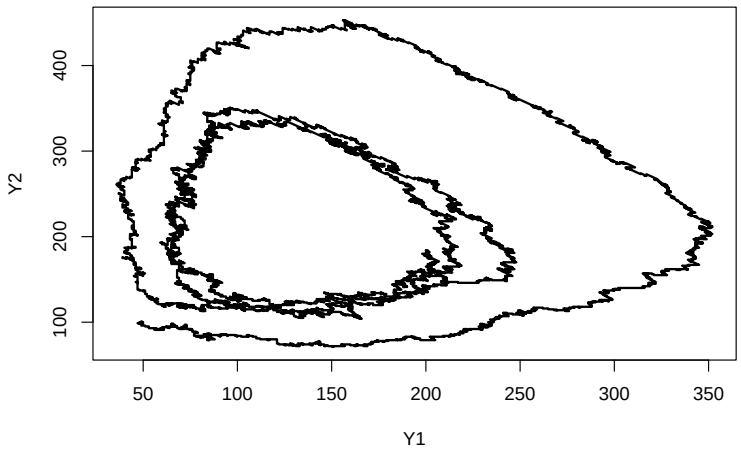
$$h_i(X_t, c_i) = c_i \prod_{j=1}^{u} \binom{X_{jt}}{p_{ij}}, \quad i = 1, \ldots, v$$

