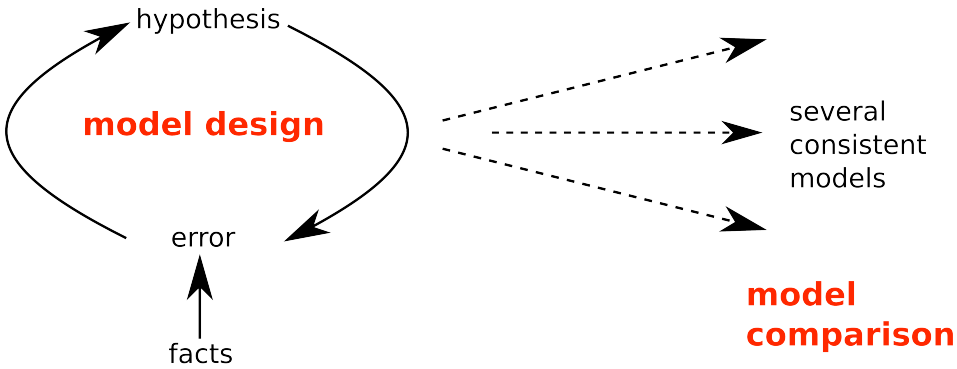
Using ABC for Model Design and Inference across Biological Scales

Oliver Ratmann (Duke Biology)

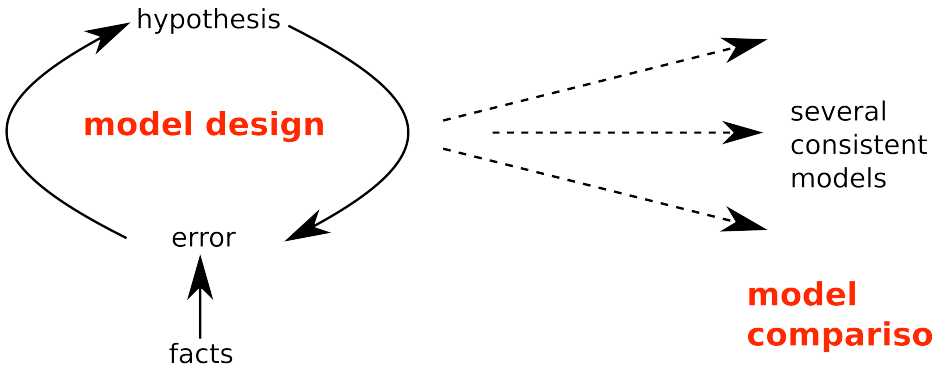
Overview

- Two aspects of model choice;
  - I will focus on methods aiding model design with ABC
  - Combine multiple lines of evidence easily with ABC
  - to drive model design



Overview

- Two aspects of model choice;
  - I will focus on methods aiding model design with ABC
  - Combine multiple lines of evidence easily with ABC to drive model design

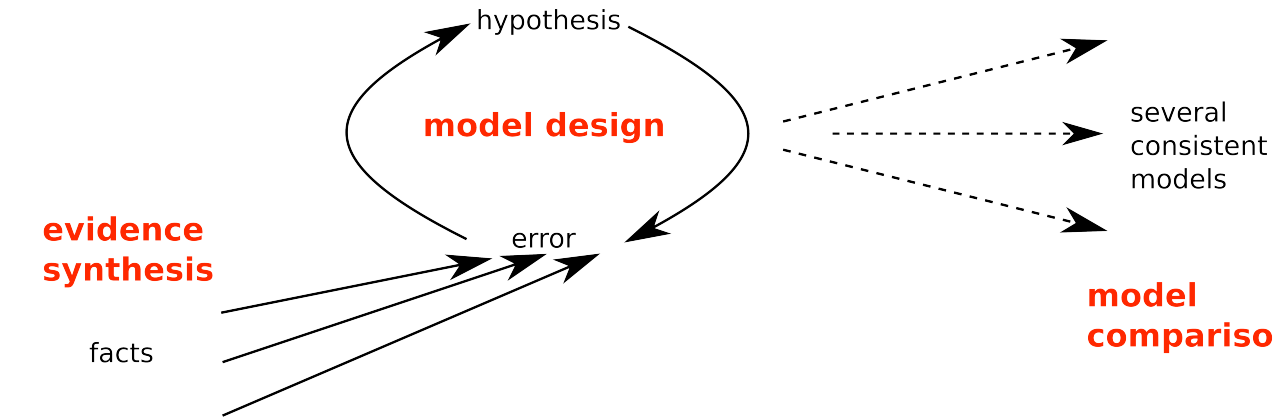


Bayesian error diagnostics
[Box JRSSA'80, Meng AnnStat '94, Ratmann PNAS '09]

Bayes factors, Post Model Probs
[Kass JASA '95, Pritchard MBE'99, Fagundes PNAS'07, Grelaud BA'09, Toni JRSI'09]

Overview

- Two aspects of model choice;
  - I will focus on methods aiding model design with ABC
- Combine multiple lines of evidence easily with ABC
  - to drive model design



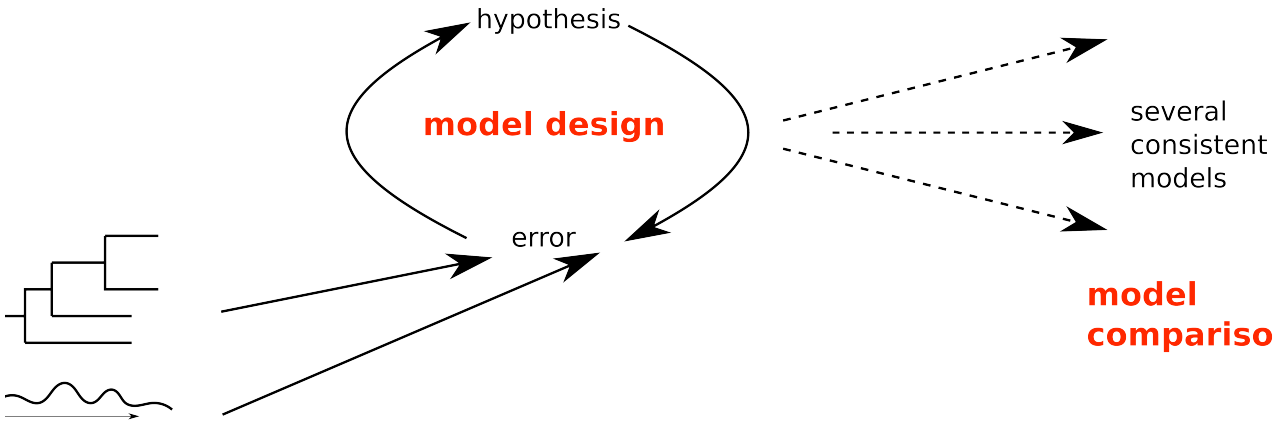
Bayesian evidence synthesis
[Ades JRSSA'06]

