Computer intensive statistical methods for Markov chain models in Biology

José Miguel Ponciano: josemi@ufl.edu

University of Florida, Biology Department

Examples: Continuous time SIS model and two SDE models

Example: a continuous time stochastic SIS model

SIS ODE model:

$$\frac{dI}{dt} = \frac{\beta}{N} S(I + \epsilon) - gI$$

$S = N - I = \#$ of susceptibles, $N =$ total pop. size (cst.)

β is the contact rate,

$\epsilon =$import of infection from an external source ($\epsilon = 0$ if pop. is isolated)

$g =$ recovery rate

Analogous to Levins 1969 metapopulation model (Hosts are empty (S) or occupied(I)).

Example: a continuous time stochastic SIS model

SIS ODE model:

$$\frac{dI}{dt} = \frac{\beta}{N} S(I + \epsilon) - gI$$

$S = N - I = \#$ of susceptibles, $N =$ total pop. size (cst.)

β is the contact rate,

$\epsilon =$import of infection from an external source ($\epsilon = 0$ if pop. is isolated)

$g =$ recovery rate

Analogous to Levins 1969 metapopulation model (Hosts are empty (S) or occupied(I)).

Stochastic version: the states are $I = 0, 1, \ldots, N$. At $t = 0$, $I(0) = k$. So $P(I(0) = k) = 1$ and

$$P(I(t) = i) = p_i(t) = P(I(t) = i | X(0) = k).$$

Kolmogorov-Forward equation

If the process is in state i at time t, then at time $t+\Delta t$ it will be either at state $i+1, i-1$ or i (Δt chosen so that at most 1 event occur). Therefore,

$$\begin{aligned} p_i(t+\Delta t) &= p_{i-1}(t)(\Delta t)\left[\tfrac{\beta}{N}S(t)(I(t)+\epsilon)\right] + p_{i+1}(t)(\Delta t)gI(t) \\ &\quad + p_i(t)\left[1-(\Delta t)\tfrac{\beta}{N}S(t)(I(t)+\epsilon)+gI(t)\right]. \end{aligned}$$

Hence

$$\begin{aligned} \tfrac{p_i(t+\Delta t)-p_i(t)}{\Delta t} &= p_{i-1}(t)\left[\tfrac{\beta}{N}S(t)(I(t)+\epsilon)\right] + p_{i+1}(t)(\Delta t)gI(t) \\ &\quad - p_i(t)\left[\tfrac{\beta}{N}S(t)(I(t)+\epsilon)+gI(t)\right], \end{aligned}$$

and since $S = N - I\,\forall\, t$, and letting $\Delta t \to 0$ we get

$$\frac{dp_i(t)}{dt} = p_{i-1}(t)\frac{\beta}{N}(N-i+1)(i-1+\epsilon) + p_{i+1}(t)g(i+1) - p_i(t)\left[\frac{\beta}{N}(N-i)(i+\epsilon)+gi\right]$$

The transition rates matrix Q

In vector notation,

$$\frac{dp_i(t)}{dt} = p_{i-1}(t)\frac{\beta}{N}(N-i+1)(i-1+\epsilon) + p_{i+1}(t)g(i+1) - p_i(t)\left[\frac{\beta}{N}(N-i)(i+\epsilon) + gi\right]$$

becomes $\frac{d\mathbf{p}}{dt} = \mathbf{p}Q$, where $\dim(\mathbf{p}) = 1 \times (N+1)$ and $\dim(Q) = (N+1) \times (N+1)$:

The transition rates matrix Q

In vector notation,

$$\frac{dp_i(t)}{dt} = p_{i-1}(t)\frac{\beta}{N}(N-i+1)(i-1+\epsilon) + p_{i+1}(t)g(i+1) - p_i(t)\left[\frac{\beta}{N}(N-i)(i+\epsilon) + gi\right]$$

becomes $\frac{d\mathbf{p}}{dt} = \mathbf{p}Q$, where $\dim(\mathbf{p}) = 1 \times (N+1)$ and $\dim(Q) = (N+1) \times (N+1)$:

$$Q = \begin{bmatrix} -\beta\epsilon & \beta\epsilon & 0 & 0 & \dots \\ g & -\left[\frac{\beta}{N}(N-1)(1+\epsilon)+g\right] & \frac{\beta}{N}(N-1)(1+\epsilon) & 0 & \dots \\ 0 & 2g & -\left[\frac{\beta}{N}(N-2)(2+\epsilon)+2g\right] & \frac{\beta}{N}(N-2)(2+\epsilon) & \dots \\ 0 & 0 & 3g & -\left[\frac{\beta}{N}(N-3)(3+\epsilon)+3g\right] & \dots \\ 0 & 0 & 0 & 4g & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

The transition rates matrix Q

In vector notation,

$$\frac{dp_i(t)}{dt} = p_{i-1}(t)\frac{\beta}{N}(N-i+1)(i-1+\epsilon) + p_{i+1}(t)g(i+1) - p_i(t)\left[\frac{\beta}{N}(N-i)(i+\epsilon) + gi\right]$$

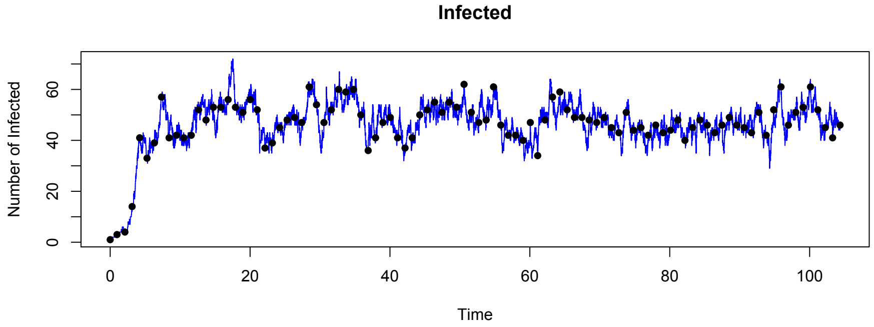
becomes $\frac{d\mathbf{p}}{dt} = \mathbf{p}Q$, where $\dim(\mathbf{p}) = 1 \times (N+1)$ and $\dim(Q) = (N+1) \times (N+1)$:

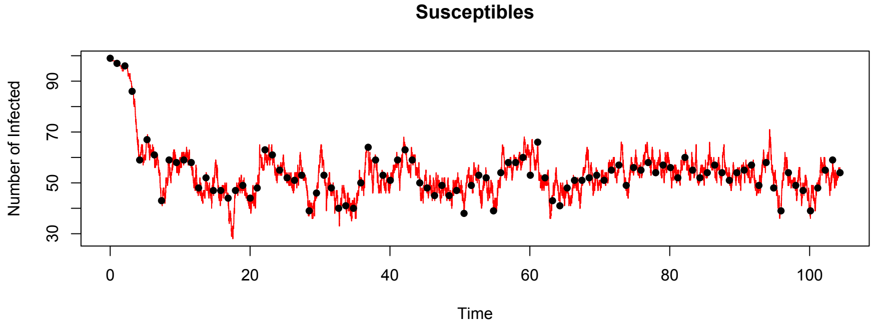
$$Q = \begin{bmatrix} -\beta\epsilon & \beta\epsilon & 0 & 0 & \dots \\ g & -\left[\frac{\beta}{N}(N-1)(1+\epsilon)+g\right] & \frac{\beta}{N}(N-1)(1+\epsilon) & 0 & \dots \\ 0 & 2g & -\left[\frac{\beta}{N}(N-2)(2+\epsilon)+2g\right] & \frac{\beta}{N}(N-2)(2+\epsilon) & \dots \\ 0 & 0 & 3g & -\left[\frac{\beta}{N}(N-3)(3+\epsilon)+3g\right] & \dots \\ 0 & 0 & 0 & 4g & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Solution to the system of ODEs: if $\mathbf{p}(0) = \mathbf{p_0}$, then $\mathbf{p_t} = \mathbf{p_0}\exp\{Qt\}$

Observing a realization of the process

- Observations at times $t_1 < t_2 < \ldots < t_{q-1} < t_q$. Let $\tau_i = t_i - t_{i-1}$ as before.
- States: $i_1, i_2, \ldots, i_{q-1}, i_q$


Infected


Susceptibles

The likelihood function

$$L(\boldsymbol{\theta}) \;=\; P(I(t_1) = i_1, I(t_2) = i_2, \ldots, I(t_{q-1}) = i_{q-1}, I(t_q) = i_q)$$

The likelihood function

$$\begin{aligned} L(\boldsymbol{\theta}) &= P(I(t_1)=i_1, I(t_2)=i_2, \ldots, I(t_{q-1})=i_{q-1}, I(t_q)=i_q) \\ &= P(I(t_1)=i_1) \times P(I(t_2)=i_2|I(t_1)=i_1) \times P(I(t_q)=i_q|I(t_{q-1})=i_{q-1}) \end{aligned}$$

The likelihood function

$$\begin{aligned} L(\boldsymbol{\theta}) &= P(I(t_1) = i_1, I(t_2) = i_2, \ldots, I(t_{q-1}) = i_{q-1}, I(t_q) = i_q) \\ &= P(I(t_1) = i_1) \times P(I(t_2) = i_2 | I(t_1) = i_1) \times P(I(t_q) = i_q | I(t_{q-1}) = i_{q-1}) \\ &= \{\mathbf{p_{t_1}}\}_{i_1} \times \{\exp((t_2 - t_1)Q)\}_{i_1,i_2} \times \{\exp((t_3 - t_2)Q)\}_{i_2,i_3} \times \ldots \end{aligned}$$

The likelihood function

$$\begin{aligned}
L(\boldsymbol{\theta}) &= P(I(t_1) = i_1, I(t_2) = i_2, \ldots, I(t_{q-1}) = i_{q-1}, I(t_q) = i_q) \\
&= P(I(t_1) = i_1) \times P(I(t_2) = i_2 | I(t_1) = i_1) \times P(I(t_q) = i_q | I(t_{q-1}) = i_{q-1}) \\
&= \{\mathbf{p_{t_1}}\}_{i_1} \times \{\exp((t_2 - t_1)Q)\}_{i_1,i_2} \times \{\exp((t_3 - t_2)Q)\}_{i_2,i_3} \times \ldots \\
&= \{\mathbf{p_{t_1}}\}_{i_1} \times \textstyle\prod_{k=2}^{q} \{\exp(\tau_k Q)\}_{i_{k-1},i_k}
\end{aligned}$$

The likelihood function

$$\begin{aligned}
L(\boldsymbol{\theta}) &= P(I(t_1) = i_1, I(t_2) = i_2, \ldots, I(t_{q-1}) = i_{q-1}, I(t_q) = i_q) \\
&= P(I(t_1) = i_1) \times P(I(t_2) = i_2 | I(t_1) = i_1) \times P(I(t_q) = i_q | I(t_{q-1}) = i_{q-1}) \\
&= \{\mathbf{p_{t_1}}\}_{i_1} \times \{\exp((t_2 - t_1)Q)\}_{i_1,i_2} \times \{\exp((t_3 - t_2)Q)\}_{i_2,i_3} \times \ldots \\
&= \{\mathbf{p_{t_1}}\}_{i_1} \times \textstyle\prod_{k=2}^{q} \{\exp(\tau_k Q)\}_{i_{k-1},i_k} \\
&= \{\mathbf{p_{t_1}}\}_{i_1} \times \textstyle\prod_{k=2}^{q} \{\mathbf{I}_{i_{k-1}} \times \exp(\tau_k Q)\}_{i_k}, \text{ where}
\end{aligned}$$

$\mathbf{I}_j$ is a vector that has zeros everywhere, except in the j^{th} position where it has a 1.

The likelihood function

$$\begin{aligned}
L(\boldsymbol{\theta}) &= P(I(t_1)=i_1, I(t_2)=i_2, \ldots, I(t_{q-1})=i_{q-1}, I(t_q)=i_q) \\
&= P(I(t_1)=i_1) \times P(I(t_2)=i_2|I(t_1)=i_1) \times P(I(t_q)=i_q|I(t_{q-1})=i_{q-1}) \\
&= \{\mathbf{p_{t_1}}\}_{i_1} \times \{\exp((t_2-t_1)Q)\}_{i_1,i_2} \times \{\exp((t_3-t_2)Q)\}_{i_2,i_3} \times \ldots \\
&= \{\mathbf{p_{t_1}}\}_{i_1} \times \textstyle\prod_{k=2}^{q}\{\exp((t_k-t_{k-1})Q)\}_{i_{k-1},i_k} \\
&= \{\mathbf{p_{t_1}}\}_{i_1} \times \textstyle\prod_{k=2}^{q}\{\mathbf{I}_{i_{k-1}} \times \exp(\tau_k Q)\}_{i_k}, \text{ where}
\end{aligned}$$

$\mathbf{I}_j$ is a vector that has zeros everywhere, except in the j^{th} position where it has a 1.
Notes: Computing $\exp(\tau Q)$ can be done using a matrix exponentiation algorithm (only once per each iteration of the maximization routine if all τ_k's are equal). However, can greatly reduce computations by calculating $\mathbf{I}_j \exp(\tau Q)$ (a vector) instead of $\exp(\tau Q)$ (a matrix). These are the so-called Krylov space methods. (Citation: On 19 dubious ways. . .)

Matrix exponentiation

A matrix exponentiation is achieved using a T.S. expansion:

$$\exp(\tau Q) = \mathbf{I} + \tau Q + \frac{\tau^2}{2!}Q^2 + \frac{\tau^3}{3!}Q^3 + \ldots$$

Matrix exponentiation

A matrix exponentiation is achieved using a T.S. expansion:

$$\exp(\tau Q) = \mathbf{I} + \tau Q + \frac{\tau^2}{2!}Q^2 + \frac{\tau^3}{3!}Q^3 + \dots$$

Hence

$$\mathbf{I}_j \exp(\tau Q) = \mathbf{I}_j\mathbf{I} + \mathbf{I}_j\tau Q + \mathbf{I}_j\frac{\tau^2}{2!}Q^2 + \mathbf{I}_j\frac{\tau^3}{3!}Q^3 + \dots$$

Matrix exponentiation

A matrix exponentiation is achieved using a T.S. expansion:

$$\exp(\tau Q) = \mathbf{I} + \tau Q + \frac{\tau^2}{2!}Q^2 + \frac{\tau^3}{3!}Q^3 + \ldots$$

Hence

$$\mathbf{I}_j \exp(\tau Q) = \mathbf{I}_j\mathbf{I} + \mathbf{I}_j\tau Q + \mathbf{I}_j\frac{\tau^2}{2!}Q^2 + \mathbf{I}_j\frac{\tau^3}{3!}Q^3 + \ldots$$

- However this is definitely NOT the way to compute it because of accumulation of numerical round-off errors!

Matrix exponentiation

A matrix exponentiation is achieved using a T.S. expansion:

$$\exp(\tau Q) = \mathbf{I} + \tau Q + \frac{\tau^2}{2!}Q^2 + \frac{\tau^3}{3!}Q^3 + \ldots$$

Hence

$$\mathbf{I}_j \exp(\tau Q) = \mathbf{I}_j\mathbf{I} + \mathbf{I}_j\tau Q + \mathbf{I}_j\frac{\tau^2}{2!}Q^2 + \mathbf{I}_j\frac{\tau^3}{3!}Q^3 + \ldots$$

- However this is definitely NOT the way to compute it because of accumulation of numerical round-off errors!
- A program to simulate and estimate parameters for this model using R will be reviewed in the computer session in the afternoon.
- About simulation: with certain computer intensive methods for parameter estimation, all we need is to simulate realizations from the process *conditioned on the ending point*. To do that, use Hobolth and Stone (2009), Annals of Applied Statistics).

Matrix exponentiation

A matrix exponentiation is achieved using a T.S. expansion:

$$\exp(\tau Q) = \mathbf{I} + \tau Q + \frac{\tau^2}{2!}Q^2 + \frac{\tau^3}{3!}Q^3 + \ldots$$

Hence

$$\mathbf{I}_j \exp(\tau Q) = \mathbf{I}_j\mathbf{I} + \mathbf{I}_j\tau Q + \mathbf{I}_j\frac{\tau^2}{2!}Q^2 + \mathbf{I}_j\frac{\tau^3}{3!}Q^3 + \ldots$$

- However this is definitely NOT the way to compute it because of accumulation of numerical round-off errors!

- A program to simulate and estimate parameters for this model using R will be reviewed in the computer session in the afternoon.

- About simulation: with certain computer intensive methods for parameter estimation, all we need is to simulate realizations from the process *conditioned on the ending point*. To do that, use Hobolth and Stone (2009), Annals of Applied Statistics).

- Sampling error?