The Lotka-Volterra model

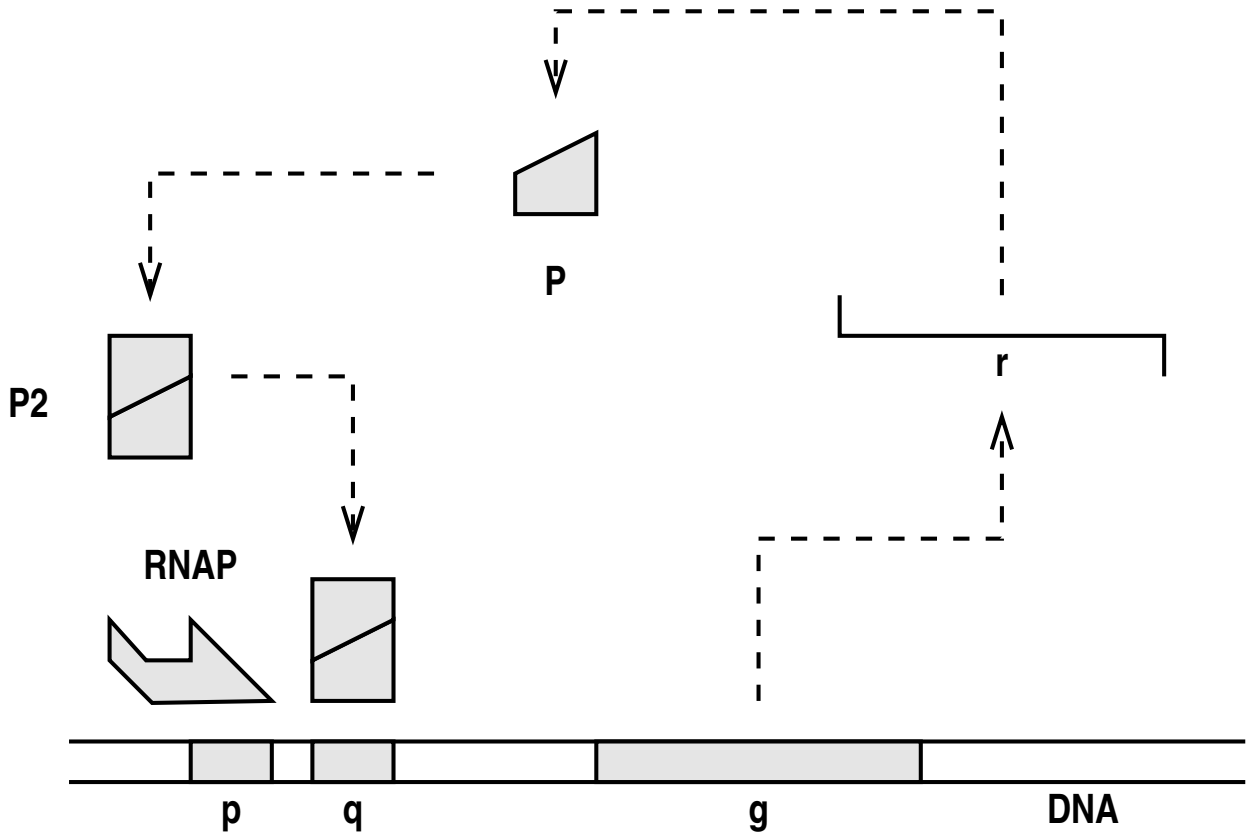








Example — genetic auto-regulation



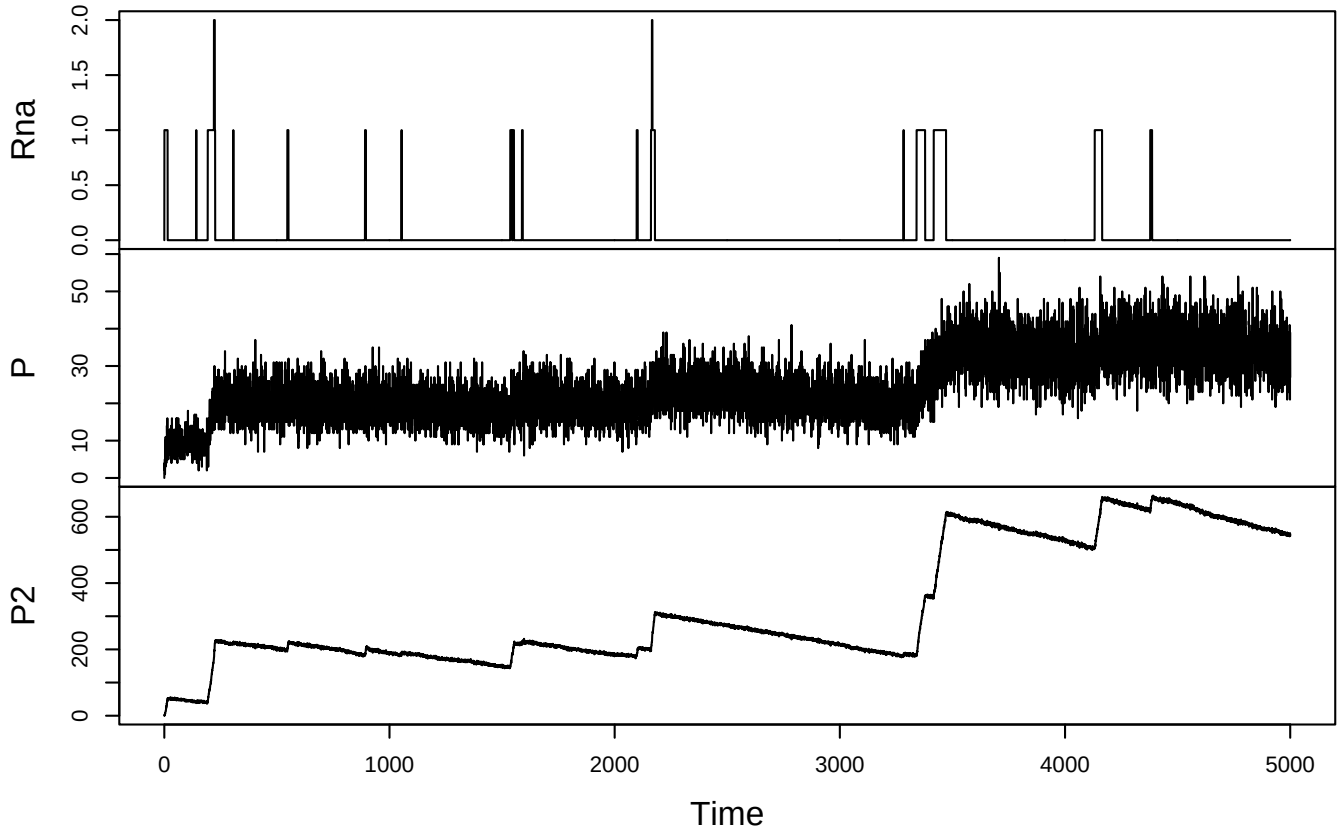
Biochemical reactions

Simplified view:

Reactions

$g + P_2 \longleftrightarrow g \cdot P_2$	Repression
$g \longrightarrow g + r$	Transcription
$r \longrightarrow r + P$	Translation
$2P \longleftrightarrow P_2$	Dimerisation
$r \longrightarrow \emptyset$	mRNA degradation
$P \longrightarrow \emptyset$	Protein degradation