Bayesian error diagnostics
[Box JRSSA'80, Meng AnnStat '94, Ratmann PNAS '09]

Bayes factors, Post Model Probs
[Kass JASA '95, Pritchard MBE'99, Fagundes PNAS'07, Grelaud BA'09, Toni JRSI'09]

Overview

- Two aspects of model choice;
  - I will focus on methods aiding model design with ABC
- Combine multiple lines of evidence easily with ABC
  - to drive model design



Bayesian error diagnostics
[Box JRSSA'80, Meng AnnStat '94,
Ratmann PNAS '09]

Bayes factors, Post Model Probs
[Kass JASA '95,
Pritchard MBE'99, Fagundes PNAS'07,
Grelaud BA'09, Toni JRSI'09]

• MODEL DESIGN WITH DIAGNOSTIC ABC ERRORS •

APPLICATION TO INFLUENZA EVOLUTION & ECOLOGY

ABC: measure theoretic construction

$f : \mathcal{B}_{\mathcal{X}} \to \mathbb{R}_0^+, \quad dx \to f(dx|\theta)$ data-generating stochastic process

$\xi_{x_0} : \mathcal{X} \to \mathbb{R}^K, \quad \varepsilon_{1:K} = (\varepsilon_1, \ldots, \varepsilon_K), \quad \varepsilon_k = \rho_k\big(S_k(x), S_k(x_0)\big)$ ABC projection

induces the multi-dim error measure

$$\xi_{x_0,\theta}(E_1 \times \ldots \times E_K)$$

$$\stackrel{\text{def}}{=} f\Big(\xi_{x_0}^{-1}(E_1 \times \ldots \times E_K) \,\Big|\, \theta \Big) = \int \mathbb{1}\left\{x \in \xi_{x_0}^{-1}(E_1 \times \ldots \times E_K)\right\} \; f(dx|\theta)$$

We explain why $K > 1$ later.

ABC: data augmentation over error space $\Theta \times \mathbb{R}^K$

To circumvent likelihood computations for given θ,

1. simulate summary errors $\varepsilon_{1:K} \sim \xi_{x_0,\theta}$
2. weight $\varepsilon_{1:K}$ according to error magnitude

$$\pi_\tau(\theta, \varepsilon_{1:K}|x_0) \propto \quad \pi(\theta) \; \xi_{x_0,\theta}(\varepsilon_{1:K}) \; \prod_{k=1}^K \kappa_k(\varepsilon_k; \tau_k)$$

e.g. $\kappa_k(\varepsilon_k; \tau_k) = 1/\tau_k \; \mathbb{1}\{|\varepsilon_k| \le \tau_k/2)$

ABC: measure theoretic construction

$f : \mathcal{B}_{\mathcal{X}} \to \mathbb{R}_0^+, \quad dx \to f(dx|\theta)$ data-generating stochastic process

$x_0 : \mathcal{X} \to \mathbb{R}^K, \quad \varepsilon_{1:K} = (\varepsilon_1, \ldots, \varepsilon_K), \quad \varepsilon_k = \rho_k\big(S_k(x), S_k(x_0)\big)$ ABC projection

induces the multi-dim error measure

$$\xi_{x_0,\theta}(E_1 \times \ldots \times E_K)$$

$$\stackrel{\text{def}}{=} f\Big(\xi_{x_0}^{-1}(E_1 \times \ldots \times E_K) \,\Big|\, \theta \Big) = \int \mathbb{1}\left\{ x \in \xi_{x_0}^{-1}(E_1 \times \ldots \times E_K) \right\} \, f(dx|\theta)$$

We explain why $K > 1$ later.

ABC: data augmentation over error space $\Theta \times \mathbb{R}^K$

To circumvent likelihood computations for given θ,

1. simulate summary errors $\varepsilon_{1:K} \sim \xi_{x_0,\theta}$
2. weight $\varepsilon_{1:K}$ according to error magnitude

$$\pi_\tau(\theta, \varepsilon_{1:K}|x_0) \propto \quad \pi(\theta) \; \xi_{x_0,\theta}(\varepsilon_{1:K}) \; \prod_{k=1}^K \kappa_k(\varepsilon_k; \tau_k)$$

e.g. $\kappa_k(\varepsilon_k; \tau_k) = 1/\tau_k \; \mathbb{1}\{|\varepsilon_k| \leq \tau_k/2)$

Easy to extend ABC algorithms to sample from $\pi_\tau(\theta, \varepsilon_{1:K}|x_0)$:

MCMC-ABC on $\Theta \times$ error space

MCMC1 If now at θ propose $\theta' \sim q(\theta \to \theta')$.

MCMC2 $x' \sim f(\cdot|\theta')$, compute errors $\varepsilon'_k = \rho_k\big(S_k(x'), S_k(x_0)\big)\ \forall k$.

MCMC3 Accept $(\theta', \varepsilon'_{1:K})$ with prob

$$mh(\theta, \varepsilon_{1:K}; \theta', \varepsilon'_{1:K}) = \min\left\{1\,,\, \frac{q(\theta' \to \theta)}{q(\theta \to \theta')} \times \frac{\pi(\theta')\ \prod_k \kappa_k(\varepsilon'_k; \tau_k)}{\pi(\theta)\ \prod_k \kappa_k(\varepsilon_k; \tau_k)}\right\},$$

o/w stay at $(\theta, \varepsilon_{1:K})$. Return to MCMC1.

Added value: model design via conflicting errors $\varepsilon_{1:K}$

Example:

x_0 is $n = 100$ samples of iid Exp(1/5). Assume $x_{0i} \sim \mathcal{N}(\mu, \sigma^2)$.
π_θ normal inverse gamma
s_k: sample mean and sample median
κ_k: std indicator kernel with $\tau_1 = \tau_2 = 1.6$

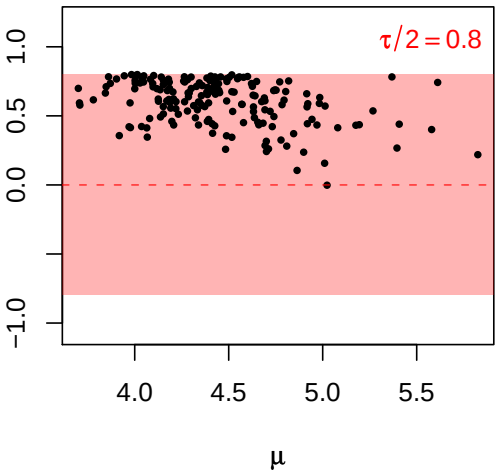
Choose hyperparameters so we can run simple rejection algorithm (10,000 iterations).

