Bayesian inference of deterministic population growth models

Luiz Max F. de Carvalho* [lmax.procc@gmail.com]

Claudio J. Struchiner [stru@fiocruz.br]

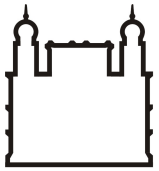
Leonardo S. Bastos [lsbastos@fiocruz.br]

Scientific Computing Programme (PROCC), Oswaldo Cruz Foundation (Fiocruz), Rio de Janeiro, Brazil

March, 2014
12th EBEB - Atibaia – SP

Nice to meet you!

- BSc. in Microbiology, UFRJ (2013);
- Statistics Assistant, Pan American Health Organization, 2010-2013;
- Currently at PROCC and DME/IM-UFRJ (MSc);
- Soon to be moving to the University of Edinburgh for a PhD in Evolutionary Biology.



Motivation

- Deterministic models are widely used in Science, let alone Biology;
  - Population Growth;
  - Disease Spreading;
  - Cell and molecular interactions.
- They provide a crude but easily interpretable representation of reality;
- Temperature is a key factor to the growth of several organisms.
  - Disease-carrying arthropds;
  - Pathogenic bacteria;
  - Economically important plants.
- With a deterministic model and some time series data at hand, **how to learn about model parameters**?

Background

- Consider a deterministic model $M(\cdot)$;
  - Let $x \in \mathcal{X} \subset \mathbb{R}^p$ be the set of model inputs and $y \in \mathcal{Y} \subset \mathbb{R}^n$ be the model outputs. The deterministic model $M(x; \boldsymbol{\theta}) = y$, where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^q$ is a q-dimensional parameter vector, completely specifies the relationship between x and y (Poole & Raftery, 2000);
  - In our particular case, we have laid our dirty hands on some data $\mathbf{y}$ and inputs $\mathbf{x}$ that we think can be modelled as $\mathbf{y} = M(\mathbf{x}; \boldsymbol{\theta})$
- We are now interested in learning about $\boldsymbol{\theta}$

Temperature-dependent Population Growth

- Consider the ordinary non-linear differential equation (Verhulst, 1838):

$$\frac{dP}{dt} = r\left(1 - \frac{P}{K}\right)P \quad \therefore \quad P(t) = \frac{K}{1 + \left(\frac{K - N_0}{N_0}\right)e^{-rt}} \tag{1}$$

- We formulate a modified version of (1), with temperature-dependent parameters

$$P(t, T) = \frac{K(T)}{1 + \left(\frac{K(T) - N_0}{N_0}\right)e^{-r(T)t}} \tag{2}$$

Temperature-dependent Population Growth (cont.)

- To complete model specification, we propose two smooth functions on temperature T:

$$K(T) = c_K exp\left(-\frac{(T - a_K)^2}{b_K}\right) \tag{3}$$

$$r(T) = c_r exp\left(-\frac{(T - a_r)^2}{b_r}\right) \tag{4}$$

We want to learn about $\boldsymbol{\theta} = \{a_K, b_K, c_K, a_r, b_r, c_r\}$

Likelihood

- Assume $P(t, T)$ to be a Gaussian process with fixed variance τ^2;
- Let $\mathbf{y} = \{y_1, y_2, \ldots, y_N\}$ be an output vector with N measurements, which we observe directly;
- Moreover, let $\mathbf{t} = \{t_1, t_2, ..., t_N\}$ and $\mathbf{T} = \{T_1, T_2, ..., T_N\}$ be the vectors with observed times and temperatures. Then

$$y_i | t_i, T_i, N_0, \boldsymbol{\theta} \sim \mathcal{N}(\mu(t_i, T_i, N_0; \boldsymbol{\theta}), \tau^2) \tag{5}$$

$$\mu(t_i, T_i, \boldsymbol{\theta}) = \frac{K(T_i; \boldsymbol{\theta}_K)}{1 + \left(\frac{K(T_i; \boldsymbol{\theta}_K) - N_0}{N_0}\right) e^{-r(T_i; \boldsymbol{\theta}_r) t_i}}, \qquad \forall i = 1, 2, \ldots, N \tag{6}$$

which is equivalent to writing $y_i = M(t_i, T_i, N_0; \boldsymbol{\theta}) + \epsilon$, $\epsilon \sim \mathcal{N}(0, \tau^2)$.

Priors

- Biologically motivated, **proper** priors, elicited to maintain functional form while remaining diffuse.

$$
\begin{aligned}
a_K, a_r &\sim Normal(20, 10) \\
b_K, b_r &\sim Gamma(4, 1/5) \\
c_K &\sim Gamma(1, 1/1000) \\
c_r &\sim Normal(1/2, 2) \\
\tau^2 &\sim Gamma(1/10, 1/10)
\end{aligned}
$$

Posterior

- From the Bayes theorem

$$p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{T}) \propto p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{t}, \boldsymbol{T})\pi(\boldsymbol{\theta}|\boldsymbol{t}, \boldsymbol{T}) \tag{7}$$

- The model for $P(t, T)$ is thus hierarchical and depends on two latent quantities, $r(T)$ and $K(T)$.