Simulated realisation of the auto-regulatory network



Partially observed Markov process (POMP) models

- Continuous-time Markov process: $\mathbf{X} = \{X_s | s \geq 0\}$ (for now, we suppress dependence on parameters, θ)
- Think about integer time observations (extension to arbitrary times is trivial): for $t \in \mathbb{N}$, $\quad \mathbf{X}_t = \{X_s | t-1 < s \leq t\}$
- Sample-path likelihoods such as $\pi(\mathbf{x}_t | x_{t-1})$ can often (but not always) be computed (but are often computationally difficult), but discrete time transitions such as $\pi(x_t | x_{t-1})$ are typically intractable
- Partial observations: $\mathcal{D} = \{d_t | t = 1, 2, \ldots, T\}$ where

$$d_t | X_t = x_t \sim \pi(d_t | x_t), \qquad t = 1, \ldots, T,$$

where we assume that $\pi(d_t | x_t)$ can be evaluated directly (simple measurement error model)

Bayesian inference for POMP models

- Most "obvious" MCMC algorithms will attempt to impute (at least) the skeleton of the Markov process: $X_0, X_1, \ldots, X_T$
- This will typically require evaluation of the intractable discrete time transition likelihoods, and this is the problem...
- Two related strategies:
  - Data augmentation: "fill in" the entire process in some way, typically exploiting the fact that the sample path likelihoods are tractable — works in principle, but difficult to "automate", and exceptionally computationally intensive due to the need to store and evaluate likelihoods of cts sample paths
  - Likelihood-free (AKA plug-and-play): exploits the fact that it is possible to forward simulate from $\pi(x_t|x_{t-1})$ (typically by simulating from $\pi(\mathbf{x}_t|x_{t-1})$), even if it can't be evaluated
- Likelihood-free is really just a special kind of augmentation strategy

Bayesian inference

- Let $\pi(\mathbf{x}|c)$ denote the (complex) likelihood of the simulation model
- Let $\pi(\mathcal{D}|\mathbf{x},\tau)$ denote the (simple) measurement error model
- Put $\theta = (c,\tau)$, and let $\pi(\theta)$ be the prior for the model parameters
- The joint density can be written

$$\pi(\theta,\mathbf{x},\mathcal{D}) = \pi(\theta)\pi(\mathbf{x}|\theta)\pi(\mathcal{D}|\mathbf{x},\theta).$$

- Interest is in the posterior distribution $\pi(\theta,\mathbf{x}|\mathcal{D})$

Marginal MH MCMC scheme

- Full model: $\pi(\theta, \mathbf{x}, \mathcal{D}) = \pi(\theta)\pi(\mathbf{x}|\theta)\pi(\mathcal{D}|\mathbf{x}, \theta)$
- Target: $\pi(\theta|\mathcal{D})$ (with $\mathbf{x}$ marginalised out)
- Generic MCMC scheme:
  - Propose $\theta^\star \sim f(\theta^\star|\theta)$
  - Accept with probability $\min\{1, A\}$, where

$$A = \frac{\pi(\theta^\star)}{\pi(\theta)} \times \frac{f(\theta|\theta^\star)}{f(\theta^\star|\theta)} \times \frac{\pi(\mathcal{D}|\theta^\star)}{\pi(\mathcal{D}|\theta)}$$

- $\pi(\mathcal{D}|\theta)$ is the "marginal likelihood" (or "observed data likelihood", or...)

LF-MCMC

- Posterior distribution $\pi(\theta, \mathbf{x}|\mathcal{D})$
- Propose a joint update for θ and $\mathbf{x}$ as follows:
  - Current state of the chain is $(\theta, \mathbf{x})$
  - First sample $\theta^\star \sim f(\theta^\star|\theta)$
  - Then sample a new path, $\mathbf{x}^\star \sim \pi(\mathbf{x}^\star|\theta^\star)$
  - Accept the pair $(\theta^\star, \mathbf{x}^\star)$ with probability $\min\{1, A\}$, where

$$A = \frac{\pi(\theta^\star)}{\pi(\theta)} \times \frac{f(\theta|\theta^\star)}{f(\theta^\star|\theta)} \times \frac{\pi(\mathcal{D}|\mathbf{x}^\star, \theta^\star)}{\pi(\mathcal{D}|\mathbf{x}, \theta)}.$$

- Note that choosing a prior independence proposal of the form $f(\theta^\star|\theta) = \pi(\theta^\star)$ leads to the simpler acceptance ratio

$$A = \frac{\pi(\mathcal{D}|\mathbf{x}^\star, \theta^\star)}{\pi(\mathcal{D}|\mathbf{x}, \theta)}$$