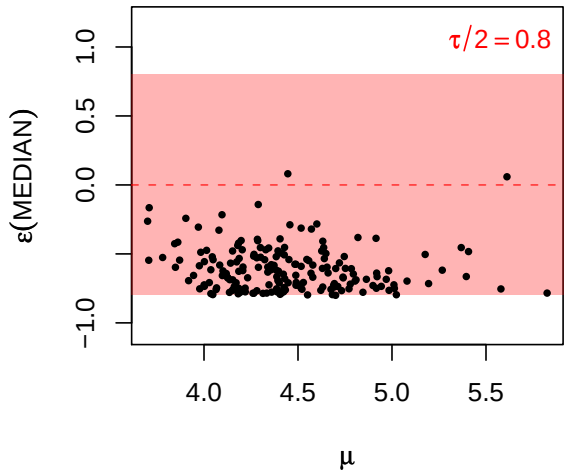
Added value: model design via conflicting errors $\varepsilon_{1:K}$

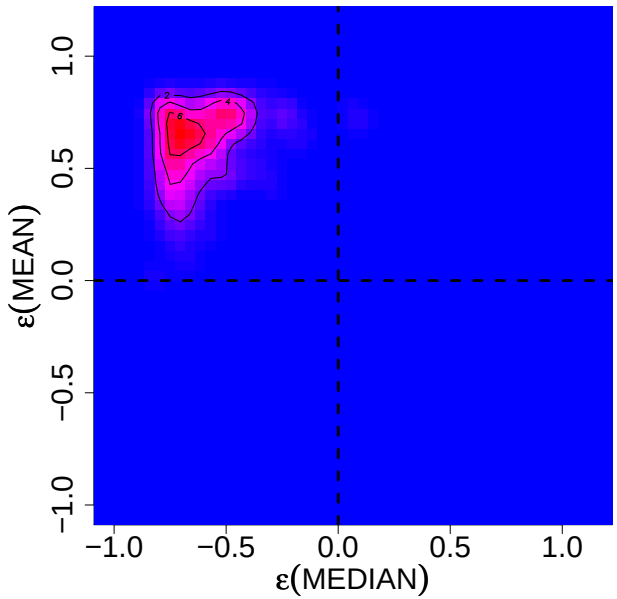
Example:

x_0 is $n = 100$ samples of iid Exp(1/5). Assume $x_{0i} \sim \mathcal{N}(\mu, \sigma^2)$.
π_θ normal inverse gamma
s_k: sample mean and sample median
κ_k: std indicator kernel with $\tau_1 = \tau_2 = 1.6$

Choose hyperparameters so we can run simple rejection algorithm (10,000 iterations).





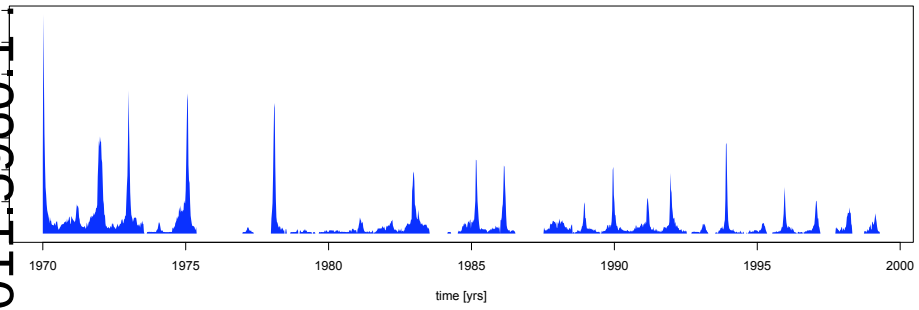


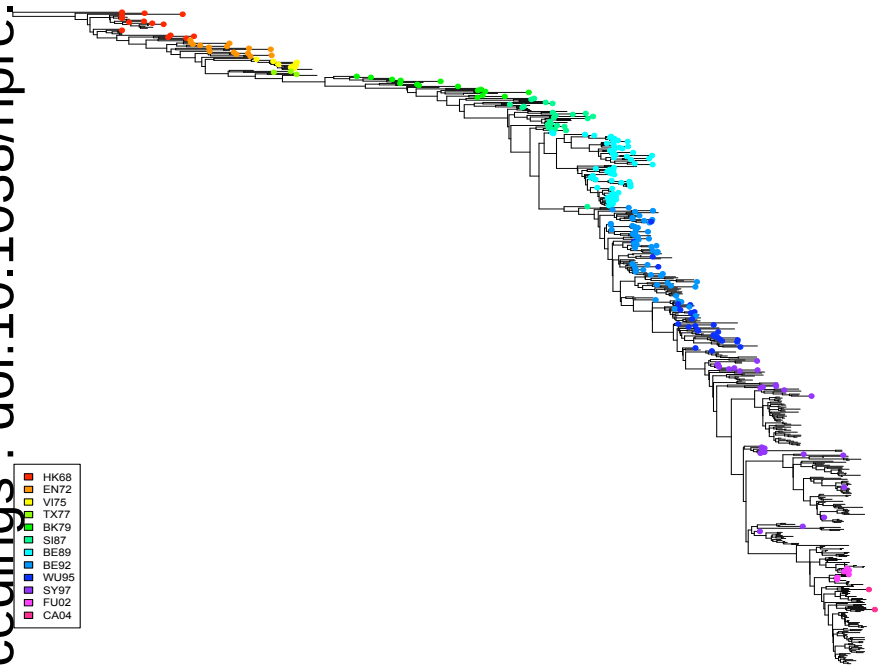
MODEL DESIGN WITH DIAGNOSTIC ABC ERRORS

• APPLICATION TO INFLUENZA EVOLUTION & ECOLOGY •

ABC to interface biological scales:

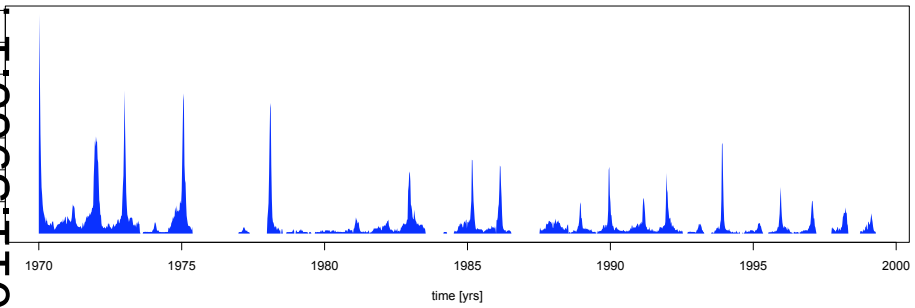
evolution & ecology of influenza A (H3N2)