SDE's I: a heuristic introduction using the Wright-Fisher model

- Let $X(k) =$ be the number of A_1 individuals in generation k, for a (constant) population size N: $X^N(k)$
- The states are $S = 0, 1, \ldots, N$
- 0 and N can be absorbing states, depending on the biological scenario (*i.e.* no mutation, no selection)
- The number of A_1's for next generation follows a Binomial law (sampling *with replacement*):

$$p_{i,j} = P(X(k+1) = j | X(k) = i) = \binom{N}{j} p_i^j (1 - p_i)^{N-j}$$

- No mutation and no selection: $p_i = \frac{i}{N}$
- Two-way mutation: $p_i = \frac{i}{N}(1 - \mu_{12}) + \left(1 - \frac{i}{N}\right) \mu_{21}$
- With selection and without mutation: $p_i = \frac{(1+s)i}{(1+s)i+(N-i)}$

Approximating the discrete process with a continuous time, continuous state MC

Recall that a diffusion process X_t satisfies:

1. $\lim_{h\to 0+} \frac{1}{h}\mathrm{E}[X_{t+h} - X_t | X_t = x] = \mu(x)$
   drift parameter
2. $\lim_{h\to 0+} \frac{1}{h}\mathrm{E}[(X_{t+h} - X_t)^2 | X_t = x] = \sigma^2(x)$
   diffusion parameter
3. $\lim_{h\to 0+} \frac{1}{h}\mathrm{E}[(X_{t+h} - X_t)^4 | X_t = x] = 0$
   needed for continuity of trajectories

Brownian Motion satisfies:

1. continuous trajectories
2. independent increments: $B_{t_1} - B_{s_1}$ indep. of $B_{t_2} - B_{s_2}$, $0 \le s_1 \le t_1 \le s_2 \le t_2$.
3. Normal increments: $B_t - B_s \sim \mathrm{N}(0, t-s)$, $s < t$
4. Drift: $\mu = 0$. Diffusion: $\sigma^2(x) = 1$.

Why 'drift' and 'diffusion' parameters?

$$\lim_{h\to 0+} \frac{1}{h} E[X_{t+h} - X_t | X_t = x] = \mu(x) \quad \text{then}$$

$$E[X_{t+h} - X_t | X_t = x] \sim \mu(x)h \quad \text{(h small)}$$

That is, when the process is in x, it changes an amount of $\mu(x)h$ in the next h units of time.
$Var(X_{t+h} - X_t | X_t = x)$

$$\begin{aligned} &= E[(X_{t+h} - X_t)^2 | X_t = x] - (E[X_{t+h} - X_t | X_t = x])^2 \\ &\approx \sigma^2(x)h - \mu(x)^2 h^2 \approx \sigma^2(x)h \end{aligned}$$

where $\sigma^2(x)$ is the infinitesimal variance of the diffusion. It tells us how much does the process tends to move when at x.

Construction of diffusions

If B_t denotes B.M., then $X_t = aB_t + bt$ is a diffusion process such that:

1. $\lim_{h\to 0+} \frac{1}{h}\mathrm{E}[X_{t+h} - X_t | X_t = x] = \mu(x) = b$
   is the drift parameter
2. $\lim_{h\to 0+} \frac{1}{h}\mathrm{E}[(X_{t+h} - X_t)^2 | X_t = x] = \sigma^2(x) = a^2$
   is the diffusion parameter
3. satisfies the condition of continuity.

Differential notation: $dX_t = adB_t + bdt$. In general:

$$dX_t = \mu(X_t)dt + \sqrt{\sigma^2(X_t)}dB_t,$$

$$X_t - X_0 = \int_0^t \mu(X_s)ds + \int_0^t \sigma(X_s)dB_s.$$

Construction of diffusions

Approximating a family of Markov Chains in discrete time:

$X^{N}(k), \quad k = 0, 1, 2, \ldots$

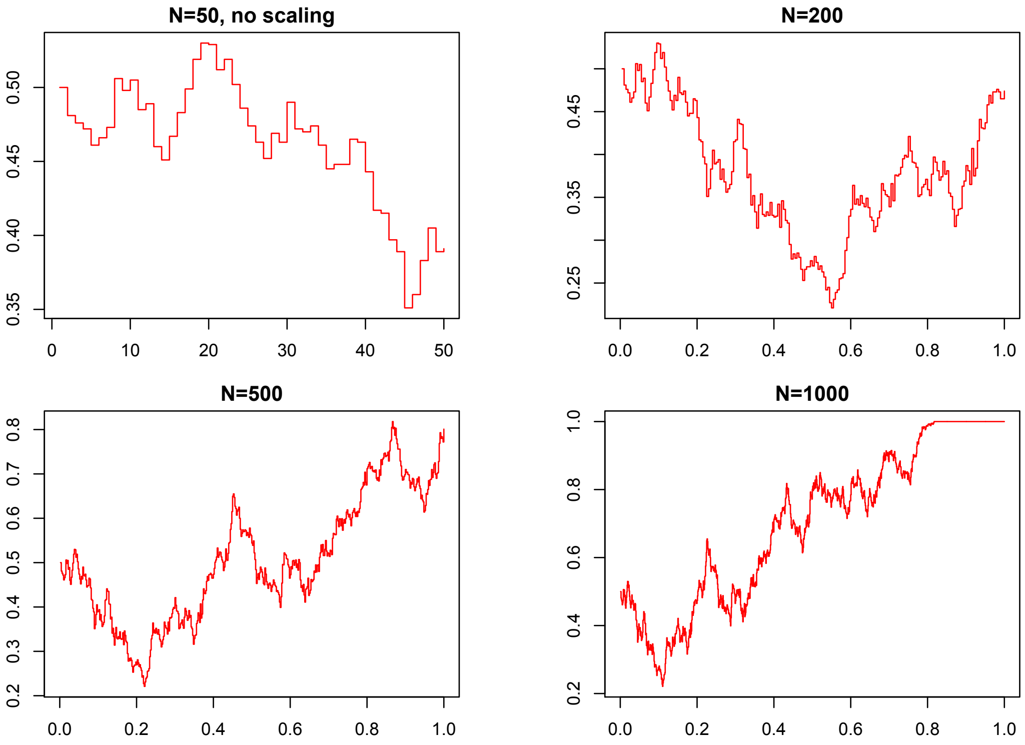
Define: $Y_t^N \equiv \frac{1}{a(N)} X^N([Nt]) \ldots$ as the scaled process.

Idea: accelerate time by a factor N and scale space by a factor $\frac{1}{a(N)}$.

With the adequate scaling, $\lim_{t \to \infty} Y_t^N = Y_t$, a limiting diffusion.

With the proper scaling, all of these processes look the same when N is large. Hence, a single limiting diffusion process approximates all the scaled Markov chains. It's like the CLT! For the W-F model, the scaling is $i/a(N) = i/N$, the fraction of A_1's is a natural scaling

Scaling W-F to a diffusion



$Y_N(t) = \frac{1}{N} X^N([Nt]) \Rightarrow Y_t \ldots$ a diffusion in $[0, 1]$

$$\lim_{t\to\infty} Y_t^N = Y_t$$

When does the approximation work? Let $h = 1/N$. If, as $N \to \infty$ and $i/a(N) \to x$ we have that:

1. $\frac{1}{1/N}\mathrm{E}[Y^N_{t+\frac{1}{N}} - Y^N_t | Y^N_t = \frac{i}{a(N)}] \to \mu(x)$,
2. $\frac{1}{1/N}\mathrm{E}[(Y^N_{t+\frac{1}{N}} - Y^N_t)^2 | Y^N_t = \frac{i}{a(N)}] \to \sigma^2(x)$,
3. $\frac{1}{1/N}\mathrm{E}[(Y^N_{t+\frac{1}{N}} - Y^N_t)^4 | Y^N_t = \frac{i}{a(N)}] \to 0$.

For the W-F model using the scaling $i/a(N) = i/N$:

- drift $\mu(x) = 0$ (neutral case)
- diffusion $\sigma^2(x) = x(1-x)$. This comes from the “binomial sampling noise” due to finite population size.

A little warning: the diffusion term is called “genetic drift” in population genetics

Drift calculation for the neutral WF diffusion

Knowing that $(X([Nt]+1)|X([Nt])=i) \sim \mathrm{Bin}(N, p_i = i/N)$ we get

$$\begin{aligned}
E\left[Y^N_{t+\frac{1}{N}} - Y^N_t | Y^N_t = \tfrac{i}{N}\right] &= E\left[\tfrac{X([N(t+\frac{1}{N})])}{N} - \tfrac{X([Nt])}{N} | \tfrac{X([Nt])}{N} = \tfrac{i}{N}\right] \\
&= \tfrac{1}{N}E[X([Nt]+1) - i | X([Nt]) = i] \\
&= \tfrac{1}{N}(Np_i - i) = \tfrac{1}{N}(N\tfrac{i}{N} - i) = 0, \ \forall N.
\end{aligned}$$

So $\frac{1}{1/N}E\left[Y^N_{t+\frac{1}{N}} - Y^N_t | Y^N_t = \frac{i}{N}\right] = 0$ and hence $\mu(x) = 0$.

Other scenarios

With mutation probabilities $u_{12} = \frac{\theta_{12}}{N}$ and $u_{21} = \frac{\theta_{21}}{N}$ we get

$$p_i = \frac{i}{N}(1 - u_{12} + (1 - \frac{i}{N})u_{21},$$

and the diffusion approximation has

- $\mu(x) = -\theta_{12}x + \theta_{21}(1 - x)$ Boundary no longer absorbing
- $\sigma^2 = x(1 - x)$

With fitness advantage $s = \frac{\sigma}{N}$ and no mutation,

$$p_i = \frac{(1 + s)i}{(1 + s)i + (N - i)}$$

and

- $\mu(x) = \sigma x(1 - x)$
- $\sigma^2 = x(1 - x)$

A hierarchical model using Wright-Fisher

$$\begin{aligned}\mathbf{Y}|\mathbf{X} &\sim \mathrm{Binom}\left(M=50,\mathbf{X}\right)\\ \mathbf{X} &\sim g(\mathbf{x};\theta),\end{aligned}$$

where

- $g(\mathbf{x}|\theta)$ is given by the transition pdf of a Wright-Fisher diffusion (which lives between $[0,1]$).
- Data example: Antibiotic resistant, antibiotic sensitive bacteria, sample size at each time step: $M = 50$.
- Likelihood:
$$L(\theta,\phi) = \int f(\mathbf{y}|\mathbf{X},\phi)g(\mathbf{x};\theta)d\mathbf{X}.$$

A hierarchical model using Wright-Fisher

$$\begin{aligned} \mathbf{Y}|\mathbf{X} &\sim \mathrm{Binom}\left(M = 50, \mathbf{X}\right) \\ \mathbf{X} &\sim g(\mathbf{x};\theta), \end{aligned}$$

where

- $g(\mathbf{x};\theta)$ is given by the product of the transition pdfs for the Wright-Fisher diffusion, evaluated at the observed time steps (Remember that the chain lives between $[0, 1]$).
- Data example: Antibiotic resistant, antibiotic sensitive bacteria, sample size at each time step: $M = 50$.
- Likelihood:
$$L(\theta, \phi) = \int f(\mathbf{y}|\mathbf{X}, \phi) g(\mathbf{x};\theta) d\mathbf{X}.$$
- Likelihood uses:
  - The transition pdf of the diffusion process or
  - the stationary density, if it exists (and data at stationarity)

$$L(\theta, \phi) = \int f(\mathbf{y}|\mathbf{X}, \phi) g(\mathbf{x}; \theta) d\mathbf{X}.$$

$g(\mathbf{x}; \theta)$ is the product of the transition pdfs, just as in the continuous MC case.

$$L(\theta, \phi) = \int f(\mathbf{y}|\mathbf{X}, \phi) g(\mathbf{x}; \theta) d\mathbf{X}.$$

$g(\mathbf{x}; \theta)$ is the product of the transition pdfs, just as in the continuous MC case.

The transition pdf $p_t(x, x'; \theta)$ satisfies the Kolmogorov-Forward equation

$$\frac{\partial}{\partial t} p_t(x, x'; \theta) = A^\star p_t(x, x'; \theta),$$ where $A^\star$ acts on x' and

$$L(\theta, \phi) = \int f(\mathbf{y}|\mathbf{X}, \phi) g(\mathbf{x}; \theta) d\mathbf{X}.$$

$g(\mathbf{x}; \theta)$ is the product of the transition pdfs, just as in the continuous MC case.

The transition pdf $p_t(x, x'; \theta)$ satisfies the Kolmogorov-Forward equation

$$\frac{\partial}{\partial t} p_t(x, x'; \theta) = A^\star p_t(x, x'; \theta), \text{ where } A^\star \text{ acts on } x' \text{ and}$$

$$A^\star p_t(x, x'; \theta) = -\frac{\partial}{\partial x'}[\mu(x') p_t(x, x'; \theta)] + \frac{1}{2}\frac{\partial^2}{\partial x'^2}[\sigma^2(x') p_t(x, x'; \theta)].$$

$$L(\theta, \phi) = \int f(\mathbf{y}|\mathbf{X}, \phi) g(\mathbf{x}; \theta) d\mathbf{X}.$$

$g(\mathbf{x}; \theta)$ is the product of the transition pdfs, just as in the continuous MC case.

The transition pdf $p_t(x, x'; \theta)$ satisfies the Kolmogorov-Forward equation

$$\frac{\partial}{\partial t} p_t(x, x'; \theta) = A^\star p_t(x, x'; \theta), \text{ where } A^\star \text{ acts on } x' \text{ and}$$

$$A^\star p_t(x, x'; \theta) = -\frac{\partial}{\partial x'}[\mu(x') p_t(x, x'; \theta)] + \frac{1}{2}\frac{\partial^2}{\partial x'^2}[\sigma^2(x') p_t(x, x'; \theta)].$$

Under certain conditions, $\lim_{t \to \infty} p_t(x, z) = \pi(z)$ exists. If this limit exists and $\pi(z)$ is a probability density, *i.e.*

$$\pi(z) \geq 0, \int_{-\infty}^{\infty} \pi(z) dz = 1, \text{ then this is the stationary density and}$$

$$L(\theta,\phi) = \int f(\mathbf{y}|\mathbf{X},\phi)g(\mathbf{x};\theta)d\mathbf{X}.$$

$g(\mathbf{x};\theta)$ is the product of the transition pdfs, just as in the continuous MC case.

The transition pdf $p_t(x,x';\theta)$ satisfies the Kolmogorov-Forward equation

$$\frac{\partial}{\partial t}p_t(x,x';\theta) = A^\star p_t(x,x';\theta),\text{ where } A^\star \text{ acts on } x' \text{ and}$$

$$A^\star p_t(x,x';\theta) = -\frac{\partial}{\partial x'}[\mu(x')p_t(x,x';\theta)] + \frac{1}{2}\frac{\partial^2}{\partial x'^2}[\sigma^2(x')p_t(x,x';\theta)].$$

Under certain conditions, $\lim_{t\to\infty} p_t(x,z) = \pi(z)$ exists. If this limit exists and $\pi(z)$ is a probability density, *i.e.*

$$\pi(z) \geq 0, \int_{-\infty}^{\infty} \pi(z)dz = 1, \text{ then this is the stationary density and}$$

$$\begin{aligned} \lim_{t\to\infty} P^x(X_t \in E) &= \textstyle\int_E \pi(z)dz \text{ and} \\ \lim_{t\to\infty} E^x f(X_t) &= \textstyle\int_{-\infty}^{\infty} f(z)\pi(z)dz \end{aligned}$$

$$L(\theta, \phi) = \int f(\mathbf{y}|\mathbf{X}, \phi) g(\mathbf{x}; \theta) d\mathbf{X}.$$

$g(\mathbf{x}; \theta)$ is the product of the transition pdfs, just as in the continuous MC case.