"Ideal" joint MCMC scheme

- LF-MCMC works by making the proposed sample path consistent with the proposed new parameters, but unfortunately not with the data
- Ideally, we would do the joint update as follows
  - First sample $\theta^\star \sim f(\theta^\star|\theta)$
  - Then sample a new path, $\mathbf{x}^\star \sim \pi(\mathbf{x}^\star|\theta^\star, \mathcal{D})$
  - Accept the pair $(\theta^\star, \mathbf{x}^\star)$ with probability $\min\{1, A\}$, where

$$A = \frac{\pi(\theta^\star)}{\pi(\theta)} \frac{\pi(\mathbf{x}^\star|\theta^\star)}{\pi(\mathbf{x}|\theta)} \frac{f(\theta|\theta^\star)}{f(\theta^\star|\theta)} \frac{\pi(\mathcal{D}|\mathbf{x}^\star, \theta^\star)}{\pi(\mathcal{D}|\mathbf{x}, \theta)} \frac{\pi(\mathbf{x}|\mathcal{D}, \theta)}{\pi(\mathbf{x}^\star|\mathcal{D}, \theta^\star)}$$
$$= \frac{\pi(\theta^\star)}{\pi(\theta)} \frac{\pi(\mathcal{D}|\theta^\star)}{\pi(\mathcal{D}|\theta)} \frac{f(\theta|\theta^\star)}{f(\theta^\star|\theta)}$$

  - This joint scheme reduces down to the marginal scheme (Chib (1995)), but will be intractable for complex models...

Particle MCMC (pMCMC)

- Of the various alternatives, pMCMC is the only obvious practical option for constructing global likelihood-free MCMC algorithms which are exact (Andreiu et al, 2010)
- Start by considering a basic marginal MH MCMC scheme with target $\pi(\theta|\mathcal{D})$ and proposal $f(\theta^\star|\theta)$ — the acceptance probability is $\min\{1, A\}$ where

$$A = \frac{\pi(\theta^\star)}{\pi(\theta)} \times \frac{f(\theta|\theta^\star)}{f(\theta^\star|\theta)} \times \frac{\pi(\mathcal{D}|\theta^\star)}{\pi(\mathcal{D}|\theta)}$$

- We can't evaluate the final terms, but if we had a way to construct a Monte Carlo estimate of the likelihood, $\hat{\pi}(\mathcal{D}|\theta)$, we could just plug this in and hope for the best:

$$A = \frac{\pi(\theta^\star)}{\pi(\theta)} \times \frac{f(\theta|\theta^\star)}{f(\theta^\star|\theta)} \times \frac{\hat{\pi}(\mathcal{D}|\theta^\star)}{\hat{\pi}(\mathcal{D}|\theta)}$$

"Exact approximate" MCMC (the pseudo-marginal approach)

- Remarkably, provided only that $\mathrm{E}[\hat{\pi}(\mathcal{D}|\theta)] = \pi(\mathcal{D}|\theta)$, the stationary distribution of the Markov chain will be exactly correct (Beaumont, 2003, Andreiu & Roberts, 2009)
- Putting $W = \hat{\pi}(\mathcal{D}|\theta)/\pi(\mathcal{D}|\theta)$ and augmenting the state space of the chain to include W, we find that the target of the chain must be

$$\propto \pi(\theta)\hat{\pi}(\mathcal{D}|\theta)\pi(w|\theta) \propto \pi(\theta|\mathcal{D})w\pi(w|\theta)$$