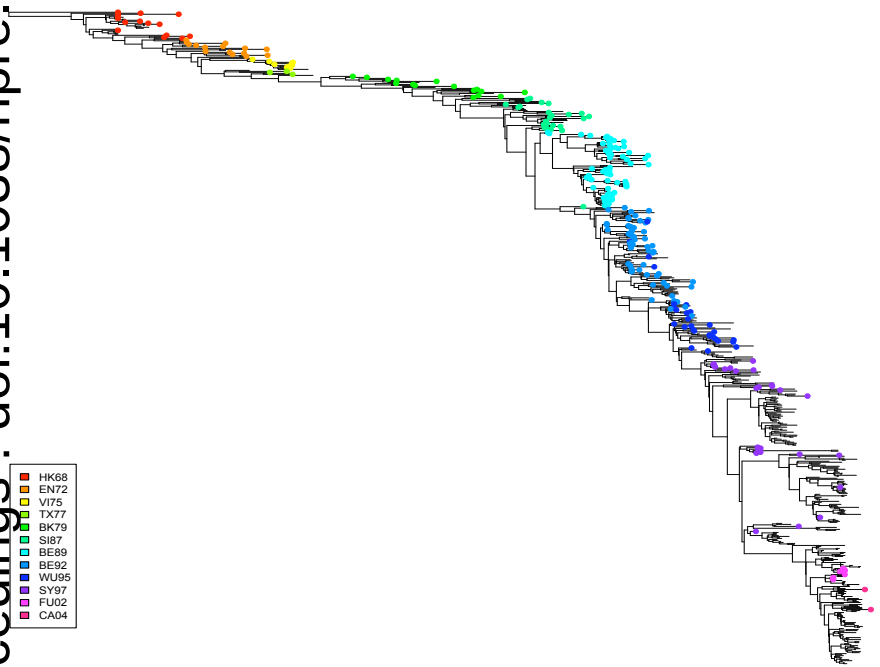


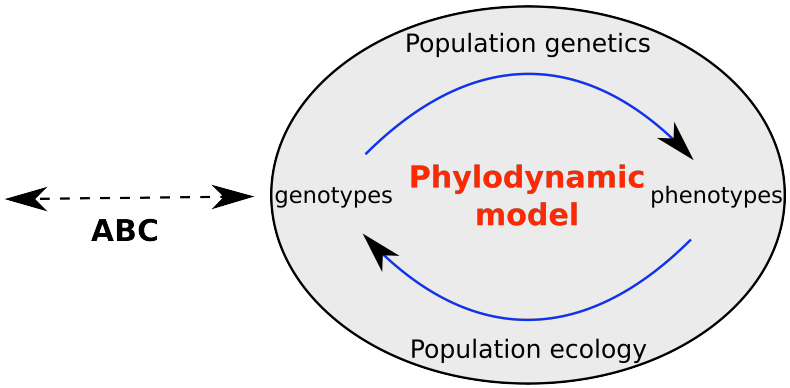


ABC to interface biological scales:

evolution & ecology of influenza A (H3N2)



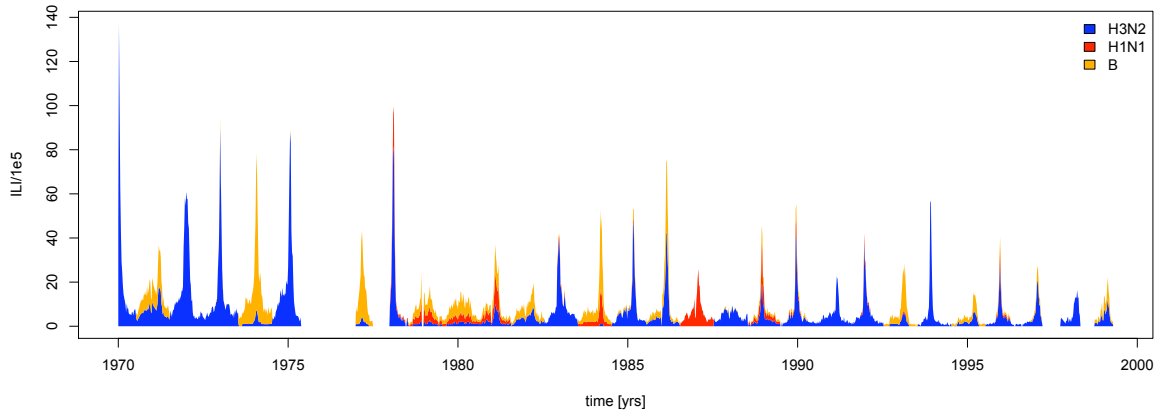




Choice of ABC summaries across scales

Summaries reflect characteristic features, eg

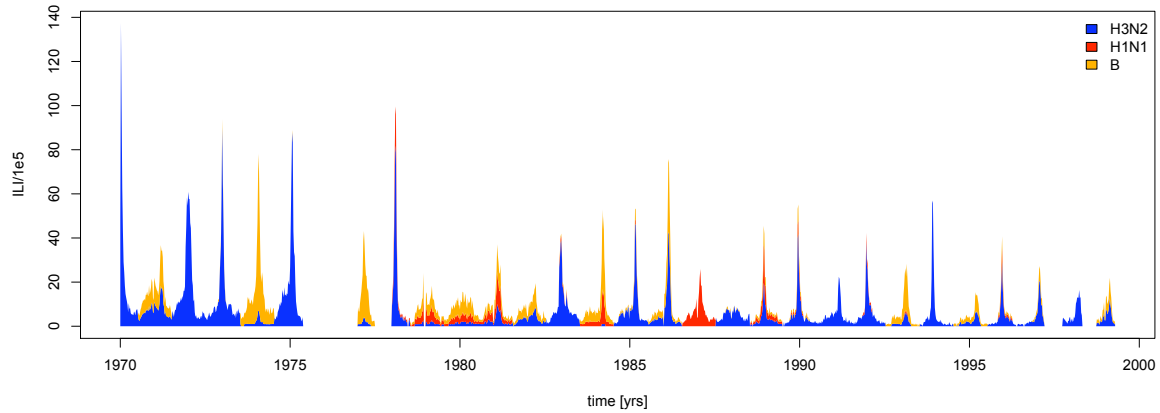
- interannual variability
- periodicity
- explosiveness
- overall magnitude