The transition pdf $p_t(x, x'; \theta)$ satisfies the Kolmogorov-Forward equation

$$\frac{\partial}{\partial t} p_t(x, x'; \theta) = A^\star p_t(x, x'; \theta), \text{ where } A^\star \text{ acts on } x' \text{ and}$$

$$A^\star p_t(x, x'; \theta) = -\frac{\partial}{\partial x'}[\mu(x') p_t(x, x'; \theta)] + \frac{1}{2}\frac{\partial^2}{\partial x'^2}[\sigma^2(x') p_t(x, x'; \theta)].$$

Under certain conditions, $\lim_{t\to\infty} p_t(x, z) = \pi(z)$ exists. If this limit exists and $\pi(z)$ is a probability density, *i.e.*

$$\pi(z) \geq 0, \int_{-\infty}^{\infty} \pi(z) dz = 1, \text{ then this is the stationary density and}$$

$$\begin{aligned} \lim\nolimits_{t\to\infty} P^x(X_t \in E) &= \textstyle\int_E \pi(z) dz \text{ and} \\ \lim\nolimits_{t\to\infty} E^x f(X_t) &= \textstyle\int_{-\infty}^{\infty} f(z) \pi(z) dz \end{aligned}$$

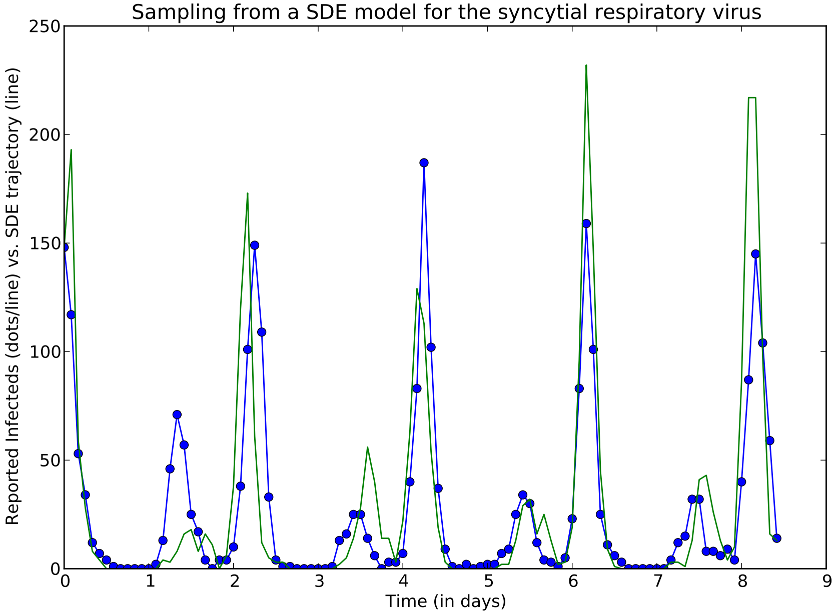
$\pi(z)$ satisfies $A^\star \pi(z) = 0$ (Letting $t \to \infty$ in the forward equation)

Writing the likelihood

Observations: $y_0, y_1, y_2, \ldots, y_q$ at times $0 < t_1 < t_2 < \ldots < t_q$

These are samples with error taken from a particular realization of the process:

$x_0, x_1, x_2, \ldots, x_q$



Writing the likelihood II

$$L(\theta, \phi) = \int f(\mathbf{y}|\mathbf{X}, \phi) g(\mathbf{x}; \theta) d\mathbf{X}.$$

Writing the likelihood II

$$L(\theta, \phi) = \int f(\mathbf{y}|\mathbf{X}, \phi) g(\mathbf{x}; \theta) d\mathbf{X}.$$

Again, let $\tau_1 = t_1 - 0, \tau_2 = t_2 - t_1, \ldots, \tau_q = t_q - t_{q-1}$. Then,

$$g(\mathbf{x}; \theta) = \prod_{i=1}^{q} p_{\tau_i}(x_{i-1}, x_i; \theta) = \prod_{i=1}^{q} p_{\tau_i}(x_i | x_{i-1}; \theta).$$

This is identical to the likelihood function without sampling error. Brauman (1983) shows ML estimation for equal time intervals, without sampling error. With sampling error, we have to integrate the statistical sampling model over all the possible realizations of the process

$$L(\theta) = \int \ldots \int \prod_{i=0}^{q} f(y_i|x_i) \prod_{i=0}^{q} p_{\tau_i}(x_i | x_{i-1}; \theta) dx_1 dx_2 \ldots dx_q$$

Writing the likelihood II

$$L(\theta, \phi) = \int f(\mathbf{y}|\mathbf{X}, \phi) g(\mathbf{x}; \theta) d\mathbf{X}.$$

Again, let $\tau_1 = t_1 - 0, \tau_2 = t_2 - t_1, \ldots, \tau_q = t_q - t_{q-1}$. Then,

$$g(\mathbf{x}; \theta) = \prod_{i=1}^{q} p_{\tau_i}(x_{i-1}, x_i; \theta) = \prod_{i=1}^{q} p_{\tau_i}(x_i | x_{i-1}; \theta).$$

This is identical to the likelihood function without sampling error. Brauman (1983) shows ML estimation for equal time intervals, without sampling error. With sampling error, we have to integrate the statistical sampling model over all the possible realizations of the process

$$L(\theta) = \int \ldots \int \underbrace{\prod_{i=0}^{q} f(y_i | x_i) \prod_{i=0}^{q} p_{\tau_i}(x_i | x_{i-1}; \theta)}_{\text{Only need to be able to write this down to run MCMC}} dx_1 dx_2 \ldots dx_q$$

Tier and Hanson 1982: Branching processes

Let Z_n be the total pop. size at time n and B_i be the offspring distribution such that $p_j(z) = P(B_i = j | Z_n = z)$. Let also

$$\begin{cases} E[B_i|Z_n = z] & = & h(z) \\ V[B_i|Z_n = z] & = & v(z) \\ Z_{n+1} & = & \sum_{i=1}^{Z_n} B_i \text{ (a randomly stopped sum)} \end{cases}$$

Then

$$\begin{aligned} E[\Delta Z_n|Z_n = z] = E[(Z_{n+1} - Z_n)|Z_n = z] &= E[\textstyle\sum_{i=1}^{z}(B_i|Z_n = z)] - z = z[h(z) - 1] \text{ and} \\ E[(\Delta Z_n)^2|Z_n = z] = E[(Z_{n+1} - Z_n)^2|Z_n = z] &= Var\left[(\textstyle\sum_{i=1}^{z}(B_i|Z_n = z)) - z\right] \\ &\quad + \left\{E\left[(\textstyle\sum_{i=1}^{z}(B_i|Z_n = z)) - z\right]\right\}^2 \\ &= zv(z) + \{z(h(z) - 1)\}^2 \end{aligned}$$

Diffusion approximation

$$\begin{aligned} E[\Delta Z_n | Z_n = z] &= z[h(z) - 1], \\ E[(\Delta Z_n)^2 | Z_n = z] &= zv(z) + \{z(h(z) - 1)\}^2 \end{aligned}$$

- Scaled process: $X(t) = \frac{Z_n}{L}$, where L is “some reference population size”
- Scaled time: $t = n\Delta t$, where $\Delta t << 1$ is the generation time.

Diffusion approximation

$$\begin{aligned} E[\Delta Z_n | Z_n = z] &= z[h(z) - 1], \\ E[(\Delta Z_n)^2 | Z_n = z] &= zv(z) + \{z(h(z) - 1)\}^2 \end{aligned}$$

- Scaled process: $X(t) = \frac{Z_n}{L}$, where L is "some reference population size"
- Scaled time: $t = n\Delta t$, where $\Delta t << 1$ is the generation time.
- Require: small changes in $X(t)$ occur in small time increments Δt (offspring mean close to replacement):

Diffusion approximation

$$\begin{aligned} E[\Delta Z_n | Z_n = z] &= z[h(z) - 1], \\ E[(\Delta Z_n)^2 | Z_n = z] &= zv(z) + \{z(h(z) - 1)\}^2 \end{aligned}$$

- Scaled process: $X(t) = \frac{Z_n}{L}$, where L is "some reference population size"
- Scaled time: $t = n\Delta t$, where $\Delta t << 1$ is the generation time.
- Require: small changes in $X(t)$ occur in small time increments Δt (offspring mean close to replacement):

  $h(z) = 1 + \delta\mu\left(\frac{z}{L}\right)$, where $\delta = (\Delta t)$ measures deviations from replacement and

Diffusion approximation

$$\begin{array}{lcl} E[\Delta Z_n | Z_n = z] & = & z[h(z) - 1], \\ E[(\Delta Z_n)^2 | Z_n = z] & = & zv(z) + \{z(h(z) - 1)\}^2 \end{array}$$

- Scaled process: $X(t) = \frac{Z_n}{L}$, where L is "some reference population size"
- Scaled time: $t = n\Delta t$, where $\Delta t << 1$ is the generation time.
- Require: small changes in $X(t)$ occur in small time increments Δt (offspring mean close to replacement):

$h(z) = 1 + \delta\mu\left(\frac{z}{L}\right)$, where $\delta = (\Delta t)$ measures deviations from replacement and

$\mu(x) = r\left(1 - \frac{x}{k}\right)$ is the per-capita growth rate of the logistic equation.

Diffusion approximation

$$\begin{aligned} E[\Delta Z_n | Z_n = z] &= z[h(z) - 1], \\ E[(\Delta Z_n)^2 | Z_n = z] &= zv(z) + \{z(h(z) - 1)\}^2 \end{aligned}$$

- Scaled process: $X(t) = \frac{Z_n}{L}$, where L is “some reference population size”
- Scaled time: $t = n\Delta t$, where $\Delta t << 1$ is the generation time.
- Require: small changes in $X(t)$ occur in small time increments Δt (offspring mean close to replacement):

$$h(z) = 1 + \delta\mu\left(\frac{z}{L}\right), \text{ where } \delta = (\Delta t) \text{ measures deviations from replacement and}$$

$$\mu(x) = r\left(1 - \frac{x}{k}\right) \text{ is the per-capita growth rate of the logistic equation.}$$

- Finally, denote $v(z) = d(\frac{z}{L})$. Then, as $\Delta t \to 0$ and $\frac{z}{L} \to x$

Diffusion approximation: let $\Delta t \to 0$ and $\frac{z}{L} \to x$

$$\tfrac{1}{\Delta t}E[\Delta X(t)|X(t)=x] \;=\; \tfrac{1}{\Delta t}E\left[\tfrac{(Z_{n+1}-Z_n)}{L}\middle|X(t)=\tfrac{z}{L}\right]=\tfrac{1}{\Delta t}\tfrac{z}{L}\left(h(z)-1\right)$$

Diffusion approximation: let $\Delta t \to 0$ and $\frac{z}{L} \to x$

$$\begin{aligned}
\frac{1}{\Delta t}E[\Delta X(t)|X(t)=x] &= \frac{1}{\Delta t}E\left[\frac{(Z_{n+1}-Z_n)}{L}\middle|X(t)=\frac{z}{L}\right] = \frac{1}{\Delta t}\frac{z}{L}\left(h(z)-1\right)\\
&= \frac{1}{\Delta t}\frac{z}{L}\left(\delta\mu\left(\frac{z}{L}\right)\right) \to x\mu(x) = xr\left(1-\frac{x}{k}\right)
\end{aligned}$$

Diffusion approximation: let $\Delta t \to 0$ and $\frac{z}{L} \to x$

$$
\begin{aligned}
\frac{1}{\Delta t}E[\Delta X(t)|X(t)=x] &= \frac{1}{\Delta t}E\left[\frac{(Z_{n+1}-Z_n)}{L}\middle|X(t)=\frac{z}{L}\right] = \frac{1}{\Delta t}\frac{z}{L}\left(h(z)-1\right)\\
&= \frac{1}{\Delta t}\frac{z}{L}\left(\delta\mu\left(\frac{z}{L}\right)\right) \to x\mu(x) = xr\left(1-\frac{x}{k}\right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{1}{\Delta t}E[(\Delta X(t))^2|X(t)=x] &= \frac{1}{\Delta t}E\left[\left(\frac{Z_{n+1}-Z_n}{L}\right)^2\middle|X(t)=\frac{z}{L}\right] = \frac{1}{\Delta t}\frac{1}{L^2}\left[zv(z)+\{z(h(z)-1)\}^2\right]\\
&= \frac{1}{\Delta t}\frac{1}{L^2}\left[zd\left(\frac{z}{L}\right)+z^2\delta^2\left(\mu\left(\frac{z}{L}\right)\right)^2\right] = \frac{1}{\Delta t}x^2\delta^2(\mu(x))^2+\frac{1}{\Delta t}\frac{1}{L}xd(x)\\
&= x^2\delta(\mu(x))^2+\frac{1}{\Delta t}(\Delta t)xd(x) \to xd(x) = x\beta \text{ (for example)}
\end{aligned}
$$

Diffusion approximation: let $\Delta t \to 0$ and $\frac{z}{L} \to x$

$$\frac{1}{\Delta t}E[\Delta X(t)|X(t)=x] = \frac{1}{\Delta t}E\left[\frac{(Z_{n+1}-Z_n)}{L}\middle|X(t)=\frac{z}{L}\right] = \frac{1}{\Delta t}\frac{z}{L}\left(h(z)-1\right)$$

$$= \frac{1}{\Delta t}\frac{z}{L}\left(\delta\mu\left(\frac{z}{L}\right)\right) \to x\mu(x) = xr\left(1-\frac{x}{k}\right)$$

$$\frac{1}{\Delta t}E[(\Delta X(t))^2|X(t)=x] = \frac{1}{\Delta t}E\left[\left(\frac{Z_{n+1}-Z_n}{L}\right)^2\middle|X(t)=\frac{z}{L}\right] = \frac{1}{\Delta t}\frac{1}{L^2}\left[zv(z)+\{z(h(z)-1)\}^2\right]$$

$$= \frac{1}{\Delta t}\frac{1}{L^2}\left[zd\left(\frac{z}{L}\right)+z^2\delta^2\left(\mu\left(\frac{z}{L}\right)\right)^2\right] = \frac{1}{\Delta t}x^2\delta^2(\mu(x))^2+\frac{1}{\Delta t}\frac{1}{L}xd(x)$$

$$= x^2\delta(\mu(x))^2+\frac{1}{\Delta t}(\Delta t)xd(x) \to xd(x) = x\beta \text{ (for example)}$$

and finally $\lim_{\Delta t\to 0}\dfrac{1}{\Delta t}E[(\Delta X(t))^j|X(t)=x] = O[(\Delta t)^{j/2-1}],\ j>2$

BP in random environments (Ludwig 1976, Keiding 1976)

$W_n =$ Random environmental fluctuations at time n such that $E(W_n) = 0$ and $V(W_n) = 1$, W_n indep. of Z_m, $m < n$.

BP in random environments (Ludwig 1976, Keiding 1976)

$W_n =$ Random environmental fluctuations at time n such that $E(W_n) = 0$ and $V(W_n) = 1$, W_n indep. of Z_m, $m < n$. Then, the offspring distribution is defined as

$$p_j(z, w) = P(B_i = j | Z_n = z, W_n = w).$$

BP in random environments (Ludwig 1976, Keiding 1976)

$W_n =$ Random environmental fluctuations at time n such that $E(W_n) = 0$ and $V(W_n) = 1$, W_n indep. of Z_m, $m < n$. Then, the offspring distribution is defined as

$$p_j(z, w) = P(B_i = j | Z_n = z, W_n = w).$$

Again, $h(z, w)$ and $v(z, w)$ are the conditional mean and variance of the offspring distribution. Assume that

BP in random environments (Ludwig 1976, Keiding 1976)

$W_n =$ Random environmental fluctuations at time n such that $E(W_n) = 0$ and $V(W_n) = 1$, W_n indep. of Z_m, $m < n$. Then, the offspring distribution is defined as

$$p_j(z, w) = P(B_i = j | Z_n = z, W_n = w).$$

Again, $h(z, w)$ and $v(z, w)$ are the conditional mean and variance of the offspring distribution. Assume that

- $E[v|Z_n = z] = E[E[v|Z_n = z, W_n]] = d\left(\frac{z}{L}\right)$, which is the expected value of the variance of the offspring distribution over the environmental process: demographic stochasticity.

BP in random environments (Ludwig 1976, Keiding 1976)

$W_n =$ Random environmental fluctuations at time n such that $E(W_n) = 0$ and $V(W_n) = 1$, W_n indep. of Z_m, $m < n$. Then, the offspring distribution is defined as

$$p_j(z, w) = P(B_i = j | Z_n = z, W_n = w).$$

Again, $h(z, w)$ and $v(z, w)$ are the conditional mean and variance of the offspring distribution. Assume that

- $E[v|Z_n = z] = E[E[v|Z_n = z, W_n]] = d\left(\frac{z}{L}\right)$, which is the expected value of the variance of the offspring distribution over the environmental process: demographic stochasticity.
- $h(z, w) = 1 + \delta\mu\left(\frac{z}{L}\right) + \sqrt{\delta e\left(\frac{z}{L}\right)}w$, where $\Delta t = \delta = \frac{1}{L}$, and the fluctuations due to the environment are of order $\sqrt{\delta}$ (a sum of a large number of iid random variables).

Diffusion approximation of a BPRE (Ludwig 1976, Keiding 1976)

Following the above conditions we get that

$$\begin{aligned} \lim_{\Delta t\to 0}\tfrac{1}{\Delta t}E[\Delta X(t)|X(t)=x] &= x\mu(x) = xr\left(1-\tfrac{x}{k}\right) \\ \lim_{\Delta t\to 0}\tfrac{1}{\Delta t}E[(\Delta X(t))^2|X(t)=x] &= xd(x)+x^2e(x) = x\beta + x^2\alpha \end{aligned}$$

Diffusion approximation of a BPRE (Ludwig 1976, Keiding 1976)

Following the above conditions we get that

$$\begin{aligned}
\lim_{\Delta t \to 0} \tfrac{1}{\Delta t} E[\Delta X(t) | X(t) = x] &= x\mu(x) = xr\left(1 - \tfrac{x}{k}\right) \\
\lim_{\Delta t \to 0} \tfrac{1}{\Delta t} E[(\Delta X(t))^2 | X(t) = x] &= xd(x) + x^2 e(x) = x\beta + x^2\alpha
\end{aligned}$$

Diffusion approximation of a BPRE (Ludwig 1976, Keiding 1976)

Following the above conditions we get that

$$\begin{aligned}
\lim_{\Delta t \to 0} \tfrac{1}{\Delta t} E[\Delta X(t) | X(t) = x] &= x\mu(x) = xr\left(1 - \tfrac{x}{k}\right) \\
\lim_{\Delta t \to 0} \tfrac{1}{\Delta t} E[(\Delta X(t))^2 | X(t) = x] &= xd(x) + x^2 e(x) = x\beta + x^2\alpha
\end{aligned}$$

- $x\beta =$ Expected value of the variance of the offspring distribution: on average, how much dos the offspring distribution varies.