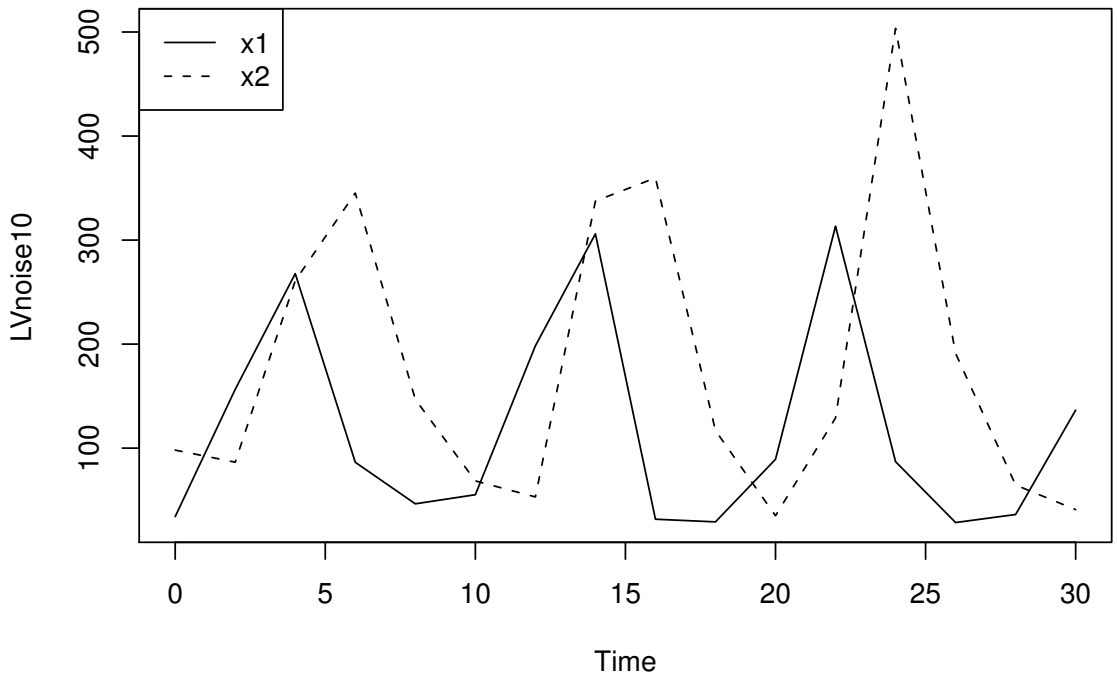
  and so then the above "unbiasedness" property implies that $\mathrm{E}(W|\theta) = 1$, which guarantees that the marginal for θ is exactly $\pi(\theta|\mathcal{D})$
- Blog post: http://tinyurl.com/6ex4xqw

Particle marginal Metropolis-Hastings (PMMH)

- Likelihood estimates constructed via importance sampling typically have this "unbiasedness" property, as do estimates constructed using a particle filter
- If a particle filter is used to construct the Monte Carlo estimate of likelihood to plug in to the acceptance probability, we get (a simple version of) the particle Marginal Metropolis Hastings (PMMH) pMCMC algorithm
- The full PMMH algorithm also uses the particle filter to construct a proposal for $\mathbf{x}$, and has target $\pi(\theta, \mathbf{x}|\mathcal{D})$ — not just $\pi(\theta|\mathcal{D})$
- The (bootstrap) particle filter relies only on the ability to forward simulate from the process, and hence the entire procedure is "likelihood-free"