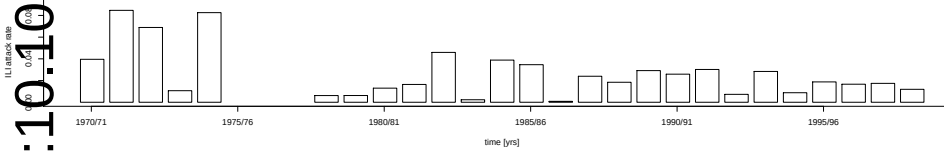
Choice of ABC summaries across scales

Summaries reflect characteristic features, eg

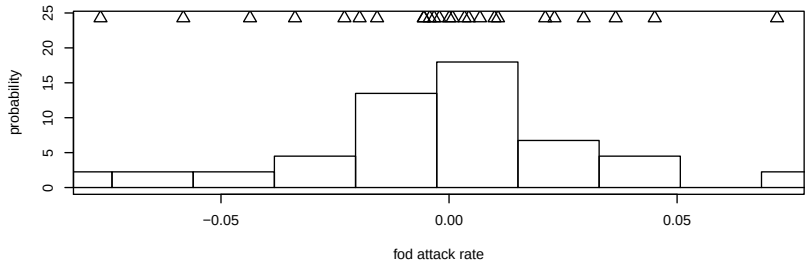
- interannual variability
- periodicity
- explosiveness
- overall magnitude



⇒ For example: distribution of differences in annual attack rate (Δ_y)



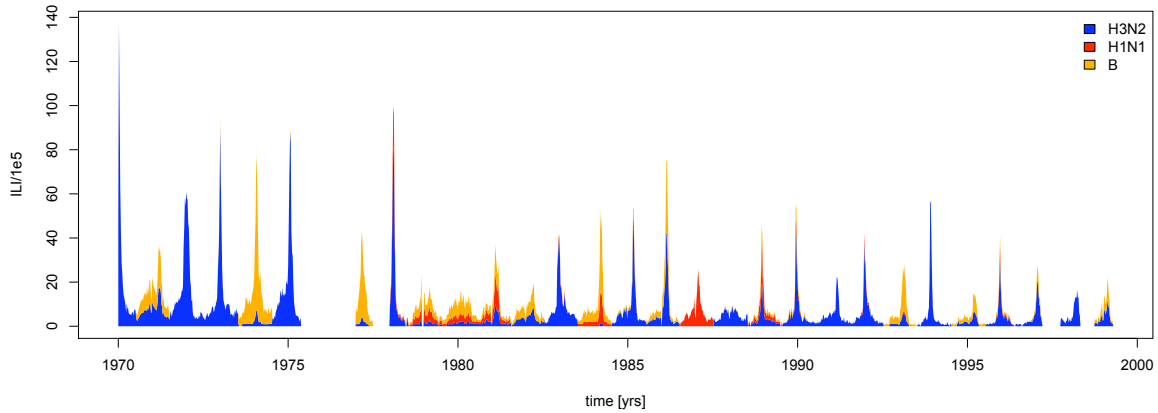
⇒

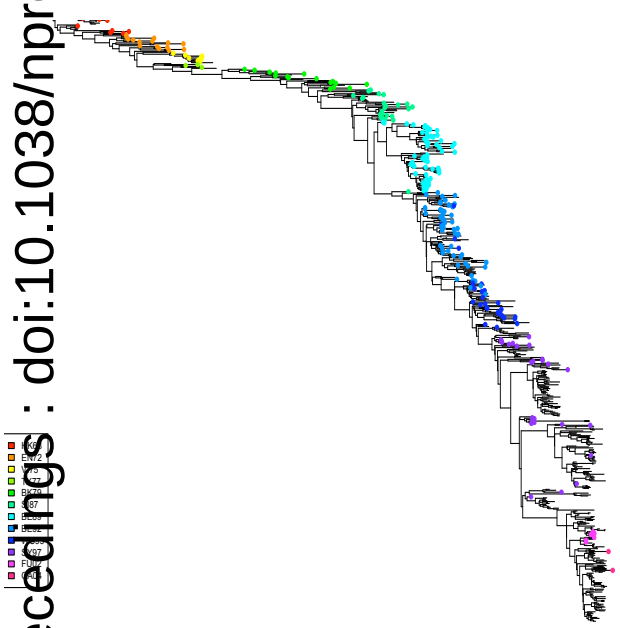


Choice of ABC summaries across scales

Summaries reflect characteristic features, eg

- interannual variability
- periodicity
- explosiveness
- overall magnitude

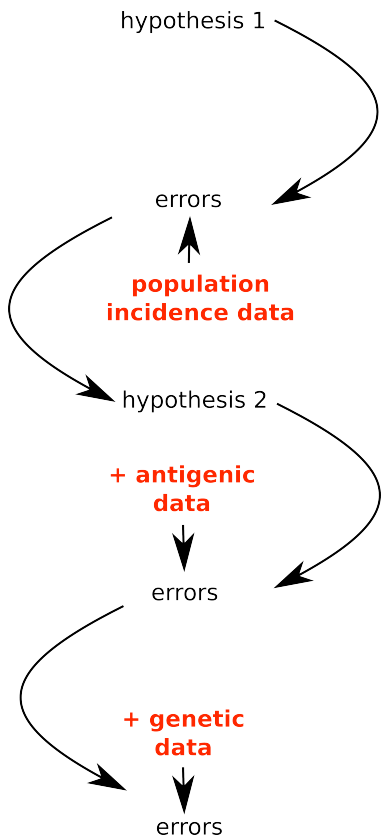




- Number of antigenic clusters (Smith et al 2004)
- pairwise diversity
- divergence of serially sampled strains to root