Diffusion approximation of a BPRE (Ludwig 1976, Keiding 1976)

Following the above conditions we get that

$$\begin{aligned}\lim_{\Delta t \to 0} \tfrac{1}{\Delta t} E[\Delta X(t) | X(t) = x] &= x\mu(x) = xr\left(1 - \tfrac{x}{k}\right) \\ \lim_{\Delta t \to 0} \tfrac{1}{\Delta t} E[(\Delta X(t))^2 | X(t) = x] &= xd(x) + x^2 e(x) = x\beta + x^2\alpha\end{aligned}$$

- $x\beta =$ Expected value of the variance of the offspring distribution: on average, how much dos the offspring distribution varies.
- $x^2\alpha =$ the variance of the expected value of the offspring distribution: how much does the **mean** of the offspring distribution changes over time

Diffusion approximation of a BPRE (Ludwig 1976, Keiding 1976)

Following the above conditions we get that

$$\begin{aligned}
\lim_{\Delta t \to 0} \frac{1}{\Delta t} E[\Delta X(t) | X(t) = x] &= x\mu(x) = xr\left(1 - \frac{x}{k}\right) \\
\lim_{\Delta t \to 0} \frac{1}{\Delta t} E[(\Delta X(t))^2 | X(t) = x] &= xd(x) + x^2 e(x) = x\beta + x^2\alpha
\end{aligned}$$

- $x\beta =$ Expected value of the variance of the offspring distribution: on average, how much does the offspring distribution varies.
- $x^2\alpha =$ the variance of the expected value of the offspring distribution: how much does the **mean** of the offspring distribution changes over time
- References: Ludwig 1976, Keiding 1976, Braumann 1983 a,b, Dennis and Patil 1984, Dennis 1989, Goel and Richter-Dyn 1974, Turelli 1977 (random environment and stochastic calculus), Ethier and Kurtz 1986, Tier and Hanson 1982, Dennis 2002 (Allee effects with environmental and demographic fluctuations)

The likelihood for the logistic SDE with environmental stochasticity

Dennis 1989: shows an approximation to the time-dependent transition distribution using matching moments from a gamma distribution. The moments came from an approximation to the Backward equation using singular perturbation methods. Wiesak (1988, PhD U of Idaho, Math Dept.) gave a rigorous justification for this approximation. The time dependent transition is then used to write down the likelihood function without sampling error.

The likelihood for the logistic SDE with environmental stochasticity

Dennis 1989: shows an approximation to the time-dependent transition distribution using matching moments from a gamma distribution. The moments came from an approximation to the Backward equation using singular perturbation methods. Wiesak (1988, PhD U of Idaho, Math Dept.) gave a rigorous justification for this approximation. The time dependent transition is then used to write down the likelihood function without sampling error.
And with sampling error? We need to review some basic concepts first.

Brief review: Bayesian statistics

What is bayesianism? The subjectivist point of view

What is meant by a statement such as “the probability that this coin will land heads up is $\frac{1}{2}$”?

- The most common interpretation is that the **long run** frequency of heads approaches $\frac{1}{2}$.

What is bayesianism? The subjectivist point of view

What is meant by a statement such as “the probability that this coin will land heads up is $\frac{1}{2}$”?

- The most common interpretation is that the **long run** frequency of heads approaches $\frac{1}{2}$.
- Bayesians who make this statement mean that their prior opinion is such that they would as soon guess heads or tails.

What is bayesianism? The subjectivist point of view

What is meant by a statement such as "the probability that this coin will land heads up is $\frac{1}{2}$"?

- The most common interpretation is that the **long run** frequency of heads approaches $\frac{1}{2}$.
- Bayesians who make this statement mean that their prior opinion is such that they would as soon guess heads or tails.
- Consider a game in which if the event A occurs, the bayesian will be paid 1 \$. Then, $P(A)$ is the amount of money he would be willing to pay to buy into the game.

What is bayesianism? The subjectivist point of view

What is meant by a statement such as "the probability that this coin will land heads up is $\frac{1}{2}$"?

- The most common interpretation is that the **long run** frequency of heads approaches $\frac{1}{2}$.
- Bayesians who make this statement mean that their prior opinion is such that they would as soon guess heads or tails.
- Consider a game in which if the event A occurs, the bayesian will be paid 1 $. Then, $P(A)$ is the amount of money he would be willing to pay to buy into the game.
- This concept of probability is personal: $P(A)$ may vary from person to person.

What is bayesianism? The subjectivist point of view

What is meant by a statement such as “the probability that this coin will land heads up is $\frac{1}{2}$”?

- The most common interpretation is that the **long run** frequency of heads approaches $\frac{1}{2}$.
- Bayesians who make this statement mean that their prior opinion is such that they would as soon guess heads or tails.
- Consider a game in which if the event A occurs, the bayesian will be paid 1 \$. Then, $P(A)$ is the amount of money he would be willing to pay to buy into the game.
- This concept of probability is personal: $P(A)$ may vary from person to person.
- For bayesians, probability is a model for quantifying the strength of personal opinions.

What is bayesianism? The subjectivist point of view

What is meant by a statement such as “the probability that this coin will land heads up is $\frac{1}{2}$”?

- The most common interpretation is that the **long run** frequency of heads approaches $\frac{1}{2}$.
- Bayesians who make this statement mean that their prior opinion is such that they would as soon guess heads or tails.
- Consider a game in which if the event A occurs, the bayesian will be paid 1 \$. Then, $P(A)$ is the amount of money he would be willing to pay to buy into the game.
- This concept of probability is personal: $P(A)$ may vary from person to person.
- For bayesians, probability is a model for quantifying the strength of personal opinions.
- In bayesian inference, evidence is collected that is meant to be consistent or inconsistent with a given hypothesis and

What is bayesianism? The subjectivist point of view

What is meant by a statement such as "the probability that this coin will land heads up is $\frac{1}{2}$"?

- The most common interpretation is that the **long run** frequency of heads approaches $\frac{1}{2}$.
- Bayesians who make this statement mean that their prior opinion is such that they would as soon guess heads or tails.
- Consider a game in which if the event A occurs, the bayesian will be paid 1 \$. Then, $P(A)$ is the amount of money he would be willing to pay to buy into the game.
- This concept of probability is personal: $P(A)$ may vary from person to person.
- For bayesians, probability is a model for quantifying the strength of personal opinions.
- In bayesian inference, evidence is collected that is meant to be consistent or inconsistent with a given hypothesis and
- as evidence accumulates, the degree of belief in a hypothesis ought to change.

What is bayesianism? The subjectivist point of view

What is meant by a statement such as "the probability that this coin will land heads up is $\frac{1}{2}$"?

- The most common interpretation is that the **long run** frequency of heads approaches $\frac{1}{2}$.
- Bayesians who make this statement mean that their prior opinion is such that they would as soon guess heads or tails.
- Consider a game in which if the event A occurs, the bayesian will be paid 1 \$. Then, $P(A)$ is the amount of money he would be willing to pay to buy into the game.
- This concept of probability is personal: $P(A)$ may vary from person to person.
- For bayesians, probability is a model for quantifying the strength of personal opinions.
- In bayesian inference, evidence is collected that is meant to be consistent or inconsistent with a given hypothesis and
- as evidence accumulates, the degree of belief in a hypothesis ought to change.
- With enough evidence, it should become very high or very low.

Bayesian inference

Bayesian inference

- uses a numerical estimate of the degree of belief in a hypothesis before evidence has been observed and

Bayesian inference

Bayesian inference

- uses a numerical estimate of the degree of belief in a hypothesis before evidence has been observed and
- calculates a numerical estimate of the degree of belief in the hypothesis after evidence has been observed.

Bayesian inference

Bayesian inference

- uses a numerical estimate of the degree of belief in a hypothesis before evidence has been observed and
- calculates a numerical estimate of the degree of belief in the hypothesis after evidence has been observed.
- This process is repeated when additional evidence is obtained.

Bayesian inference

Bayesian inference

- uses a numerical estimate of the degree of belief in a hypothesis before evidence has been observed and
- calculates a numerical estimate of the degree of belief in the hypothesis after evidence has been observed.
- This process is repeated when additional evidence is obtained.
- Bayesian inference usually relies on degrees of belief, or subjective probabilities in the induction process.

Bayesian inference

Bayesian inference

- uses a numerical estimate of the degree of belief in a hypothesis before evidence has been observed and
- calculates a numerical estimate of the degree of belief in the hypothesis after evidence has been observed.
- This process is repeated when additional evidence is obtained.
- Bayesian inference usually relies on degrees of belief, or subjective probabilities in the induction process.

Suppose the prior prob. of A is $P(A)$. Upon observing event C, the opinion about A changes to $P(A|C)$:

$$P(A|C) = \frac{P(A \,\text{and}\, C)}{P(C)} = \frac{P(C|A)P(A)}{P(C)}.$$

Bayesian inference: the posterior distribution

$$P(A|C) = \frac{P(A\,\textsf{and}\,C)}{P(C)} = \frac{P(C|A)P(A)}{P(C)}.$$

Bayesian inference: the posterior distribution

$$P(A|C) = \frac{P(A \,\textsf{and}\, C)}{P(C)} = \frac{P(C|A)P(A)}{P(C)}.$$

- Bayesian's prior opinion about the value of θ is given by a density $g(\theta)$.

Bayesian inference: the posterior distribution

$$P(A|C) = \frac{P(A\,\text{and}\,C)}{P(C)} = \frac{P(C|A)P(A)}{P(C)}.$$

- Bayesian's prior opinion about the value of θ is given by a density $g(\theta)$.
- Having observed the data $\mathbf{X} = \mathbf{x}$, where $\mathbf{X}$ has density function $f(\mathbf{x}|\theta)$

Bayesian inference: the posterior distribution

$$P(A|C) = \frac{P(A\,\text{and}\,C)}{P(C)} = \frac{P(C|A)P(A)}{P(C)}.$$

- Bayesian's prior opinion about the value of θ is given by a density $g(\theta)$.
- Having observed the data $\mathbf{X} = \mathbf{x}$, where $\mathbf{X}$ has density function $f(\mathbf{x}|\theta)$, the new opinion about θ is

$$h(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)g(\theta)}{\int f(\mathbf{x}|\theta)g(\theta)d\theta} \propto f(\mathbf{x}|\theta)g(\theta).$$

Bayesian inference: the posterior distribution

$$P(A|C) = \frac{P(A \text{ and } C)}{P(C)} = \frac{P(C|A)P(A)}{P(C)}.$$

- Bayesian's prior opinion about the value of θ is given by a density $g(\theta)$.
- Having observed the data $\mathbf{X} = \mathbf{x}$, where $\mathbf{X}$ has density function $f(\mathbf{x}|\theta)$, the new opinion about θ is
$$h(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)g(\theta)}{\int f(\mathbf{x}|\theta)g(\theta)d\theta} \propto f(\mathbf{x}|\theta)g(\theta).$$
- Note that: the tool for making inference, the posterior distribution $h(\mathbf{x}|\theta)$ is defined in terms of the degree of belief about θ.

Bayesian inference: the posterior distribution

$$P(A|C) = \frac{P(A \text{ and } C)}{P(C)} = \frac{P(C|A)P(A)}{P(C)}.$$

- Bayesian's prior opinion about the value of θ is given by a density $g(\theta)$.

- Having observed the data $\mathbf{X} = \mathbf{x}$, where $\mathbf{X}$ has density function $f(\mathbf{x}|\theta)$, the new opinion about θ is

$$h(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)g(\theta)}{\int f(\mathbf{x}|\theta)g(\theta)d\theta} \propto f(\mathbf{x}|\theta)g(\theta).$$

- Note that: the tool for making inference, the posterior distribution $h(\mathbf{x}|\theta)$ is defined in terms of the degree of belief about θ.

- Sample space probabilities are no longer used, hypothetical repetitions of a random process are no longer considered.

Bayesian inference: the posterior distribution

$$P(A|C) = \frac{P(A \text{ and } C)}{P(C)} = \frac{P(C|A)P(A)}{P(C)}.$$

- Bayesian's prior opinion about the value of θ is given by a density $g(\theta)$.
- Having observed the data $\mathbf{X} = \mathbf{x}$, where $\mathbf{X}$ has density function $f(\mathbf{x}|\theta)$, the new opinion about θ is

$$h(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)g(\theta)}{\int f(\mathbf{x}|\theta)g(\theta)d\theta} \propto f(\mathbf{x}|\theta)g(\theta)$$

- Note that: the tool for making inference, the posterior distribution $h(\mathbf{x}|\theta)$ is defined in terms of the degree of belief about θ.
- Sample space probabilities are no longer used, hypothetical repetitions of a random process are no longer considered.
- Deciding between competing hypotheses in light of data reduces to compute their posterior probabilities.

An example: Binomial survival

Suppose we follow the fate of n individuals during a single time interval.

An example: Binomial survival

Suppose we follow the fate of n individuals during a single time interval. Let $X =$ number of survivors be binomially distributed:

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

An example: Binomial survival

Suppose we follow the fate of n individuals during a single time interval. Let $X =$ number of survivors be binomially distributed:

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Now, in the absence of previous information we may postulate that the prior distribution of p is uniform:

$$g(p) = 1, \quad 0 \leq p \leq 1.$$

An example: Binomial survival

Suppose we follow the fate of n individuals during a single time interval. Let $X =$ number of survivors be binomially distributed:

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Now, in the absence of previous information we may postulate that the prior distribution of p is uniform:

$$g(p) = 1, \quad 0 \leq p \leq 1.$$

Suppose $n = 20$ and $x = 5$. The posterior distribution is

$$h(p|x) \propto f(x|p)g(p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

An example: Binomial survival

Suppose we follow the fate of n individuals during a single time interval. Let $X =$ number of survivors be binomially distributed:

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Now, in the absence of previous information we may postulate that the prior distribution of p is uniform:

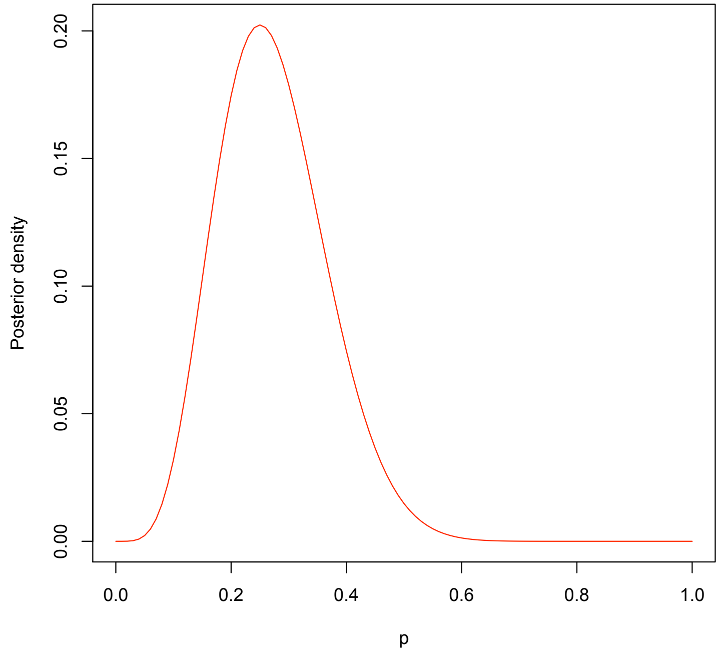
$$g(p) = 1, \quad 0 \leq p \leq 1.$$

Suppose $n = 20$ and $x = 5$. The posterior distribution is

$$h(p|x) \propto f(x|p)g(p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

This posterior distribution represents the new opinion of a Bayesian who was initially indifferent to the value of p *after* observing 5 survivals in 20 trials.

Posterior distribution for binomial survival:



The uniform prior expresses an indifference about the possible values of p, which is modified after the experiment.

Quantifying prior beliefs using a beta distribution

Using a prior beta distribution for p we get:

$$g(p) \propto p^{a-1}(1-p)^{b-1}$$

Quantifying prior beliefs using a beta distribution

Using a prior beta distribution for p we get:

$$g(p) \propto p^{a-1}(1-p)^{b-1}$$

and the posterior distribution of p given x is:

$$h(p|x) \propto f(x|p)g(p) = \binom{n}{x} p^x (1-p)^{n-x} p^{a-1}(1-p)^{b-1} \propto p^{a+x-1}(1-p)^{n+b-x-1},$$

a new beta distribution with parameters $a' = a + x$ and $b' = b + n - x$.