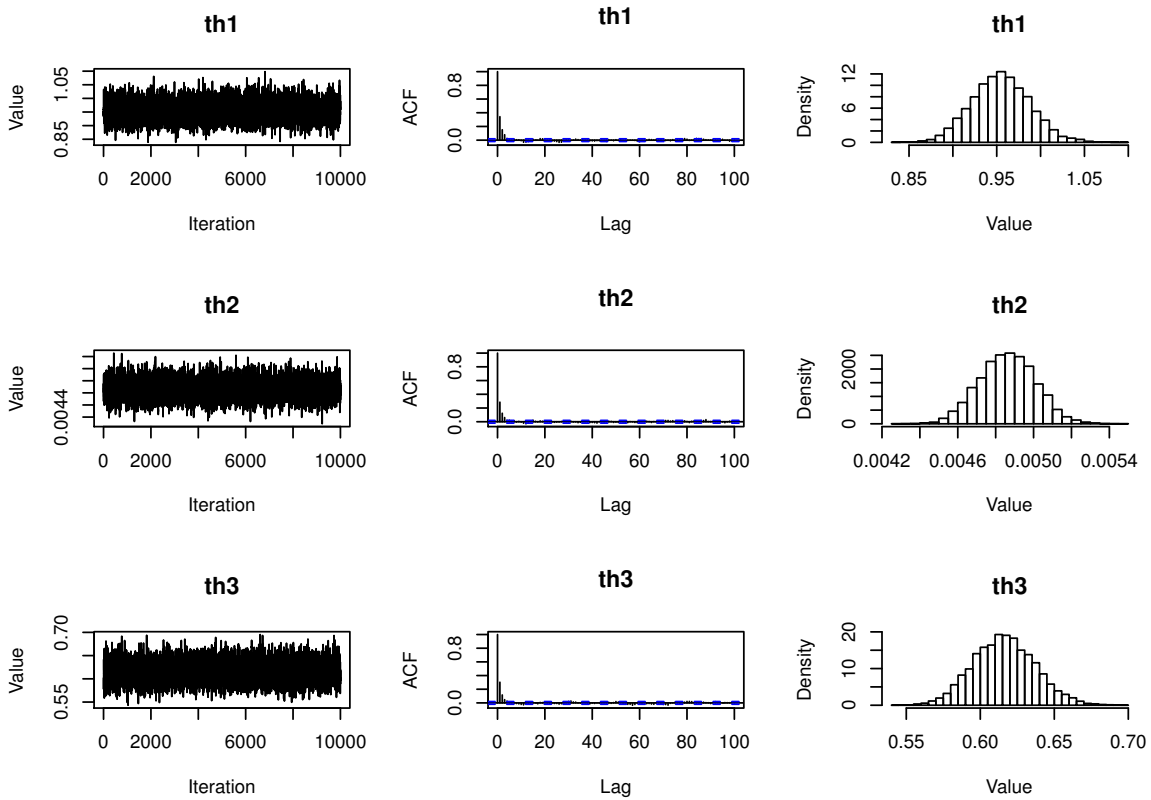
Blog post: http://bit.ly/kvznmq

Test problem: Lotka-Volterra model



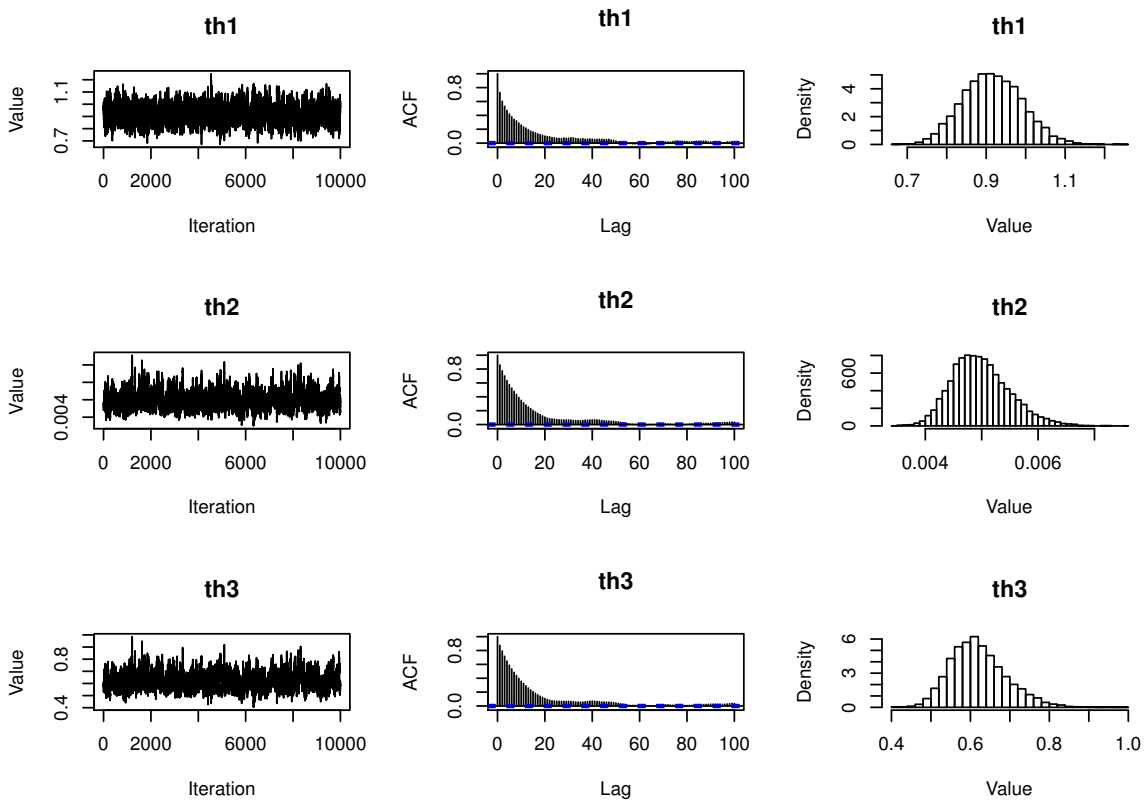
Simulated time series data set consisting of 16 equally spaced observations subject to Gaussian measurement error with a standard deviation of 10.