Combining multiple lines of evidence

can drive model design



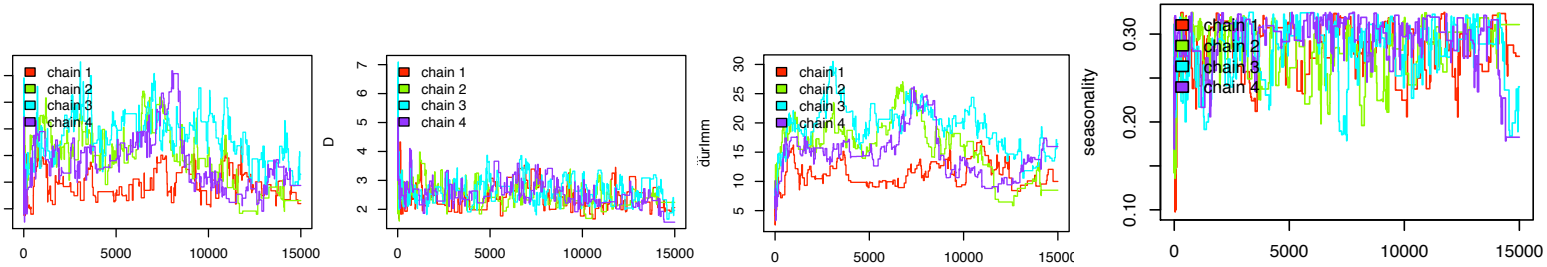
Scale: epidemiological data

- SIRS with sinusoidal seasonal forcing on transm pa
- MCMC

Scale: epidemiological data

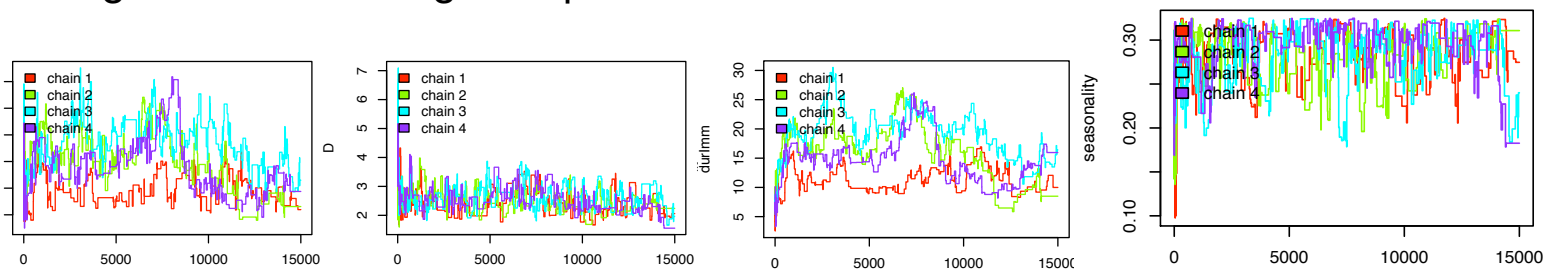
- SIRS with sinusoidal seasonal forcing on transm pa
- MCMC

- strong seasonal forcing to explain interannual seasonal variation

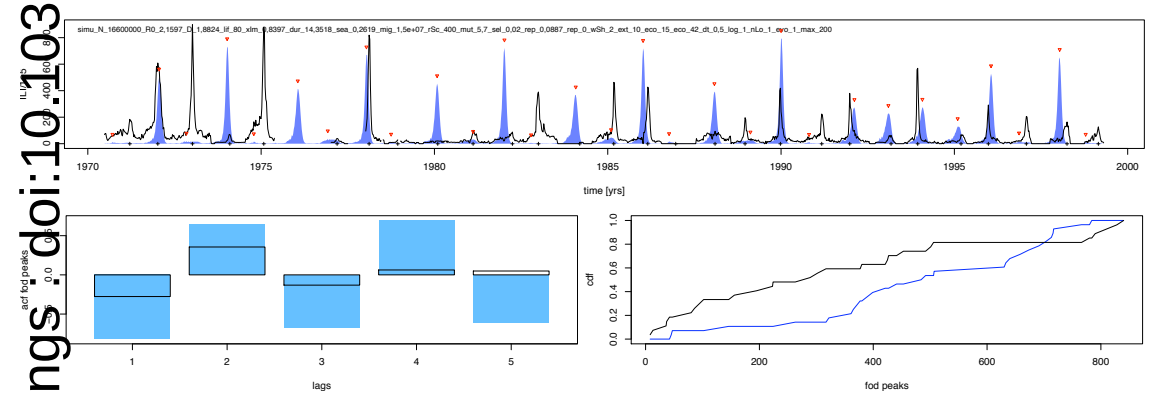


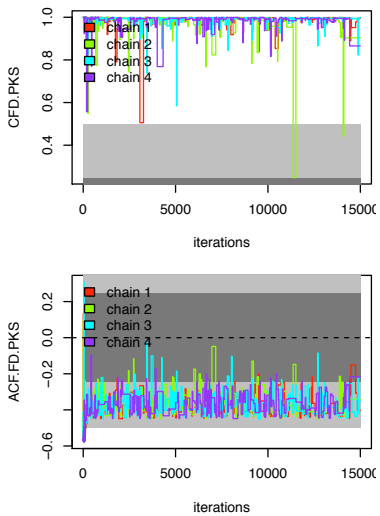
Scale: epidemiological data

- SIRS with sinusoidal seasonal forcing on transm pa
- MCMC
- strong seasonal forcing to explain interannual seasonal variation

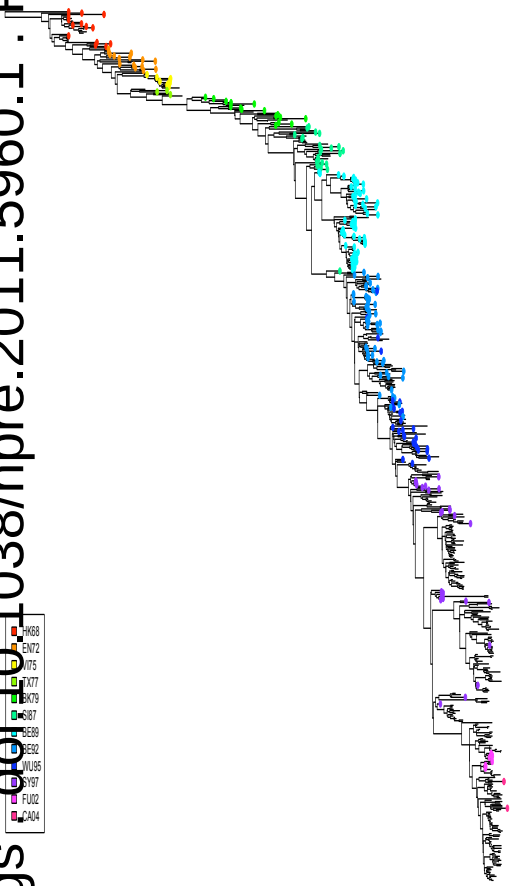


- too regular and too strong sustained oscillations





Phylodynamic multi-SIRS model (Koelle et al JRoySoc 2010)