Quantifying prior beliefs using a beta distribution

Using a prior beta distribution for p we get:

$$g(p) \propto p^{a-1}(1-p)^{b-1}$$

and the posterior distribution of p given x is:

$$h(p|x) \propto f(x|p)g(p) = \binom{n}{x} p^x (1-p)^{n-x} p^{a-1}(1-p)^{b-1} \propto p^{a+x-1}(1-p)^{n+b-x-1},$$

a new beta distribution with parameters $a' = a + x$ and $b' = b + n - x$. Note that

$$\mu_{\text{prior}} = \tfrac{a}{a+b} \quad \text{whereas} \quad \mu_{post} \; = \; \tfrac{a'}{a'+b'}$$

Quantifying prior beliefs using a beta distribution

Using a prior beta distribution for p we get:

$$g(p) \propto p^{a-1}(1-p)^{b-1}$$

and the posterior distribution of p given x is:

$$h(p|x) \propto f(x|p)g(p) = \binom{n}{x} p^x (1-p)^{n-x} p^{a-1}(1-p)^{b-1} \propto p^{a+x-1}(1-p)^{n+b-x-1},$$

a new beta distribution with parameters $a' = a + x$ and $b' = b + n - x$. Note that

$$\mu_{\text{prior}} = \tfrac{a}{a+b} \quad \text{whereas} \quad \mu_{post} = \tfrac{a'}{a'+b'}$$

$$= \tfrac{a+x}{a+b+n} = \tfrac{a+b}{a+b+n}\left(\tfrac{a}{a+b}\right) + \tfrac{n}{a+b+n}\bar{x},$$

where $\bar{x} = x/n$ is the sample mean.

Quantifying prior beliefs using a beta distribution

Using a prior beta distribution for p we get:

$$g(p) \propto p^{a-1}(1-p)^{b-1}$$

and the posterior distribution of p given x is:

$$h(p|x) \propto f(x|p)g(p) = \binom{n}{x} p^x (1-p)^{n-x} p^{a-1}(1-p)^{b-1} \propto p^{a+x-1}(1-p)^{n+b-x-1},$$

a new beta distribution with parameters $a' = a + x$ and $b' = b + n - x$. Note that

$$\mu_{\text{prior}} = \tfrac{a}{a+b} \quad \text{whereas} \quad \mu_{post} = \tfrac{a'}{a'+b'}$$

$$= \tfrac{a+x}{a+b+n} = \tfrac{a+b}{a+b+n}\left(\tfrac{a}{a+b}\right) + \tfrac{n}{a+b+n}\bar{x},$$

where $\bar{x} = x/n$ is the sample mean.

The posterior mean is a weighted average of the prior mean and the sample mean!

Quantifying prior beliefs using a beta distribution

Using a prior beta distribution for p we get:

$$g(p) \propto p^{a-1}(1-p)^{b-1}$$

and the posterior distribution of p given x is:

$$h(p|x) \propto f(x|p)g(p) = \binom{n}{x} p^x(1-p)^{n-x}p^{a-1}(1-p)^{b-1} \propto p^{a+x-1}(1-p)^{n+b-x-1},$$

a new beta distribution with parameters $a' = a + x$ and $b' = b + n - x$. Note that

$$\mu_{\text{prior}} = \tfrac{a}{a+b} \quad \text{whereas} \quad \mu_{post} = \tfrac{a'}{a'+b'}$$

$$= \tfrac{a+x}{a+b+n} = \tfrac{a+b}{a+b+n}\left(\tfrac{a}{a+b}\right) + \tfrac{n}{a+b+n}\bar{x},$$

where $\bar{x} = x/n$ is the sample mean.

The posterior mean is a weighted average of the prior mean and the sample mean!

As n grows large, x/n approaches true p_0 value and μ_{post} approaches p_0 (and var. goes to 0).

Bayesian inference for more ‘realistic’ problems:

- Suppose that at time t, the survival probability x_t is drawn from a stochastic process $X_t(\theta)$ that lives between 0 and 1.

Bayesian inference for more ‘realistic’ problems:

- Suppose that at time t, the survival probability x_t is drawn from a stochastic process $X_t(\theta)$ that lives between 0 and 1.
- Furthermore, let us suppose that the form of distribution of $X_t(\theta)$ depends only on the previous realization $X_{t-1}(\theta) = x_{t-1}$ ($X_t(\theta)$ is a Markov process that you *do not observe*).

Bayesian inference for more ‘realistic’ problems:

- Suppose that at time t, the survival probability x_t is drawn from a stochastic process $X_t(\theta)$ that lives between 0 and 1.

- Furthermore, let us suppose that the form of distribution of $X_t(\theta)$ depends only on the previous realization $X_{t-1}(\theta) = x_{t-1}$ ($X_t(\theta)$ is a Markov process that you *do not observe*).

- Researcher interested in the *per unit of time* survival probability

Bayesian inference for more ‘realistic’ problems:

- Suppose that at time t, the survival probability x_t is drawn from a stochastic process $X_t(\theta)$ that lives between 0 and 1.
- Furthermore, let us suppose that the form of distribution of $X_t(\theta)$ depends only on the previous realization $X_{t-1}(\theta) = x_{t-1}$ ($X_t(\theta)$ is a Markov process that you *do not observe*).
- Researcher interested in the *per unit of time* survival probability
- At each time step, take a random sample of n_t individuals from the population at large, $t = 0, \ldots, k$ and record their fate during one time unit (*e.g.*: one-year olds survival).

Bayesian inference for more ‘realistic’ problems:

- Suppose that at time t, the survival probability x_t is drawn from a stochastic process $X_t(\theta)$ that lives between 0 and 1.
- Furthermore, let us suppose that the form of distribution of $X_t(\theta)$ depends only on the previous realization $X_{t-1}(\theta) = x_{t-1}$ ($X_t(\theta)$ is a Markov process that you *do not observe*).
- Researcher interested in the *per unit of time* survival probability
- At each time step, take a random sample of n_t individuals from the population at large, $t = 0, \ldots, k$ and record their fate during one time unit (*e.g.*: one-year olds survival).
- Data: pairs $(n_0, y_0), (n_1, y_1), \ldots, (n_k, y_k)$.

Stochastic time-varying survival

Unobserved Markov process $X_t(\theta) = f(X_{t-1}(\theta))$

Observations $(\mathbf{Y}|\mathbf{X} = \mathbf{x}) \sim \mathrm{Binom}\left(\mathbf{n}, \mathbf{x}\right)$

Stochastic time-varying survival

Unobserved Markov process $X_t(\theta) = f(X_{t-1}(\theta))$

Observations $(\mathbf{Y}|\mathbf{X} = \mathbf{x}) \sim \mathrm{Binom}\,(\mathbf{n}, \mathbf{x})$

$t = 0, 1, 2, \ldots, k$. The vector of unobserved trajectory of the survival process is $\mathbf{X}_t$.

Stochastic time-varying survival

Unobserved Markov process $X_t(\theta) = f(X_{t-1}(\theta))$

Observations $(\mathbf{Y}|\mathbf{X} = \mathbf{x}) \sim \mathrm{Binom}\,(\mathbf{n}, \mathbf{x})$

$t = 0, 1, 2, \ldots, k$. The vector of unobserved trajectory of the survival process is $\mathbf{X}_t$.

This defines a State-Space model or Hidden-Markov process

Stochastic time-varying survival

Unobserved Markov process $X_t(\theta) = f(X_{t-1}(\theta))$

Observations $(\mathbf{Y}|\mathbf{X} = \mathbf{x}) \sim \mathrm{Binom}\,(\mathbf{n}, \mathbf{x})$

$t = 0, 1, 2, \ldots, k$. The vector of unobserved trajectory of the survival process is $\mathbf{X}_t$.

This defines a State-Space model or Hidden-Markov process

The likelihood of a single replicated time series of observations is

$$L\,(\theta) = \int P(\mathbf{Y}|\mathbf{X})g(\mathbf{X};\theta)d\mathbf{X},$$

where $g(\mathbf{X};\theta)$ is the joint distribution of a trajectory $\mathbf{X}$ of the Markov Chain, starting at X_0.

Hierarchical models in Ecology

$$\begin{aligned} \mathbf{Y} &\sim f(\mathbf{y}|\mathbf{X}, \phi) \\ \mathbf{X} &\sim g(\mathbf{x}|\theta) \end{aligned}$$

it is known that the likelihood is

$$L(\theta, \phi) = \int f(\mathbf{y}|\mathbf{X}, \phi) g(\mathbf{x}; \theta) d\mathbf{X}.$$

A few examples include:

- Stochastic population models with added observation error (De Valpine and Hastings 2002, Clark and Bjornstad 2004, Newman et al. 2006, Dennis et al 2006)
- Stochastic models of species abundance distributions (Etienne and Olff 2005)
- Capture-recapture models with uncertain capture probabilities (George and Robert 1992)

Non-linear, non-Gaussian SSM

$$\begin{aligned}\mathbf{Y} &\sim f(\mathbf{y}|\mathbf{X},\phi)\\ \mathbf{X} &\sim g(\mathbf{x}|\theta)\end{aligned}$$

it is known that the likelihood is

$$L(\theta,\phi)=\int f(\mathbf{y}|\mathbf{X},\phi)g(\mathbf{x};\theta)d\mathbf{X}.$$

- Maximum likelihood was known to be very difficult for these models.

Non-linear, non-Gaussian SSM

$$\begin{aligned}\mathbf{Y} &\sim f(\mathbf{y}|\mathbf{X},\phi)\\ \mathbf{X} &\sim g(\mathbf{x}|\theta)\end{aligned}$$

it is known that the likelihood is

$$L(\theta,\phi)=\int f(\mathbf{y}|\mathbf{X},\phi)g(\mathbf{x};\theta)d\mathbf{X}.$$

- Maximum likelihood was known to be very difficult for these models.
- Bayesian solutions to the study of hierarchical population models were much easier to implement.

Non-linear, non-Gaussian SSM

$$\begin{aligned} \mathbf{Y} &\sim f(\mathbf{y}|\mathbf{X}, \phi) \\ \mathbf{X} &\sim g(\mathbf{x}|\theta) \end{aligned}$$

it is known that the likelihood is

$$L(\theta, \phi) = \int f(\mathbf{y}|\mathbf{X}, \phi) g(\mathbf{x}; \theta) d\mathbf{X}.$$

- Maximum likelihood was known to be very difficult for these models.
- Bayesian solutions to the study of hierarchical population models were much easier to implement.
- However, it can be very difficult to specify non-informative priors to do "objective bayesian statistics" for hierarchical models (Nancy Reid, Mexico, 2008):
  - Bayesian hierarchical Poisson models, (Gelman et al 2007)
  - Heinrich 2005, Proceedings of Phystat05 ($\mathrm{Poisson}\,(\epsilon s + b)$, s of interest, additional Poisson measurements of b and ϵ)
  - Bayesian probit regression (Jones 2008, Siddhartha and Chib 1984)

The Bayesian solution I

- Circumvents the problem of high dimensional integration
- Assumes (θ, ϕ) are random variables.
- Also assume that $\mathbf{X}$ are unknown and random
- Uses Bayes's rule and MCMC to sample from:

$$\pi(\theta, \phi, \mathbf{X}|\mathbf{y}) = \frac{f(\mathbf{y}|\mathbf{X}, \phi)g(\mathbf{x}|\theta)\pi(\theta, \phi)}{\int f(\mathbf{y}|\mathbf{X}, \phi)g(\mathbf{x}|\theta)\pi(\theta, \phi)d\mathbf{X}d\theta d\phi}$$

- The marginal posterior distribution $\pi(\theta, \phi|\mathbf{y})$ is obtained by integrating the above posterior over $\mathbf{X}$

The Bayesian solution II

Markov Chain Monte Carlo algorithms yield B independent samples from $\pi(\theta, \phi, \mathbf{X}|\mathbf{y})$:

$$\begin{array}{ccc}
\phi^{(1)} & \theta^{(1)} & \mathbf{X}^{(1)} \\
\phi^{(2)} & \theta^{(2)} & \mathbf{X}^{(2)} \\
\vdots & \vdots & \vdots \\
\phi^{(B)} & \theta^{(B)} & \mathbf{X}^{(B)}
\end{array}$$

The Bayesian solution II

Markov Chain Monte Carlo algorithms yield B independent samples from $\pi(\theta, \phi, \mathbf{X}|\mathbf{y})$:

$$\begin{array}{ccc} \phi^{(1)} & \theta^{(1)} & \mathbf{X}^{(1)} \\ \phi^{(2)} & \theta^{(2)} & \mathbf{X}^{(2)} \\ \vdots & \vdots & \vdots \\ \phi^{(B)} & \theta^{(B)} & \mathbf{X}^{(B)} \end{array}$$

The marginal posterior distribution $\pi(\theta, \phi|\mathbf{y})$ is simply obtained by discarding the $\mathbf{X}$ from

$$\{\phi^{(i)},\, \theta^{(i)},\, \mathbf{X}^{(i)}\}_{i=1}^{B},$$

leaving

$$\{\phi^{(i)},\, \theta^{(i)}\}_{i=1}^{B},$$

and no integration is needed.

The Bayesian solution II

Markov Chain Monte Carlo algorithms yield B independent samples from $\pi(\theta, \phi, \mathbf{X}|\mathbf{y})$:

$$\begin{array}{ccc} \phi^{(1)} & \theta^{(1)} & \mathbf{X}^{(1)} \\ \phi^{(2)} & \theta^{(2)} & \mathbf{X}^{(2)} \\ \vdots & \vdots & \vdots \\ \phi^{(B)} & \theta^{(B)} & \mathbf{X}^{(B)} \end{array}$$

The marginal posterior distribution $\pi(\theta, \phi|\mathbf{y})$ is simply obtained by discarding the $\mathbf{X}$ from

$$\{\phi^{(i)},\, \theta^{(i)},\, \mathbf{X}^{(i)}\}_{i=1}^{B},$$

leaving

$$\{\phi^{(i)},\, \theta^{(i)}\}_{i=1}^{B},$$

and no integration is needed. The mean values and variances of $\pi(\theta, \phi|\mathbf{y})$ are simply the mean values and variances of

$$\{\phi^{(i)},\, \theta^{(i)}\}_{i=1}^{B}.$$

The Metropolis Hastings algorithm

Purpose: to draw samples from a pdf $\pi(x)$.**How?** By implementing four steps:

The Metropolis Hastings algorithm

Purpose: to draw samples from a pdf $\pi(x)$.**How?** By implementing four steps:

1. If in x initially, propose a move from x to y from $q(x \to y)$.

The Metropolis Hastings algorithm

Purpose: to draw samples from a pdf $\pi(x)$.**How?** By implementing four steps:

1. If in x initially, propose a move from x to y from $q(x \to y)$.
2. Calculate the Hastings ratio:

$$a(x,y) = \min\left(1, \frac{\pi(y)q(y \to x)}{\pi(x)q(x \to y)}\right)$$

The Metropolis Hastings algorithm

Purpose: to draw samples from a pdf $\pi(x)$.**How?** By implementing four steps:

1. If in x initially, propose a move from x to y from $q(x \to y)$.
2. Calculate the Hastings ratio:
$$a(x, y) = \min\left(1, \frac{\pi(y)q(y \to x)}{\pi(x)q(x \to y)}\right)$$
3. Accept the move with probability $a(x, y)$. Else, return x.

The Metropolis Hastings algorithm

Purpose: to draw samples from a pdf $\pi(x)$.**How?** By implementing four steps:

1. If in x initially, propose a move from x to y from $q(x \to y)$.
2. Calculate the Hastings ratio:

$$a(x, y) = \min\left(1, \frac{\pi(y)q(y \to x)}{\pi(x)q(x \to y)}\right)$$

3. Accept the move with probability $a(x, y)$. Else, return x.
4. Repeat many times.

The Metropolis Hastings algorithm

Purpose: to draw samples from a pdf $\pi(x)$.**How?** By implementing four steps:

1. If in x initially, propose a move from x to y from $q(x \to y)$.
2. Calculate the Hastings ratio:
$$a(x, y) = \min\left(1, \frac{\pi(y)q(y \to x)}{\pi(x)q(x \to y)}\right)$$
3. Accept the move with probability $a(x, y)$. Else, return x.
4. Repeat many times.

The process of generating x's from those steps is a Markov Chain whose stationary pdf is guaranteed to be $\pi(x)$. So all you have to do is *wait long enough* to get the desired samples.

When is M-H. useful?

- When there's no closed expression for $\pi(x)$, yet the ratio $\pi(x)/\pi(y)$ has a closed expression.
- Example: our bayesian posterior:

$$\pi(\theta, \phi, \mathbf{X}|\mathbf{y}) = \frac{f(\mathbf{y}|\mathbf{X}, \phi)g(\mathbf{x}|\theta)\pi(\theta, \phi)}{\int f(\mathbf{y}|\mathbf{X}, \phi)g(\mathbf{x}|\theta)\pi(\theta, \phi)d\mathbf{X}d\theta d\phi}$$

and

$$\frac{\pi(\theta, \phi, \mathbf{X}|\mathbf{y})}{\pi(\theta', \phi', \mathbf{X}'|\mathbf{y})} = \frac{f(\mathbf{y}|\mathbf{X}, \phi)g(\mathbf{x}|\theta)\pi(\theta, \phi)}{f(\mathbf{y}|\mathbf{X}', \phi')g(\mathbf{x}'|\theta')\pi(\theta', \phi')}$$

Why it works? -Some informal heuristics...