Marginal posteriors for the Lotka-Volterra model



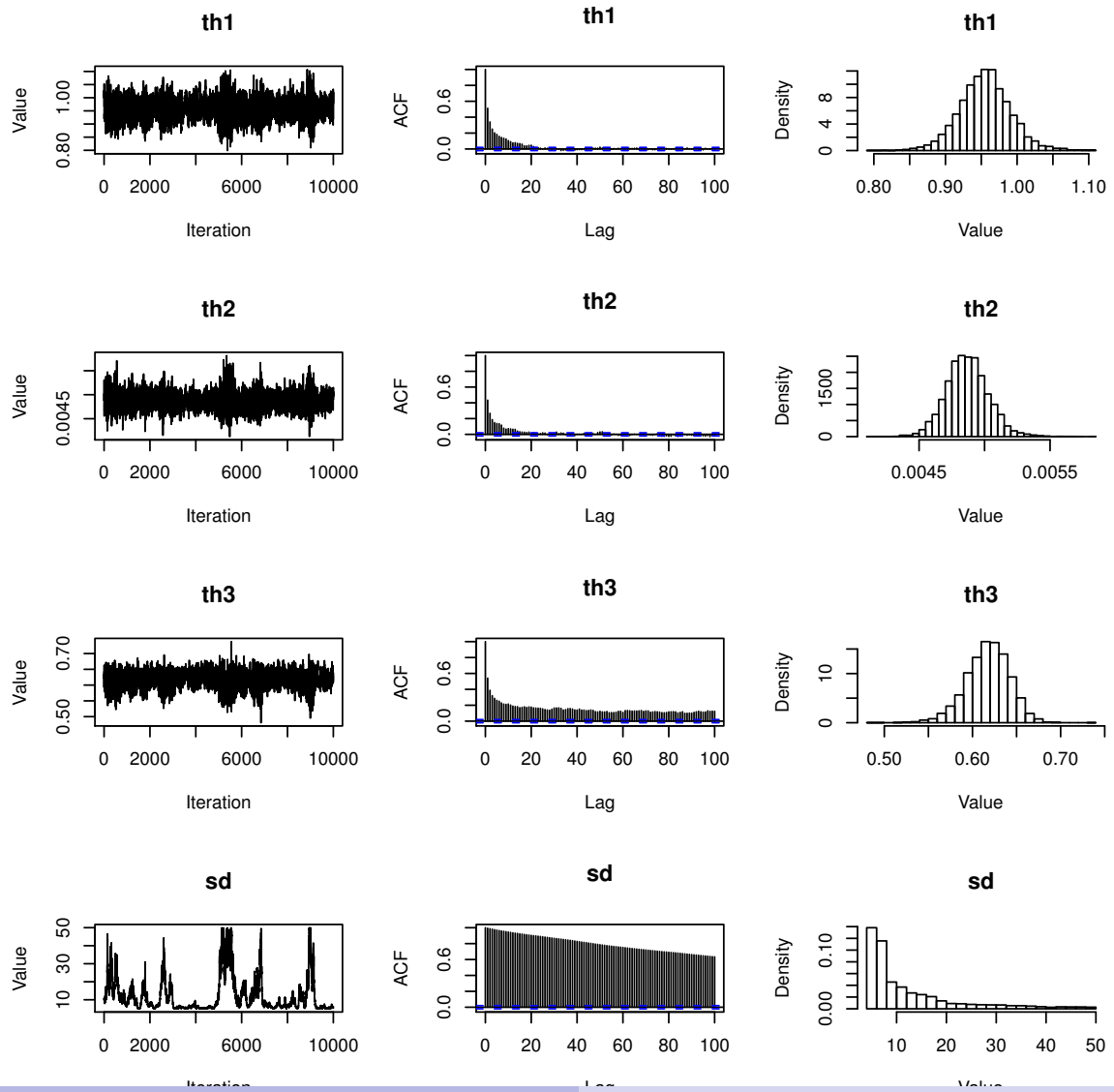
Note that the true parameters, $\theta = (1, 0.005, 0.6)$ are well identified by the data

Marginal posteriors observing only prey



Note that the mixing of the MCMC sampler is reasonable, and that the true parameters, $\theta = (1, 0.005, 0.6)$ are quite well identified by the data

Marginal posteriors for unknown measurement error



R package: smfsb

- Free, open source, well-documented software package for R, `smfsb`, associated with the forthcoming second edition of "Stochastic modelling for systems biology"
- Code for stochastic simulation and of (biochemical) reaction networks (Markov jump processes and chemical Langevin), and pMCMC-based Bayesian inference for POMP models
- Full installation and "getting started" instructions at `http://tinyurl.com/smfsb2e`
- Once the package is installed and loaded, running `demo("PMCMC")` at the R prompt will run a PMMH algorithm for the Lotka-Volterra model discussed here

Hitting the data...