Posterior Approximation – Stan

- Hamiltonian Monte Carlo (HMC):
  - Avoids Random Walk behaviour;
  - Allows big moves, with high acceptance probability;
  - Does not suffer with highly correlated posteriors;
- We used the **stan** package of the R Statistical Computing Environment to approximate the posterior through HMC.
  - Fast C++ implementation;
  - No-U-Turn Sampler (NUTS);
  - Neat BUGS-like syntax for model specification;
  - Smooth interface with R.
- MCMC was run for $50,000$ iterations with $25,000$ burn-in and convergence was assessed by inspecting the trace- and autocorrelation plots and potential scale reduction factor.

Results I – Simulation study

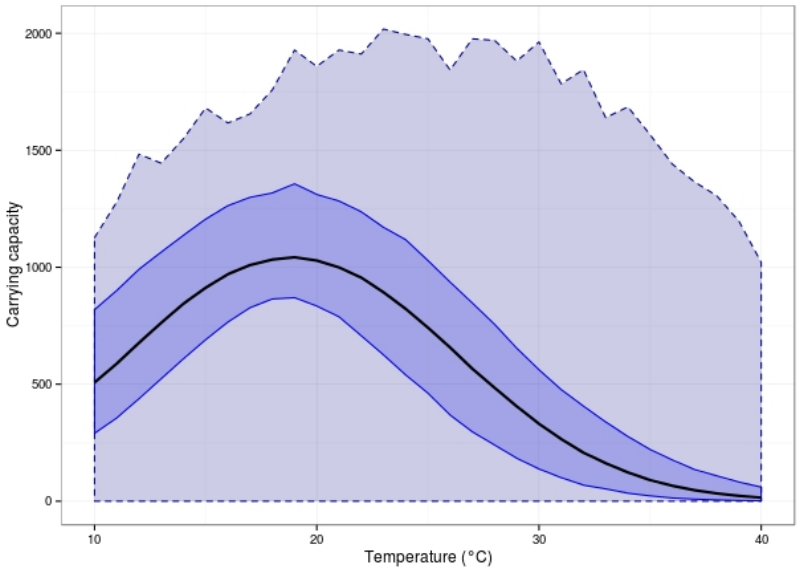
- From a set of parameters $\boldsymbol{\theta}^*$ and a grid $\{\boldsymbol{t}, \boldsymbol{T}\}$ we generate Q data sets of size $N = n_t \times n_T$ by sampling from $\boldsymbol{y}|\boldsymbol{\theta}^*, \boldsymbol{t}, \boldsymbol{T}$;
- We then obtain Q posterior estimates and calculate MSE, normalized bias and coverage probability of the 95% credibility intervals;

Parameter	Value	Posterior Mean	Bias	MSE	Coverage
a_K	30.00	29.44	0.01	7.99	0.93
a_r	23.00	22.71	0.00	3.31	0.86
b_K	10.00	13.08	0.95	450.26	0.94
b_r	15.00	16.82	0.22	16.77	0.88
c_K	700.00	692.17	0.09	7203.06	0.96
c_r	0.40	0.43	0.00	0.04	0.85
τ	3.16	4.89	0.94	67.02	0.88

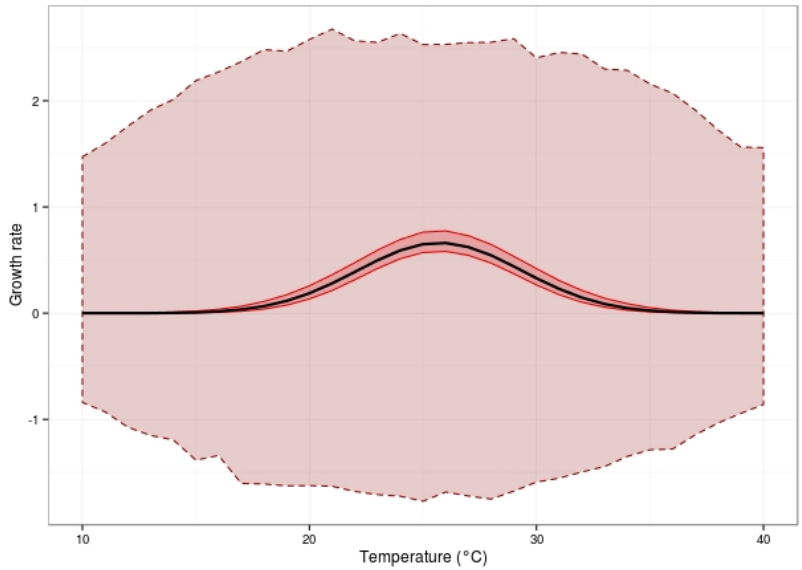
  - Consistent results for $N = 350, 630$ and 1600.