- Let $r(x \to y)$ be the transition pdf of the Markov chain generated by M-H.
- Pick $y \neq x$ such that

$$\frac{\pi(y)q(y \to x)}{\pi(x)q(x \to y)} \leq 1, \quad \text{i.e} \quad a(x,y) = \min\left(1, \frac{\pi(y)q(y \to x)}{\pi(x)q(x \to y)}\right) \leq 1$$

Why it works? -Some informal heuristics...

- Let $r(x \to y)$ be the transition pdf of the Markov chain generated by M-H.
- Pick $y \neq x$ such that

$$\frac{\pi(y)q(y \to x)}{\pi(x)q(x \to y)} \leq 1, \quad \text{i.e} \quad a(x,y) = \min\left(1, \frac{\pi(y)q(y \to x)}{\pi(x)q(x \to y)}\right) \leq 1$$

Then:

$$\pi(x)r(x \to y) \; =$$

Why it works? -Some informal heuristics. . .

- Let $r(x \to y)$ be the transition pdf of the Markov chain generated by M-H.
- Pick $y \neq x$ such that

$$\frac{\pi(y)q(y \to x)}{\pi(x)q(x \to y)} \leq 1, \quad \text{i.e} \quad a(x,y) = \min\left(1, \frac{\pi(y)q(y \to x)}{\pi(x)q(x \to y)}\right) \leq 1$$

Then:

$$\pi(x)r(x \to y) \;=\; \pi(x)q(x \to y)a(x,y)$$

Why it works? -Some informal heuristics...

- Let $r(x \to y)$ be the transition pdf of the Markov chain generated by M-H.
- Pick $y \neq x$ such that

$$\frac{\pi(y)q(y \to x)}{\pi(x)q(x \to y)} \leq 1, \quad \text{i.e} \quad a(x,y) = \min\left(1, \frac{\pi(y)q(y \to x)}{\pi(x)q(x \to y)}\right) \leq 1$$

Then:

$$\begin{aligned} \pi(x)r(x \to y) &= \pi(x)q(x \to y)a(x,y) \\ &= \pi(x)q(x \to y)\frac{\pi(y)q(y \to x)}{\pi(x)q(x \to y)} \end{aligned}$$

Why it works? -Some informal heuristics...

- Let $r(x \to y)$ be the transition pdf of the Markov chain generated by M-H.
- Pick $y \neq x$ such that

$$\frac{\pi(y)q(y \to x)}{\pi(x)q(x \to y)} \leq 1, \quad \text{i.e} \quad a(x,y) = \min\left(1, \frac{\pi(y)q(y \to x)}{\pi(x)q(x \to y)}\right) \leq 1$$

Then:

$$\begin{aligned} \pi(x)r(x \to y) &= \pi(x)q(x \to y)a(x,y) \\ &= \pi(x)q(x \to y)\frac{\pi(y)q(y \to x)}{\pi(x)q(x \to y)} \\ &= \pi(y)q(y \to x) \end{aligned}$$

Why it works? -Some informal heuristics...

- Let $r(x \to y)$ be the transition pdf of the Markov chain generated by M-H.
- Pick $y \neq x$ such that

$$\frac{\pi(y)q(y \to x)}{\pi(x)q(x \to y)} \leq 1, \quad \text{i.e} \quad a(x,y) = \min\left(1, \frac{\pi(y)q(y \to x)}{\pi(x)q(x \to y)}\right) \leq 1$$

Then:

$$\begin{aligned}
\pi(x)r(x \to y) &= \pi(x)q(x \to y)a(x,y) \\
&= \pi(x)q(x \to y)\frac{\pi(y)q(y \to x)}{\pi(x)q(x \to y)} \\
&= \pi(y)q(y \to x) \\
&= \pi(y)q(y \to x)a(y,x), \text{ since } \quad a(x,y) \leq 1 \Rightarrow a(y,x) = 1
\end{aligned}$$

Why it works? -Some informal heuristics...

- Let $r(x \to y)$ be the transition pdf of the Markov chain generated by M-H.
- Pick $y \neq x$ such that

$$\frac{\pi(y)q(y \to x)}{\pi(x)q(x \to y)} \leq 1, \quad \text{i.e} \quad a(x,y) = \min\left(1, \frac{\pi(y)q(y \to x)}{\pi(x)q(x \to y)}\right) \leq 1$$

Then:

$$\begin{aligned}
\pi(x)r(x \to y) &= \pi(x)q(x \to y)a(x,y) \\
&= \pi(x)q(x \to y)\frac{\pi(y)q(y \to x)}{\pi(x)q(x \to y)} \\
&= \pi(y)q(y \to x) \\
&= \pi(y)q(y \to x)a(y,x), \text{ since} \quad a(x,y) \leq 1 \Rightarrow a(y,x) = 1 \\
&= \pi(y)r(y \to x), \quad \text{which is the detailed balance equation.}
\end{aligned}$$

The data cloning method heuristics (Lele et al. 2007) - Robert (1993):

Recall that for the general model

$$\begin{aligned} \mathbf{Y} &\sim f(\mathbf{y}|\mathbf{X},\phi) \\ \mathbf{X} &\sim g(\mathbf{x}|\theta) \end{aligned}$$

the likelihood is

$$L(\theta,\phi) = \int f(\mathbf{y}|\mathbf{X},\phi)g(\mathbf{x}|\theta)d\mathbf{X}$$

and

$$\pi^{(1)}(\theta,\phi|\mathbf{y}) = \frac{\left\{\int f(\mathbf{y}|\mathbf{X},\phi)g(\mathbf{x}|\theta)d\mathbf{X}\right\}\pi(\theta,\phi)}{h(\mathbf{y})}$$

where

$$h(\mathbf{y}) = \int f(\mathbf{y}|\mathbf{X},\phi)g(\mathbf{x}|\theta)\pi(\theta,\phi)d\mathbf{X}d\theta d\phi.$$

The data cloning method heuristics (Lele et al. 2007) - Robert (1993):

Substitute original posterior back again as a prior and keep doing that:

$$\begin{aligned}
\pi^{(2)}(\theta,\phi|\mathbf{y}) &= \frac{\left\{\int f(\mathbf{y}|\mathbf{X},\phi)g(\mathbf{x}|\theta)d\mathbf{X}\right\}\pi^{(1)}(\theta,\phi)}{h^{(2)}(\mathbf{y})} \\
&= \frac{\left\{\int f(\mathbf{y}|\mathbf{X},\phi)g(\mathbf{x}|\theta)d\mathbf{X}\right\}^2\pi(\theta,\phi)}{h^{(2)}(\mathbf{y})} \\
&= \frac{\{L(\theta,\phi)\}^2\pi(\theta,\phi)}{h^{(2)}(\mathbf{y})}
\end{aligned}$$

and continuing in this fashion:

$$\pi^{(k)}(\theta,\phi|\mathbf{y}) = \frac{\{L(\theta,\phi)\}^k\pi(\theta,\phi)}{h^{(k)}(\mathbf{y})}.$$

The data cloning method heuristics (Lele et al. 2007):

Let $(\hat{\theta}, \hat{\phi})$ be such that $L(\hat{\theta}, \hat{\phi}; \mathbf{y}) > L(\theta, \phi; \mathbf{y})$ for all (θ, ϕ). (MLE def.) Given that $\pi(\theta, \phi)$ is positive everywhere on the parameter space, as k grows large

$$\frac{\pi^{(k)}(\theta, \phi|\mathbf{y})}{\pi^{(k)}(\hat{\theta}, \hat{\phi}|\mathbf{y})} = \frac{[L(\theta, \phi; \mathbf{y})]^k}{\left[L(\hat{\theta}, \hat{\phi}; \mathbf{y})\right]^k} \rightarrow \begin{cases} 0 & \text{if} \quad (\theta, \phi) \neq (\hat{\theta}, \hat{\phi}) \\ 1 & \text{if} \quad (\theta, \phi) = (\hat{\theta}, \hat{\phi}) \end{cases}$$

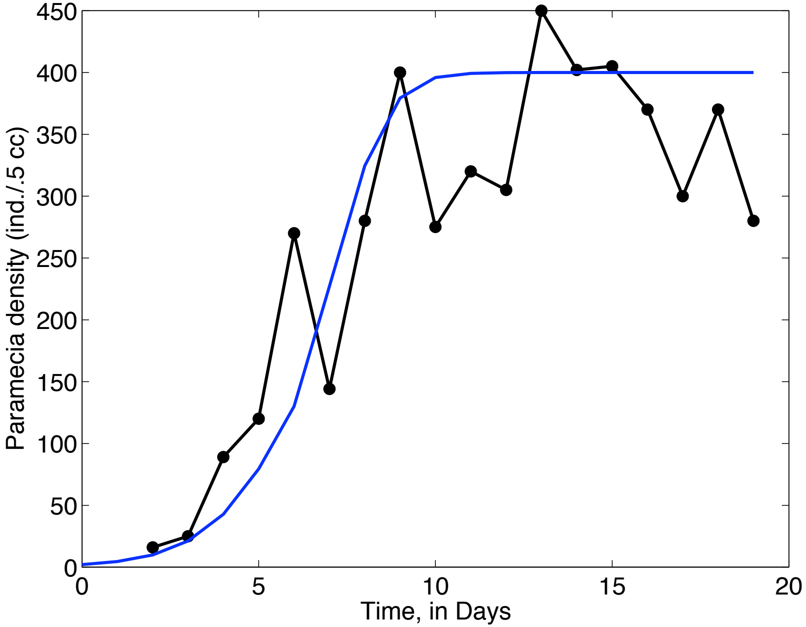
- That is, the fixed point for the iterated map is a degenerate distribution at $(\hat{\theta}, \hat{\phi})$ and independent of the initial distribution π.
- Fact: the mean of a degenerate distribution is the point at which it degenerates.
- So the mean of the k^{th} posterior distribution for large enough k approaches the MLE of (θ, ϕ).
- Finally, Lele's result: as $k \rightarrow \infty$, $\pi^{(k)}(\theta, \phi|\mathbf{y})$ converges to a $MVN([\hat{\theta}, \hat{\phi}]', \frac{1}{k}I^{-1}(\hat{\theta}, \hat{\phi}))$, where $I(\hat{\theta}, \hat{\phi})$ is the Fisher information matrix from the original likelihood function regardless of $\pi(\theta, \phi)$.

Testing Data Cloning: stochastic population growth

The Stochastic Gompertz Model (Dennis et al. 2006):

$$N_t = N_{t-1} \exp\left[a + b \ln N_{t-1} + \sigma Z_t\right] \quad \text{where} \quad Z_t \sim \text{ iid N}(0,1)$$

$$\text{and} \quad Y_t = \ln N_t + F_t \quad \text{where} \quad F_t \sim \text{ iid N}(0,\tau^2)$$



Testing Data Cloning: The stochastic Gompertz model with closed-form likelihood

$$N_t = N_{t-1} e^{[(a+b\ln(N_{t-1})+\sigma E_t]}$$

Let $x_t = \ln(n_t)$ and take $c = b+1$, then we have a first-order autoregressive process (Reddingius, 1971, Dennis and Taper 1994):

$$\begin{aligned} X_t &= X_{t-1} + a + bX_{t-1} + E_t \\ &= a + cX_{t-1} + E_t \end{aligned}$$

Density independence is expressed through $b = 0$ or $c = 1$. For $|c| < 1$ the stationary distribution exists and:

$$\begin{aligned} E[X_\infty] &= \lim_{t\to\infty} E[X_t] = \frac{a}{1-c} \\ Var[X_\infty] &= \lim_{t\to\infty} Var[X_t] = \frac{\sigma^2}{1-c^2} \end{aligned}$$

Stochastic Gompertz with observation error (Dennis et al 2006):

- Let Y_t be the estimated logarithmic population abundance, such that:

$$\begin{aligned} Y_t &= X_t + F_t \\ &= a + cX_{t-1} + E_t + F_t \\ &= a + c(Y_{t-1} - F_{t-1}) + E_t + F_t, \end{aligned}$$

  where $F_t \sim \mathrm{N}(0, \tau^2)$.

- The Markov property is lost: it is an ARMA model (Autorregresive Moving Average process).
- There is extra info. in the autocorrelation structure about σ^2 and τ^2.
- The ML parameter estimates are obtained via the Kalman filter (lots of conditioning) or using MVN:

The Multivariate Normal model:

No observation error: we have a series of recorded observations

$$x_0, x_1, \dots x_q.$$

Assuming X_0 arises from the stationary distribution, the joint pdf of $X_0, X_1, \dots X_q = \mathbf{X}$ has the following distribution:

$$\mathbf{X} \sim \mathbf{MVN}(\mu, \mathbf{\Sigma})$$

where

$$\Sigma = \frac{\sigma^2}{1-c^2} \begin{pmatrix} 1 & c & c^2 & \dots & c^q \\ c & 1 & c & \dots & c^{q-1} \\ c^2 & c & 1 & \dots & c^{q-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c^q & c^{q-1} & c^{q-2} & \dots & c \end{pmatrix}$$

and

$$\mu = \frac{a}{1-c}\mathbf{j},$$

$\mathbf{j}$ being a $(q+1) \times 1$ vector of ones.

The Multivariate Normal model:

With observation error: given the observations,$y_0, y_1, \ldots y_q$, the joint pdf of $Y_0, Y_1, \ldots Y_q$ is multivariate normal: writing $\mathbf{Y} = \mathbf{X} + \mathbf{F}$, we get

$$\mathbf{Y} \sim \mathbf{MVN}(\mu, \mathbf{V})$$

where $\mu = \frac{a}{1-c}\mathbf{j}$, $\mathbf{j}$ being a $(q+1) \times 1$ vector of ones, and $\mathbf{V} = \mathbf{\Sigma} + \tau^2\mathbf{I}$. The variance covariance matrix of the process is:

$$\mathbf{V} = \begin{bmatrix} \frac{\sigma^2}{1-c^2} + \tau^2 & \frac{c\sigma^2}{1-c^2} & \frac{c^2\sigma^2}{1-c^2} & \cdots & \frac{c^q\sigma^2}{1-c^2} \\ \frac{c\sigma^2}{1-c^2} & \frac{\sigma^2}{1-c^2} + \tau^2 & \frac{c\sigma^2}{1-c^2} & \cdots & \frac{c^{q-1}\sigma^2}{1-c^2} \\ \frac{c^2\sigma^2}{1-c^2} & \frac{c\sigma^2}{1-c^2} & \frac{\sigma^2}{1-c^2} + \tau^2 & \cdots & \frac{c^{q-2}\sigma^2}{1-c^2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{c^q\sigma^2}{1-c^2} & \frac{c^{q-1}\sigma^2}{1-c^2} & \frac{c^{q-2}\sigma^2}{1-c^2} & \cdots & \frac{\sigma^2}{1-c^2} + \tau^2 \end{bmatrix}.$$

Therefore, the log-likelihood needed for parameter estimation is:

$$\ln L(a, c, \sigma^2, \tau^2) = -\frac{q+1}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mu)'\mathbf{V}^{-1}(\mathbf{y} - \mu)$$

(First differences log-likelihood -REML- can also be obtained and behave nicely)

Testing Data Cloning: comparing with Gompertz SSM results

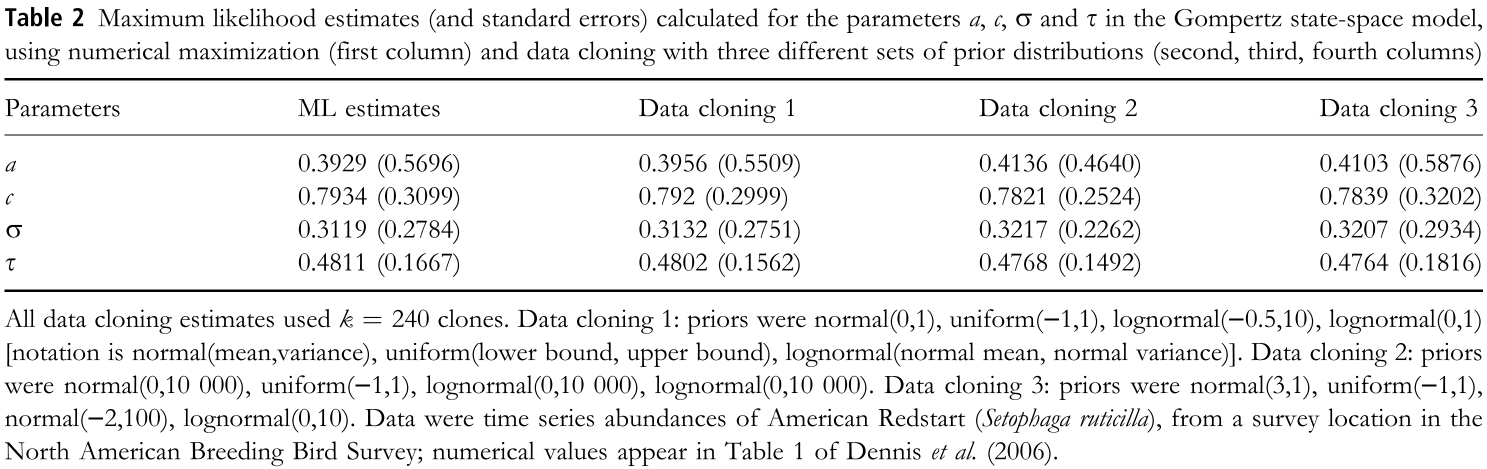


Table 2 Maximum likelihood estimates (and standard errors) calculated for the parameters *a*, *c*, σ and τ in the Gompertz state-space model, using numerical maximization (first column) and data cloning with three different sets of prior distributions (second, third, fourth columns)

Parameters	ML estimates	Data cloning 1	Data cloning 2	Data cloning 3
a	0.3929 (0.5696)	0.3956 (0.5509)	0.4136 (0.4640)	0.4103 (0.5876)
c	0.7934 (0.3099)	0.792 (0.2999)	0.7821 (0.2524)	0.7839 (0.3202)
σ	0.3119 (0.2784)	0.3132 (0.2751)	0.3217 (0.2262)	0.3207 (0.2934)
τ	0.4811 (0.1667)	0.4802 (0.1562)	0.4768 (0.1492)	0.4764 (0.1816)

All data cloning estimates used $k = 240$ clones. Data cloning 1: priors were normal(0,1), uniform(−1,1), lognormal(−0.5,10), lognormal(0,1) [notation is normal(mean,variance), uniform(lower bound, upper bound), lognormal(normal mean, normal variance)]. Data cloning 2: priors were normal(0,10 000), uniform(−1,1), lognormal(0,10 000), lognormal(0,10 000). Data cloning 3: priors were normal(3,1), uniform(−1,1), normal(−2,100), lognormal(0,10). Data were time series abundances of American Redstart (*Setophaga ruticilla*), from a survey location in the North American Breeding Bird Survey; numerical values appear in Table 1 of Dennis *et al.* (2006).