- The above algorithm works well in many cases, and is extremely general (works for any Markov process)
- In the case of no measurement error, the probability of hitting the data (and accepting the proposal) is very small (possibly zero), and so the mixing of the MCMC scheme is very poor
- ABC (approximate Bayesian computation) strategy is to accept if

$$\|x^{\star}_{t+1} - d_{t+1}\| < \varepsilon$$

  but this forces a trade-off between accuracy and efficiency which can be unpleasant (cf. noisy ABC)
- Same problem in the case of low measurement error
- Particularly problematic in the context of high-dimensional data
- Would like a strategy which copes better in this case

The chemical Langevin equation (CLE)

- The CLE is a diffusion approximation to the true Markov jump process
- Start with the time change representation

$$X_t - X_0 = S\,N\left(\int_0^t h(X_\tau, c)d\tau\right)$$

and approximate $N_i(t) \simeq t + W_i(t)$, where $W_i(t)$ is an independent Wiener process for each i

- Substituting in and using a little stochastic calculus gives:

The CLE as an Itô SDE:

$$dX_t = Sh(X_t, c)\,dt + \sqrt{S\,\mathrm{diag}\{h(X_t, c)\}S'}\,dW_t$$

Improved particle filters for SDEs

- The "bootstrap" particle filter uses blind forward simulation from the model
- If we are able to evaluate the "likelihood" of sample paths, we can use other proposals
- The particle filter weights then depend on the Radon-Nikodym derivative of law of the proposed path wrt the true conditioned process
- For SDEs, the weight will degenerate unless the proposed process is absolutely continuous wrt the true conditioned process
- Ideally we would like to sample from $\pi(\mathbf{x}^\star_{t+1}|c^\star, x^\star_t, d_{t+1})$, but this is not tractable for nonlinear SDEs such as the CLE

Modified diffusion bridge (MDB)

- Need a tractable process $q(\mathbf{x}^\star_{t+1}|c^\star, x^\star_t, d_{t+1})$ that is locally equivalent to $\pi(\mathbf{x}^\star_{t+1}|c^\star, x^\star_t, d_{t+1})$
- Diffusion $dX_t = \mu(X_t)dt + \beta(X_t)^{\frac{1}{2}}dW_t$
- The nonlinear diffusion bridge

$$dX_t = \frac{x_1 - X_t}{1-t}dt + \beta(X_t)^{\frac{1}{2}}dW_t$$

  hits x_1 at $t = 1$, yet is locally equivalent to the true diffusion as it has the same diffusion coefficient
- This forms the basis of an efficient proposal; see Durham & Gallant (2002), Chib, Pitt & Shephard (2004), Delyon & Hu (2006), and Stramer & Yan (2007) for technical details

Summary

- POMP models form a large, important and interesting class of models, with many applications
- It is possible, and often desirable, to develop inferential algorithms which are "likelihood free" or "plug-and-play", as this allows the separation of the modelling from the inferential algorithm, allowing more rapid model exploration
- Many likelihood free approaches are possible, including sequential LF-MCMC, PMMH (pMCMC), (sequential) ABC for Bayesian inference and iterative filtering for maximum likelihood estimation
- Much work needs to be done to properly understand the strengths and weaknesses of these competing approaches


Golightly, A. and D. J. Wilkinson (2008) Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statistics and Data Analysis,* **52**(3), 1674–1693.


Golightly, A. and D. J. Wilkinson (2011) Bayesian parameter inference for stochastic biochemical network models using particle MCMC. *In submission*.


Wilkinson, D. J. (2009) Stochastic modelling for quantitative description of heterogeneous biological systems, *Nature Reviews Genetics.* **10**(2):122-133.


Wilkinson, D. J. (2010) Parameter inference for stochastic kinetic models of bacterial gene regulation: a Bayesian approach to systems biology (with discussion), in J.-M. Bernardo et al (eds) *Bayesian Statistics 9,* OUP, in press.


Wilkinson, D. J. (2011) *Stochastic Modelling for Systems Biology, second edition*. Chapman & Hall/CRC Press, in press.

Contact details...

email: darren.wilkinson@ncl.ac.uk
www: tinyurl.com/darrenjw