Results II – *Rhodnius prolixus data*

- Important Chagas disease vector;
  - We fear climatic change may increase suitability in previously uncolonized areas;
- Population sizes measured in 10 temperatures in the range $16 - 38\,^{\circ}\mathrm{C}$ for 35 days ($N_0 = 30$ for all T);



(a) K(T)

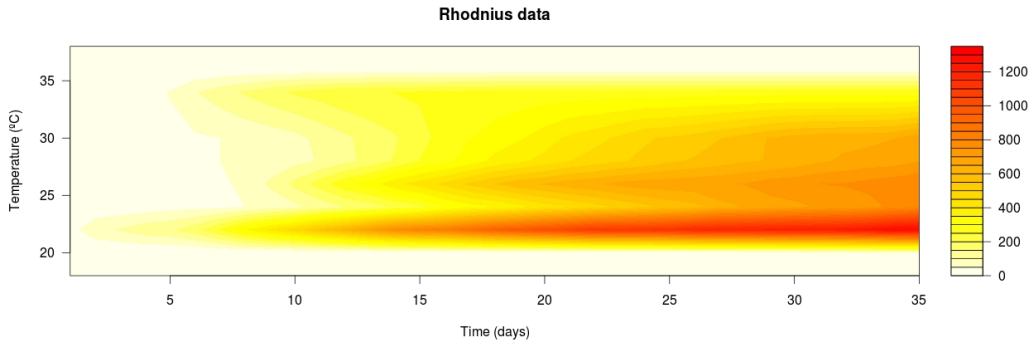


(b) r(T)

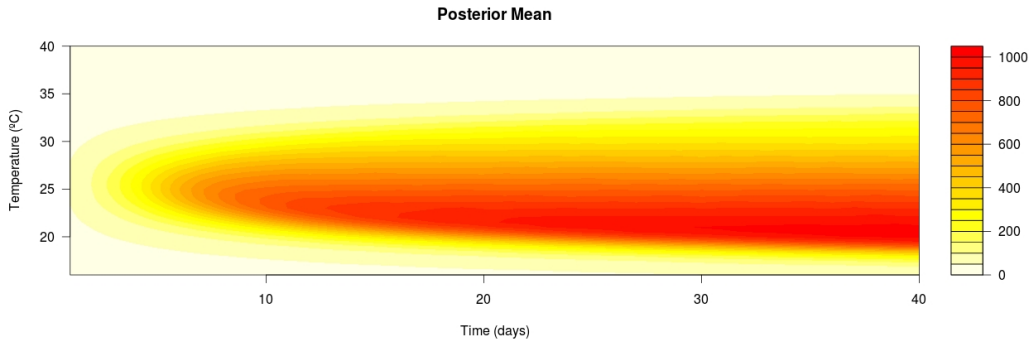
Results II – *Rhodnius prolixus data* (cont.)

	Posterior Mean (95% C.I.)	Prior Mean (95% C.I.)
a_K	19.23 (17.56 – 21.09)	25.00 (5.40 – 44.60)
a_r	25.73 (25.44 – 26.10)	25.00 (5.40 – 44.60)
b_K	106.17 (75.25 – 137.31)	20.00 (5.44 – 43.84)
b_r	26.77 (22.59 – 32.19)	20.00 (5.44 – 43.84)
c_K	1023.32 (898.28 – 1165.40)	1000.00 (25.31 – 3688.87)
c_r	0.66 (0.58 – 0.76)	0.50 (-3.41 – 4.41)
τ	177.33 (166.10 – 191.78)	1.00 (0.00 – 9.78)

Results II – *Rhodnius prolixus data* (cont.)



(c) Data



(d) Posterior

Conclusions and Perspectives

- We stress the importance of using Bayesian Inference to learn about model parameters
  - Parameters retain direct interpretability
- Stan
  - Efficient sampling through HMC;
  - NUTS drops the need for hand-tuning;
  - Consirable speed-up and quicker convergence.
- Perspectives
  - Dynamic variance, $\tau^2(t)$;
  - Other data sets, e.g, bacterial growth;
  - Complete treatment of uncertainty: Bayesian melding.

Thank you!

- References

 D. Poole and A. E. Raftery, "Inference for deterministic simulation models: the bayesian melding approach," *Journal of the American Statistical Association*, vol. 95, no. 452, pp. 1244–1255, 2000.

 P.-F. Verhulst, "Notice sur la loi que la population suit dans son accroissement. correspondance mathématique et physique publiée par a," *Quetelet*, vol. 10, pp. 113–121, 1838.

- Acknowledgements

  - My advisors, Claudio and Leo;
  - Leonardo B. Santos, INPE;
  - PROCC, who provided an office and **a Gourmet coffee machine**!

Questions