Lele et al. (2007)

Data Cloning continued

The method apparently

- Relies on asymptotic symmetric confidence intervals that can have coverage failures when using small data sets and are symmetric.
- Cannot easily get likelihood function evaluated at the maximum, so cannot:
  - Perform likelihood ratio tests
  - Draw profile likelihoods
  - Do model selection via information criteria (AIC, BIC, . . .)

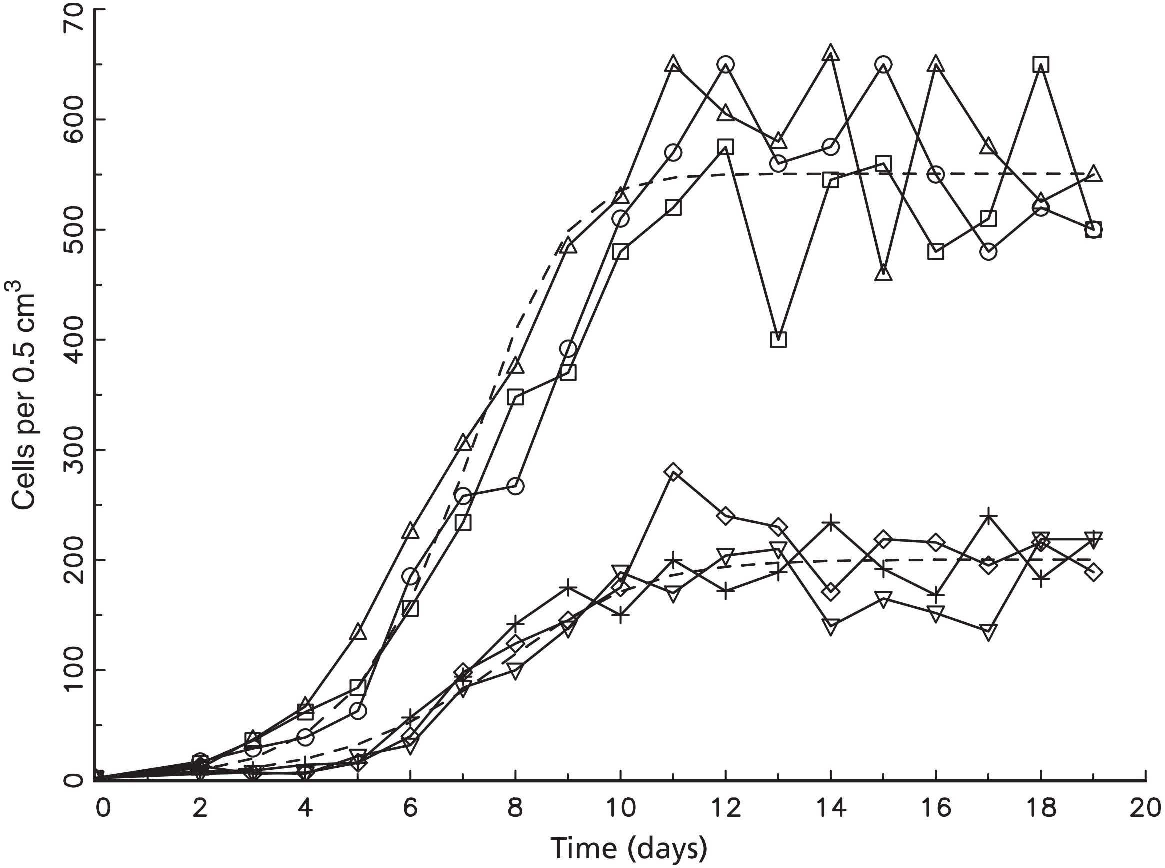
So cannot answer many scientific/biological questions !!!

Three problems of interest:

1. Drawing a profile likelihood (better CI's) seems computationally prohibitive (1 DC run + 1 MC integral at each value of the profiled parameter).
2. If DC does not yield the value of the maximum $L(\hat{\theta})$, how do we carry model selection via IC like AIC:
$$AIC_1 - AIC_2 = -2\ln\frac{\hat{L}_1}{\hat{L}_2} + 2(d_1 - d_2) \qquad = ?$$
3. If one posits that
$$H_0 : \theta_1 = \theta_2 = \theta_3 \quad \text{or that}$$
$$H_1 : \theta_1 \neq \theta_2 \neq \theta_3.$$
How do we compute the ratio of integrals
$$\Lambda = \frac{L_0(\hat{\theta})}{L_1(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)} = ?$$

To answer these questions all we need is to know how to compute likelihood ratios (an example next).

Back to Gause's 1934 data:



The latent variable model component

Let the log-population abundance be $X_t = \ln N_t$. Specify a family of models of the form:

$$X_t = m(X_{t-1}) + \sigma Z_t, \quad \text{where} \quad Z_t \sim \mathrm{N}(0,1).$$

Two forms of density-dependence are:

$$m(x) = \begin{cases} x + a - be^{x} \quad \text{(Ricker)} \\ x + \ln(\lambda) - \ln(1+\beta e^{x}) \quad \text{(Beverton-Holt)} \end{cases}$$

The joint distribution for a single time series of log-abundances $x_i,\ i = 0, \ldots, q$ is:

$$g(\mathbf{x}|\theta) = \prod_{t=1}^{q} (2\pi\sigma^2)^{-1/2} \mathrm{exp}\left(-\frac{(x_t - m(x_{t-1}))^2}{2\sigma^2} \right).$$

For the Ricker model $\theta = [a\, b\, \sigma^2]'$ and for the B-H $\theta = [\lambda\, \beta\, \sigma^2]'$.

The hierarchical model:

Sampling from well-mixed liquid cultures suggests using a Poisson sampling model:

$$f(\mathbf{y}|\mathbf{x}, \phi) = \prod_{t=2}^{q} \frac{e^{-n_t} n_t^{y_t}}{y_t!},$$

where $n_t = \exp(x_t)$. The distribution $f(\mathbf{y}|\mathbf{x}, \phi)$ will serve as the observation component in the Likelihood function for three population cultures:

$$L(\theta_1, \theta_2, \theta_3) = \prod_{j=1}^{3} \int f(\mathbf{y}_j|\mathbf{x}_j) g(\mathbf{x}_j|\theta_j) d\mathbf{x}_j.$$

Three problems of interest:

1. Drawing a profile likelihood (better CI's) seems computationally prohibitive (1 DC run $+$ 1 MC integral at each value of the profiled parameter).

2. If DC does not yield the value of the maximum $L(\hat{\theta})$, how do we carry model selection via IC like AIC:

$$AIC_1 - AIC_2 = -2\ln\frac{\hat{L}_1}{\hat{L}_2} + 2(d_1 - d_2) \qquad = ?$$

3. If one posits that

$$H_0 : \theta_1 = \theta_2 = \theta_3 \quad \text{or that}$$

$$H_1 : \theta_1 \neq \theta_2 \neq \theta_3.$$

   How do we compute the ratio of integrals

$$\Lambda = \frac{L_0(\hat{\theta})}{L_1(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)} = ?$$

To answer these questions all we need is to know how to compute likelihood ratios.

Likelihood ratios for data cloning:

Let $(\theta^{(0)}, \phi^{(0)})$ and $(\theta^{(1)}, \phi^{(1)})$ be two particular sets of parameter values. To compute

$$\frac{L(\theta^{(0)}, \phi^{(0)})}{L(\theta^{(1)}, \phi^{(1)})}, \quad \text{do:}$$

1. Generate m samples $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(m)}$ from the posterior of the latent variables via MCMC (straightforward):

$$h(\mathbf{x}|\mathbf{y}, \theta^{(1)}, \phi^{(1)}) \propto f(\mathbf{y}|\mathbf{x}, \phi^{(1)}) g(\mathbf{x}|\theta^{(1)})$$

2. Calculate the LR as:

$$\frac{L(\theta^{(0)}, \phi^{(0)})}{L(\theta^{(1)}, \phi^{(1)})} \approx \frac{1}{m} \sum_{j=1}^{m} \frac{f(\mathbf{y}|\mathbf{x}^{(j)}, \phi^{(0)}) g(\mathbf{x}^{(j)}|\theta^{(0)})}{f(\mathbf{y}|\mathbf{x}^{(j)}, \phi^{(1)}) g(\mathbf{x}^{(j)}|\theta^{(1)})}.$$

Profile likelihood for DC:

Let $\theta = [\theta_S, \theta_C]$. To draw profile for θ_S, do

1. Calculate ML estimates $(\hat{\theta}, \hat{\phi})$ using DC.
2. For θ_S select an array $\theta_S^{(1)}, \theta_S^{(2)}, \ldots, \theta_S^{(J)}$ bracketing the ML estimates broadly enough.
3. For each value $\theta_S^{(1)}, \theta_S^{(2)}, \ldots, \theta_S^{(J)}$ in turn, carry DC to maximize the likelihood w.r. to θ_C getting
$$\left(\{\hat{\theta}_C^{(1)}, \hat{\phi}^{(1)}\}, \{\hat{\theta}_C^{(2)}, \hat{\phi}^{(2)}\}, \ldots, \{\hat{\theta}_C^{(J)}, \hat{\phi}^{(J)}\}\right).$$
4. Generate m samples $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(m)}$ from $h(\mathbf{x}|\mathbf{y}, \hat{\theta}, \hat{\phi})$.
5. Then, for each $\theta_S^{(i)}, i = 1, 2, \ldots, J$ calculate the sample average:
$$\frac{L(\theta_S^{(i)}, \hat{\theta}_C^{(i)}, \hat{\phi}^{(i)})}{L(\hat{\theta}, \hat{\phi})} \approx \frac{1}{m} \sum_{j=1}^{m} \frac{f(\mathbf{y}|\mathbf{x}^{(j)}, \hat{\phi}^{(i)}) g(\mathbf{x}^{(j)}|\theta_S^{(i)}, \hat{\theta}_C^{(i)})}{f(\mathbf{y}|\mathbf{x}^{(j)}, \hat{\phi}) g(\mathbf{x}^{(j)}|\hat{\theta})}.$$

Use a single MCMC chain + vectorized calculations this algorithm is fast!

Why the LR and PL algorithms work? (E. Thompson, U.W., 1991)

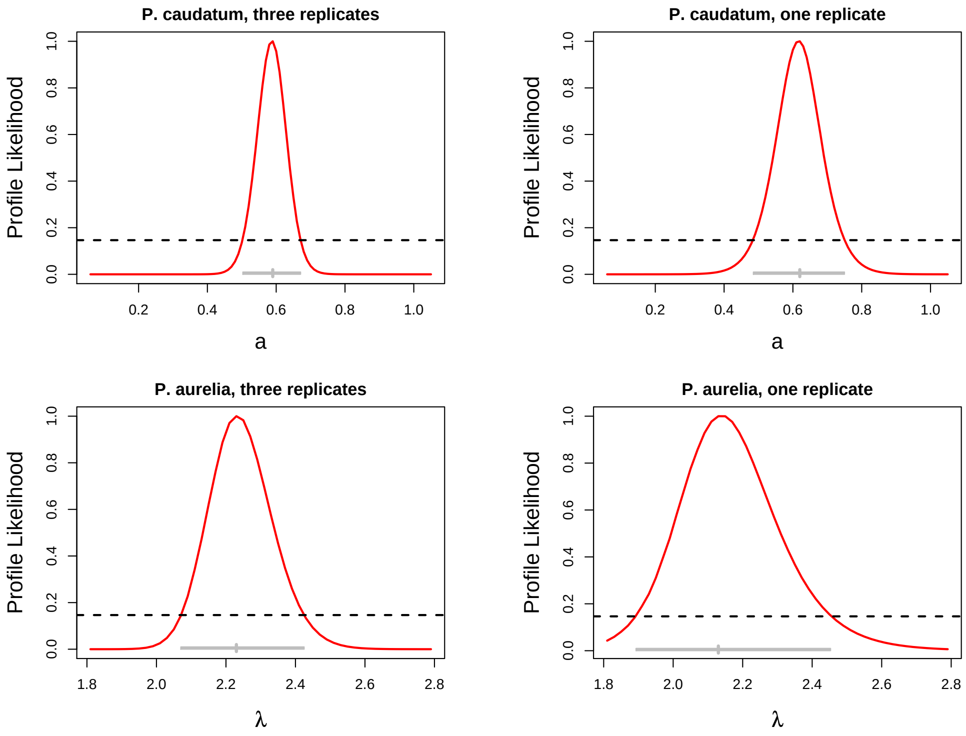
$$
\begin{aligned}
\frac{L(\theta,\phi,\mathbf{y})}{L(\theta_0,\phi_0,\mathbf{y})} &= \frac{L(\theta,\phi,\mathbf{y})}{L(\theta_0,\phi_0,\mathbf{y})} \int_{\mathbf{x}\in S} h(\mathbf{x}|\mathbf{y},\theta,\phi)d\mathbf{x} \\
&= \frac{L(\theta,\phi,\mathbf{y})}{L(\theta_0,\phi_0,\mathbf{y})} \int_{\mathbf{x}\in S} h(\mathbf{x}|\mathbf{y},\theta,\phi)\frac{h(\mathbf{x}|\mathbf{y},\theta_0,\phi_0)}{h(\mathbf{x}|\mathbf{y},\theta_0,\phi_0)}d\mathbf{x} \\
&= \int_{\mathbf{x}\in S} \frac{L(\theta,\phi,\mathbf{y})h(\mathbf{x}|\mathbf{y},\theta,\phi)}{L(\theta_0,\phi_0,\mathbf{y})h(\mathbf{x}|\mathbf{y},\theta_0,\phi_0)} h(\mathbf{x}|\mathbf{y},\theta_0,\phi_0)d\mathbf{x} \\
&= \int_{\mathbf{x}\in S} \frac{f(\mathbf{y},\mathbf{x}|\theta,\phi)}{f(\mathbf{y},\mathbf{x}|\theta_0,\phi_0)} h(\mathbf{x}|\mathbf{y},\theta_0,\phi_0)d\mathbf{x} \\
&= \int_{\mathbf{x}\in S} \frac{f(\mathbf{y}|\mathbf{x},\phi)g(\mathbf{x}|\theta)}{f(\mathbf{y}|\mathbf{x},\phi_0)g(\mathbf{x}|\theta_0)} h(\mathbf{x}|\mathbf{y},\theta_0,\phi_0)d\mathbf{x}
\end{aligned}
$$

and

$$
R.H.S. \approx \frac{1}{m}\sum_{j=1}^{m} \frac{f(\mathbf{y}|\mathbf{x}^{(j)},\phi)g(\mathbf{x}^{(j)}|\theta)}{f(\mathbf{y}|\mathbf{x}^{(j)},\phi_0)g(\mathbf{x}^{(j)}|\theta_0)}.
$$

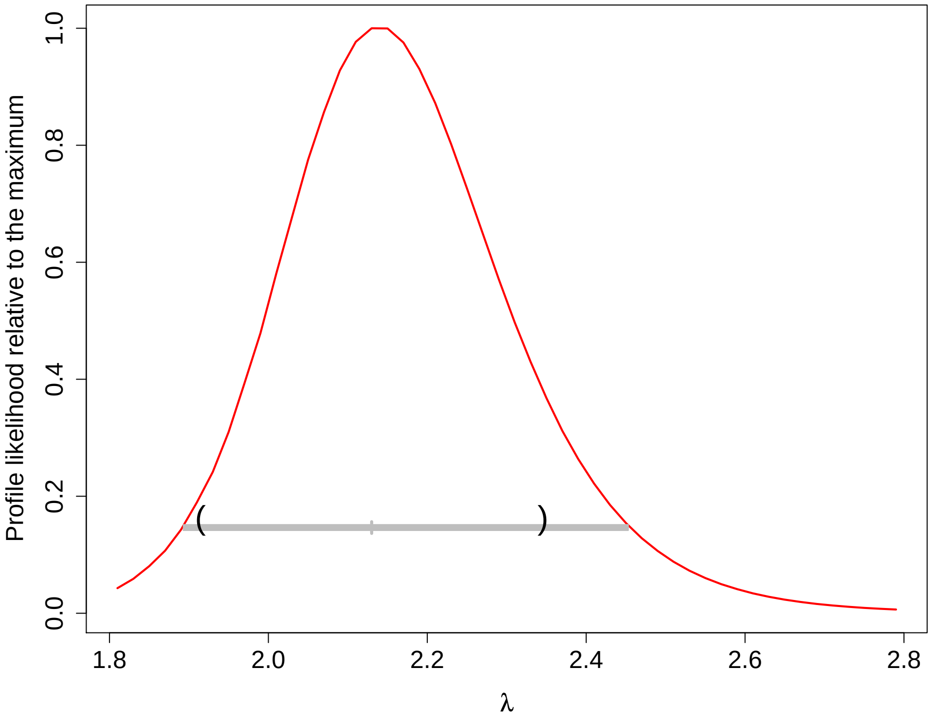
from which a likelihood profile can be easily computed using vectorized calculations.