- track status of infection with multiple phenot distinct variants

$$i = 1, \ldots, n: \quad \frac{dS_i}{dt} = \mu(N - S_i) - \beta_t \frac{S_i}{N} \sum_{j=1}^{n} \sigma_{ij} I_j + \gamma(N - S_i - I_i)$$

$$\frac{dI_i}{dt} = \beta_t \frac{S_i}{N} I_i - (\mu + \nu) I_i$$

- specify only tempo with which variants emerge

$$\frac{dI_i}{dt} = \beta_t \frac{S_i}{N} I_i - (\mu + \nu) I_i - h(\text{age}_i) I_i$$

$$h(a) = \kappa/\lambda \, (a/\lambda)^{\kappa - 1}$$

- simulate strains of each variant

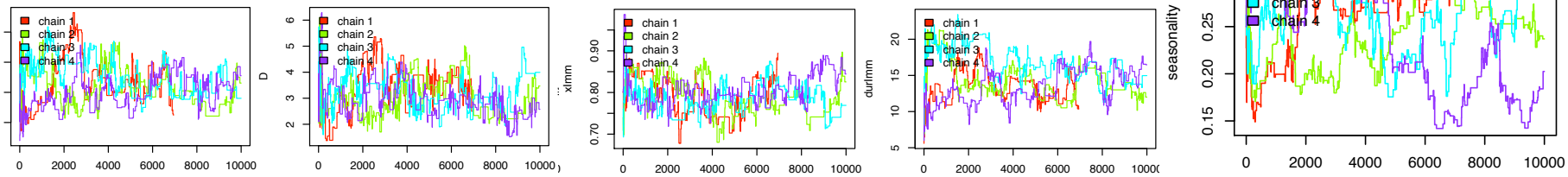
Scale: epidemiological & antigenic data

- phylodynamic multi-SIRS model
- MCMC

Scale: epidemiological & antigenic data

- phylodynamic multi-SIRS model
- MCMC

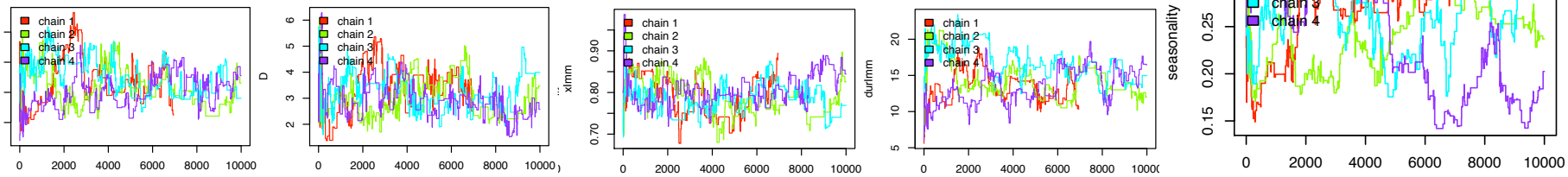
intrinsic dynamics contribute to overall seasonality



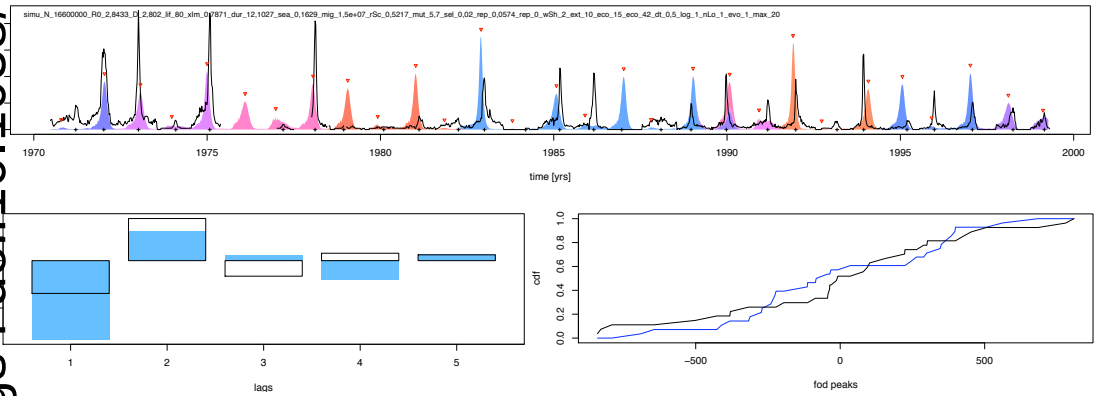
Scale: epidemiological & antigenic data

- phylodynamic multi-SIRS model
- MCMC

intrinsic dynamics contribute to overall seasonality



consistent with observed summaries



Scale: epidemiological & antigenic & genetic data

- phylodynamic multi-SIRS model; more θ:
  - nucl mut rate 5.7×10^{-3}/site/yr (fix);
  - population multiplier $e \sim [0, 250]$;
  - rel sel advantage $s \sim [0, 0.1]$
- MCMC

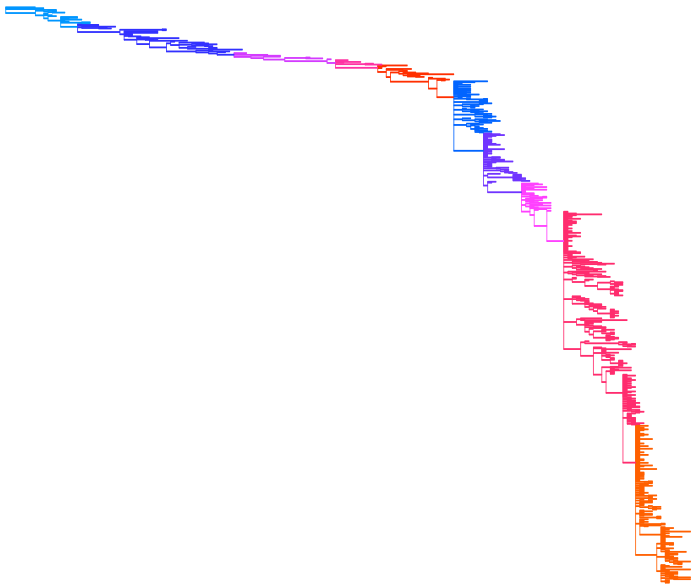
Scale: epidemiological & antigenic & genetic data

- phylodynamic multi-SIRS model; more θ:

  nucl mut rate 5.7×10^{-3}/site/yr (fix); population multiplier $e \sim [0, 250]$; rel sel advantage $s \sim [0, 0.1]$