Likelihood ratios for data cloning:



Ponciano, J.M., Taper, M.L., Dennis, B. and Lele, S.R. *in prep.* Inference for hierarchical models in ecology: confidence intervals, hypothesis testing and model selection using data cloning.

Lele et al CI's vs. 95% profile likelihood intervals:



Ponciano et al 2009 (Ecology).

Lele et al CI's vs. 15% likelihood intervals: some relevant conclusions

- For σ^2: if size of its variability is under estimated, the variability around the estimated probability of crossing a critical pop. threshold is under-estimated!!
- LRT for
$$H_0 : \theta_1 = \theta_2 = \theta_3 \quad \text{vs.}$$
$$H_1 : \theta_1 \neq \theta_2 \neq \theta_3$$
fails to reject H_0 so Gause's replicates likely arose under the same process.
- The stochastic Beverton-Holt model (2) explains the data better than the stochastic Ricker model (1) because
$$AIC_1 - AIC_2 = -2\ln\frac{\hat{L}_1}{\hat{L}_2} + 2(d_1 - d_2) = 3.7387,$$
thus a particular form of density-dependence seems to explain the data the best (scramble vs. contest intra-specific competition). Model selection process does NOT stop here!

Other current computer intensive methods

1. Filtering: sequential factorization of the likelihood (Newman et al 2009)
2. Gride-based methods (Kitagawa 1987, de Valpine and Hastings 2002) using numerical integration methods
3. Sequential Monte Carlo methods (Particle Filtering), Newman et al 2009, Liu 2001, Thomas et al 2005)
4. Importance sampling (E. Thompson 1994, UW, genetics of pedigrees, Donnelly, Nordborg and Joyce 2001 -Genetics- the coalescent process)
5. Monte Carlo Expectation Maximization (EM) algorithm (Liu 2001)
6. Iterated Filtering methodology: "plug-and-play" inference (Ionides, Breto and King 2006, Breto et al 2009), examples with SDE's.

Simulating Markov Chains and carrying inference

- Is there a way to take advantage of the ease with which simulations are done to carry statistical inference?

Simulating Markov Chains and carrying inference

- Is there a way to take advantage of the ease with which simulations are done to carry statistical inference?
- Yes, according to the so-called “Likelihood-free” inference methods (in a bit).

Simulating Markov Chains and carrying inference

- Is there a way to take advantage of the ease with which simulations are done to carry statistical inference?
- Yes, according to the so-called "Likelihood-free" inference methods (in a bit).
- Regardless of the method, don't try to pool a big Hidden-Markov model into a big, computer intensive bag!!
- ALWAYS seek ways to diagnose the inference methods and results, try to keep it simple!!.

MCMC without likelihoods

Marjoram, Molitor, Plagnol y Tavare 2003 PNAS 100:15324-15328. Objective: sampling from $f(\theta|D) \propto P(D|\theta)\pi(\theta)$

1. Now in θ
2. Propose a change to θ' according to $q(\theta \to \theta')$
3. Generate D' using θ'
4. If $D' = D$, go to next step, else return θ
5. Calculate
$$a(\theta, \theta') = \min\left(1, \frac{\pi(\theta')q(\theta' \to \theta)}{\pi(\theta)q(\theta \to \theta')}\right)$$
6. Accept θ' with probability $a(\theta, \theta')$, else return θ

Repeat until obtaining many (independent) samples from the posterior $f(\theta|D)$.

Why it works?

- Let $r(\theta \to \theta')$ be the transition pdf of the Markov Chain.
- Pick $\theta' \neq \theta$ such that

$$\frac{\pi(\theta')q(\theta' \to \theta)}{\pi(\theta)q(\theta \to \theta')} \leq 1, \quad \text{i.e.} \quad a(\theta, \theta') = \min\left(1, \frac{\pi(\theta')q(\theta' \to \theta)}{\pi(\theta)q(\theta \to \theta')}\right) \leq 1$$

Why it works?

- Let $r(\theta \to \theta')$ be the transition pdf of the Markov Chain.
- Pick $\theta' \neq \theta$ such that

$$\frac{\pi(\theta')q(\theta' \to \theta)}{\pi(\theta)q(\theta \to \theta')} \leq 1, \quad \text{i.e.} \quad a(\theta, \theta') = \min\left(1, \frac{\pi(\theta')q(\theta' \to \theta)}{\pi(\theta)q(\theta \to \theta')}\right) \leq 1$$

Then:

$$f(\theta|D)r(\theta \to \theta') \;=\; f(\theta|D)q(\theta \to \theta')P(D|\theta')a(\theta, \theta')$$

Why it works?

- Let $r(\theta \to \theta')$ be the transition pdf of the Markov Chain.
- Pick $\theta' \neq \theta$ such that

$$\frac{\pi(\theta')q(\theta' \to \theta)}{\pi(\theta)q(\theta \to \theta')} \leq 1, \quad \text{i.e.} \quad a(\theta, \theta') = \min\left(1, \frac{\pi(\theta')q(\theta' \to \theta)}{\pi(\theta)q(\theta \to \theta')}\right) \leq 1$$

Then:

$$\begin{aligned} f(\theta|D)r(\theta \to \theta') &= f(\theta|D)q(\theta \to \theta')P(D|\theta')a(\theta, \theta') \\ &= \tfrac{P(D|\theta)\pi(\theta)}{P(D)} q(\theta \to \theta')P(D|\theta')\tfrac{\pi(\theta')q(\theta' \to \theta)}{\pi(\theta)q(\theta \to \theta')} \end{aligned}$$

Why it works?

- Let $r(\theta \to \theta')$ be the transition pdf of the Markov Chain.
- Pick $\theta' \neq \theta$ such that

$$\frac{\pi(\theta')q(\theta' \to \theta)}{\pi(\theta)q(\theta \to \theta')} \leq 1, \quad \text{i.e.} \quad a(\theta, \theta') = \min\left(1, \frac{\pi(\theta')q(\theta' \to \theta)}{\pi(\theta)q(\theta \to \theta')}\right) \leq 1$$

Then:

$$\begin{aligned}
f(\theta|D)r(\theta \to \theta') &= f(\theta|D)q(\theta \to \theta')P(D|\theta')a(\theta, \theta') \\
&= \tfrac{P(D|\theta)\pi(\theta)}{P(D)}q(\theta \to \theta')P(D|\theta')\tfrac{\pi(\theta')q(\theta' \to \theta)}{\pi(\theta)q(\theta \to \theta')} \\
&= \tfrac{P(D|\theta')\pi(\theta')}{P(D)}q(\theta' \to \theta)P(D|\theta)
\end{aligned}$$

Why it works?

- Let $r(\theta \to \theta')$ be the transition pdf of the Markov Chain.
- Pick $\theta' \neq \theta$ such that

$$\frac{\pi(\theta')q(\theta' \to \theta)}{\pi(\theta)q(\theta \to \theta')} \leq 1, \quad \text{i.e.} \quad a(\theta,\theta') = \min\left(1, \frac{\pi(\theta')q(\theta' \to \theta)}{\pi(\theta)q(\theta \to \theta')}\right) \leq 1$$

Then:

$$\begin{aligned}
f(\theta|D)r(\theta \to \theta') &= f(\theta|D)q(\theta \to \theta')P(D|\theta')a(\theta,\theta') \\
&= \tfrac{P(D|\theta)\pi(\theta)}{P(D)}q(\theta \to \theta')P(D|\theta')\tfrac{\pi(\theta')q(\theta'\to\theta)}{\pi(\theta)q(\theta\to\theta')} \\
&= \tfrac{P(D|\theta')\pi(\theta')}{P(D)}q(\theta' \to \theta)P(D|\theta) \\
&= f(\theta'|D)q(\theta' \to \theta)P(D|\theta)a(\theta',\theta)
\end{aligned}$$

Why it works?

- Let $r(\theta \to \theta')$ be the transition pdf of the Markov Chain.
- Pick $\theta' \neq \theta$ such that

$$\frac{\pi(\theta')q(\theta' \to \theta)}{\pi(\theta)q(\theta \to \theta')} \leq 1, \quad \text{i.e.} \quad a(\theta, \theta') = \min\left(1, \frac{\pi(\theta')q(\theta' \to \theta)}{\pi(\theta)q(\theta \to \theta')}\right) \leq 1$$

Then:

$$\begin{aligned}
f(\theta|D)r(\theta \to \theta') &= f(\theta|D)q(\theta \to \theta')P(D|\theta')a(\theta, \theta') \\
&= \tfrac{P(D|\theta)\pi(\theta)}{P(D)}q(\theta \to \theta')P(D|\theta')\tfrac{\pi(\theta')q(\theta' \to \theta)}{\pi(\theta)q(\theta \to \theta')} \\
&= \tfrac{P(D|\theta')\pi(\theta')}{P(D)}q(\theta' \to \theta)P(D|\theta) \\
&= f(\theta'|D)q(\theta' \to \theta)P(D|\theta)a(\theta', \theta) \\
&= f(\theta'|D)r(\theta' \to \theta)
\end{aligned}$$

Practical version: ABC -methods

Marjoram, Molitor, Plagnol y Tavare 2003 PNAS 100:15324-15328

1. Now in θ
2. Propose a change to θ' according to $q(\theta \to \theta')$
3. Generate D' using θ'
4. If $\rho(D', D) \leq \epsilon$ go to the next step, else, return θ
5. Calculate
$$a(\theta, \theta') = \min\left(1, \frac{\pi(\theta')q(\theta' \to \theta)}{\pi(\theta)q(\theta \to \theta')}\right)$$
6. Accept θ' with probability $a(\theta, \theta')$, else return θ

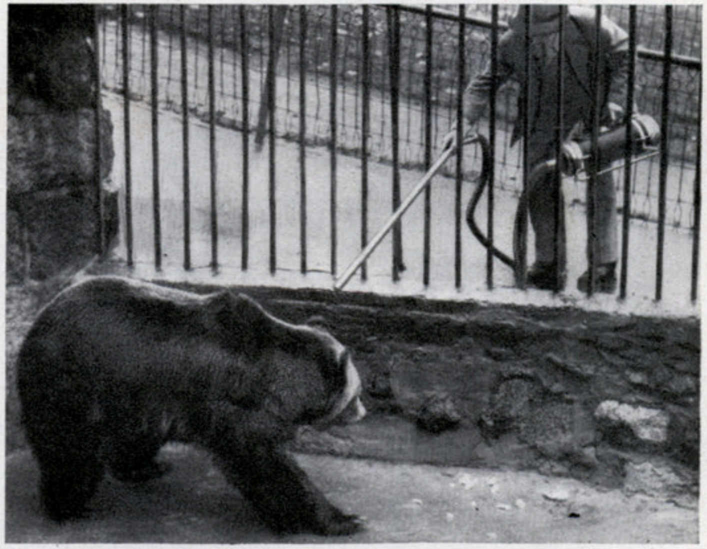
Obs come from $f(\theta|\rho(D'|D) \leq \epsilon)$. Variants: If S is a sufficient statistic, use: 4. If $\rho(S', S) \leq \epsilon$ go to next step. How do we find an approximately sufficient statistic? See Beaumont et al 2002 Genetics 162:20025-2-35, Joyce and Marjoram 2008.

General conclusions and future directions

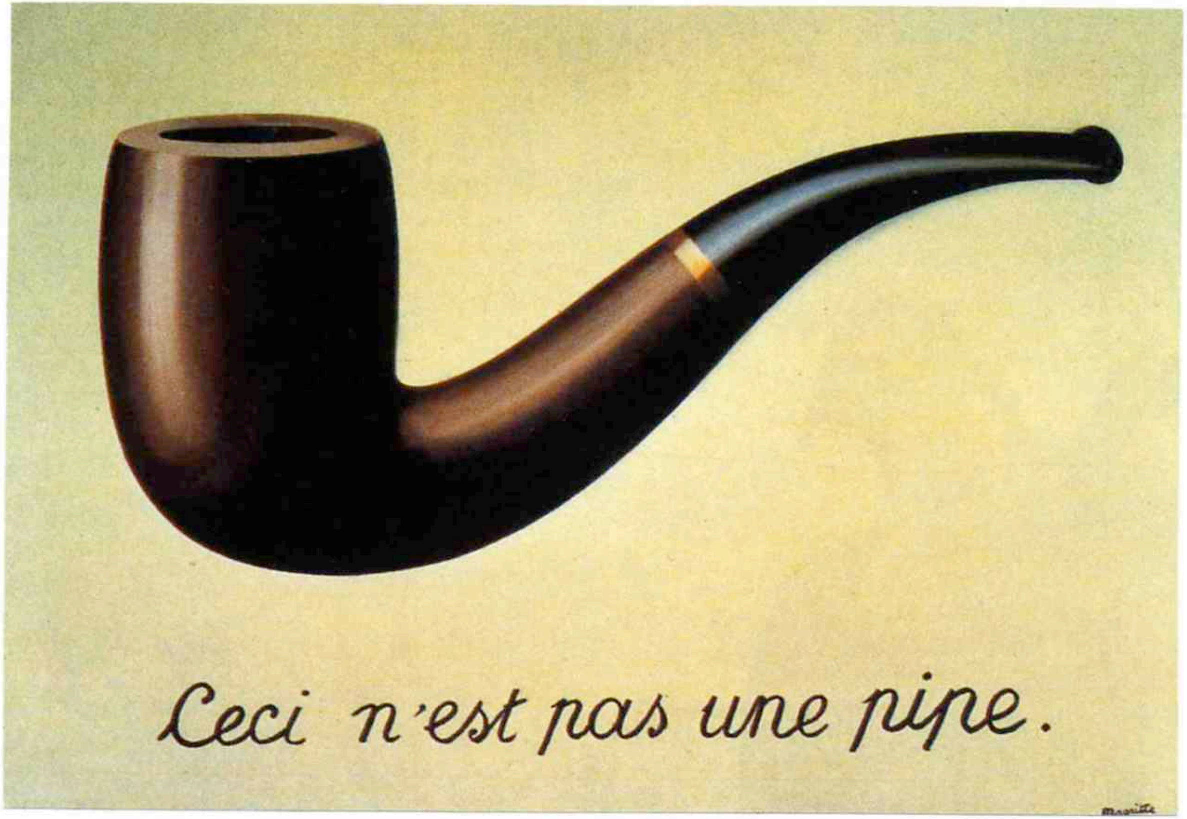
- A suite of methods for classical complete inference is now available for complex biological problems that are modeled using hierarchical statistical models.
- Note that the emphasis of the statistical Data Cloning method was shifted from a 'point-estimation-only method' to solving important biological questions via model selection, LRT, profile likelihood !!!
- The choice between Bayesian and frequentist approaches is not a matter of feasibility or convenience but rather can and should be based on the philosophical foundations of statistical inference preferred by the investigator.
- Using Data Cloning to test for parameter identifiability (as diagnostic tool)!

How To Gather Fleas from a Grizzly Bear

How to get fleas from a grizzly bear might puzzle a less resourceful man than Walt Sutter of Tacoma, Wash. From a radio program he learned that a wealthy Englishwoman was in the market for grizzly-bear fleas, to complete a collection taken from various wild animals. So he went to a zoo with a long-nozzled vacuum cleaner, and soon the coveted specimens were in the bag, ready for a purchaser.



Walt Sutter harvesting grizzly-bear fleas to sell to a flea collector


Ceci n'est pas une pipe.
Magritte