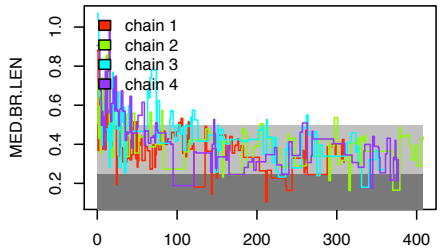
- MCMC

- in principle, punctuated antigenic change can reproduce limited diversity



$\varepsilon = log(S(x_0)/S(x))$

$exp(0.4) \approx 1.5$ fold, $exp(0.7) \approx 2$ fold discrepancy



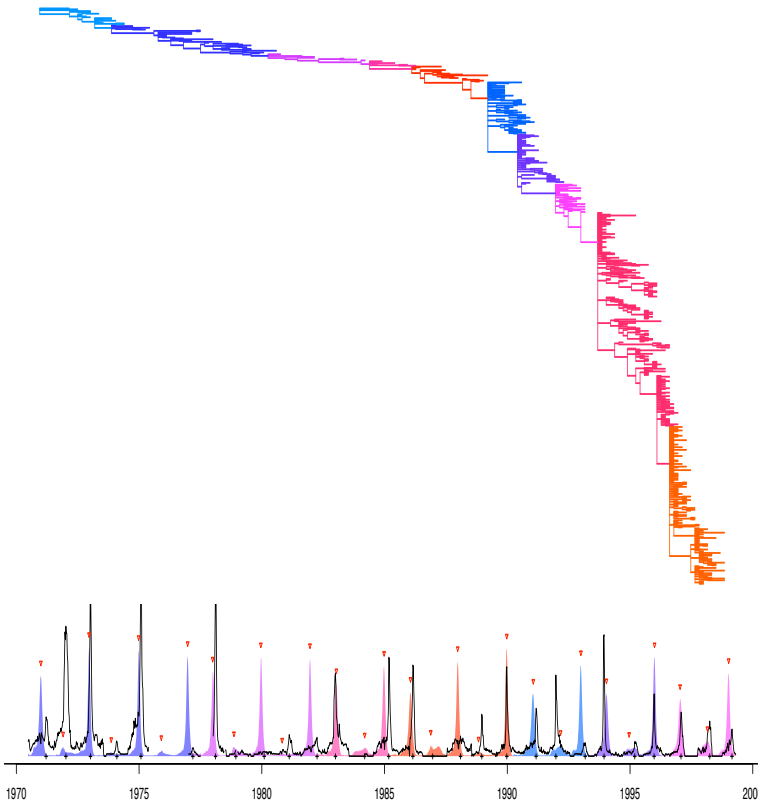
Scale: epidemiological & antigenic & genetic data

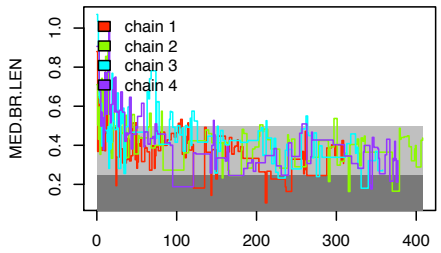
- phylodynamic multi-SIRS model; more θ:

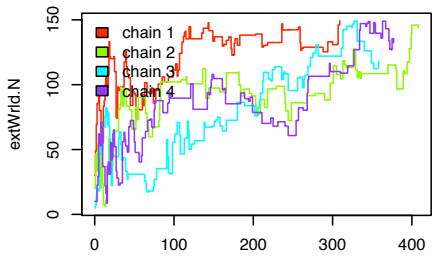
  nucl mut rate 5.7×10^{-3}/site/yr (fix); population multiplier $e \sim [0, 250]$; rel sel advantage $s \sim [0, 0.1]$

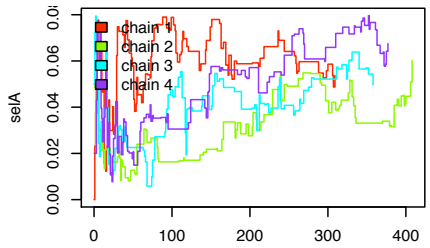
- MCMC

→ but only for unrealistic population size, $eN = 1500e6 \gg 400e6$









Take home

- 2-dim ABC error density useful to diagnose conflict
- existing ABC algorithms are easily extended
- straightforward to include multiple lines of evidence into ABC

Take home

- 2-dim ABC error density useful to diagnose conflict
- existing ABC algorithms are easily extended
- straightforward to include multiple lines of evidence into ABC

Collaborators

Sylvia Richardson

Christophe Andrieu

Katia Koelle

Christophe Fraser

Acknowledgments

Christian Robert

funding:

Wellcome Trust, Sir Henry Wellcome fellowship

NSF Advancing Theory in Biology grant NSF-EF-08-27416

Added value: model design via conflicting errors $\varepsilon_{1:K}$

ABC error = prior predictive error weighted by error magnitude

$$\pi_\tau(\varepsilon_{1:K}|x_0) \propto \pi_{x_0}(\varepsilon_{1:K}) \times \prod_k \kappa_k(\varepsilon_k;\tau_k)$$

Consider mean ABC error $\mathbb{E}_{\pi_\tau}(\varepsilon_{1:K}|x_0)$

Vector-valued integral:

$$\mathbb{E}_\tau\left(\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\end{pmatrix}\right) = \begin{pmatrix}\int \varepsilon_1 \pi_\tau(\varepsilon_{1:2}|x_0)d\varepsilon_{1:2}\\ \int \varepsilon_2 \pi_\tau(\varepsilon_{1:2}|x_0)d\varepsilon_{1:2}\end{pmatrix}$$

kth component re-weighted according to mutual constraints

$$\left[\mathbb{E}_{\pi_\tau}(\varepsilon_{1:K}|x_0)\right]_k = \int \rho_k(S_k(x),S_k(x_0))\prod_k \kappa_k\Big(\rho_k(S_k(x),S_k(x_0))\ ;\ \tau_k\Big)\pi(dx)$$

No conflict possible with independent summaries

$$\left[\mathbb{E}_{\pi_\tau}(\varepsilon_{1:K}|x_0)\right]_k = \int \rho_k(S_k(x),S_k(x_0))\kappa_k(\rho_k(S_k(x),S_k(x_0));\tau_k)\pi(dx)$$

Added value: model design via conflicting errors $\varepsilon_{1:K}$

ABC error = prior predictive error weighted by error magnitude

$$\pi_\tau(\varepsilon_{1:K}|x_0) \propto \pi_{x_0}(\varepsilon_{1:K}) \times \prod_k \kappa_k(\varepsilon_k;\tau_k)$$

Consider mean ABC error $\mathbb{E}_{\pi_\tau}(\varepsilon_{1:K}|x_0)$

à vector-valued integral:

$$\mathbb{E}_{\pi_\tau}\left(\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\end{pmatrix}\right) = \begin{pmatrix}\int \varepsilon_1 \pi_\tau(\varepsilon_{1:2}|x_0)d\varepsilon_{1:2}\\ \int \varepsilon_2 \pi_\tau(\varepsilon_{1:2}|x_0)d\varepsilon_{1:2}\end{pmatrix}$$

- kth component re-weighted according to mutual constraints

$$\left[\mathbb{E}_{\pi_\tau}(\varepsilon_{1:K}|x_0)\right]_k = \int \rho_k(S_k(x),S_k(x_0))\prod_k \kappa_k\Big(\rho_k(S_k(x),S_k(x_0))\ ;\ \tau_k\Big)\pi(dx)$$

- No conflict possible with independent summaries

$$\left[\mathbb{E}_{\pi_\tau}(\varepsilon_{1:K}|x_0)\right]_k = \int \rho_k(S_k(x),S_k(x_0))\kappa_k\big(\rho_k(S_k(x),S_k(x_0));\tau_k\big)\pi(dx)$$