Data cloning
(a.k.a.)
How to trick Bayesians into giving frequentist answers

Subhash R Lele
(Joint work with Brian Dennis and Frithjof Lutscher)

WHY BOTHER?

Some simple examples:

1) **Gause's experiment with *paramecia*:** One of the standard models for population time series of the number of individuals, N_t, is the Stochastic Ricker model:

$$\log N_{t+1} - \log N_t = a + bN_t + \varepsilon_{t+1}$$

where $\varepsilon_t \sim N(0,\sigma^2)$ are independent random variables corresponding to the 'environmental' variation.

If we have the exact number of individuals at each time point, one can estimate the unknown parameters (a,b,σ^2) using the standard method of the maximum likelihood. These are obtained by maximizing the likelihood function:

$$L(a,b,\sigma^2;\underline{N}) = \prod_{t=1}^{T-1} f(\log N_{t+1} \mid \log N_t; a,b,\sigma^2)$$

This is pretty easy to write because we know that

$$f(\log N_{t+1} \mid \log N_t; a,b,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp\{\frac{-1}{2\sigma^2}(\log N_{t+1} - \log N_t - a - bN_t)^2\}$$

REALITY BITES!

- Gause did not count all the individuals in the test tube.
- He took a small portion of the medium and counted the number of individuals. He then multiplied the number to scale it to test tube.
- This is *an estimate* (indeed a good one) of the number of individuals in the test tube. As such there is some sampling error associated with this estimated number of individuals.

- Assuming that the individuals were distributed randomly throughout the medium, a reasonable sampling model is:

$$\hat{N}_t \mid N_t \sim Poisson(N_t)$$

Thus, the observed time series is NOT $\{N_1, N_2, ..., N_T\}$ but $\{\hat{N}_1, \hat{N}_2, ..., \hat{N}_T\}$.

The likelihood function, hence, should be written in terms of the joint distribution of $\{\hat{N}_1, \hat{N}_2, ..., \hat{N}_T\}$.

Following standard probability rules, we can write the joint distribution as:

$$f(\underline{\hat{N}}; a, b, \sigma^2) = \int f(\underline{\hat{N}} \mid \underline{N}) f(\underline{N}; a, b, \sigma^2) d\underline{N}$$

- This integral is over a T dimensional (=20 for Gause's data) space.
- The integral has no analytic form.
- We could try doing it numerically. But it has to be done for various values of the parameters (a, b, σ^2) to maximize the function.

Are there any takers for this task? Tomas, where are you when we need you?

Such a situation is not unique to Gause's data. In fact, these situations arise routinely when one tries to confront models with data.

1) **Compartmental models**: The standard epidemiological models such as the SIR model is an example of a compartmental model. If we want to fit the SIR model to real data, there are two issues that arise. One, we cannot observe the number of individuals in the S or R stages. Second, most of times we only have the number of reported cases, may be every two weeks.

The likelihood computation involves integrating over the unobserved values of the S and R stages and taking into account the sampling error in the reported cases.

2) **Latent variable models**: Suppose we are modeling population time series using the stochastic Ricker model but we have multiple spatial locations (ignore the dispersal for the time being).

$$\log N_{s,t+1} - \log N_{s,t} = a_s + bN_{s,t} + \varepsilon_{s,t+1}$$

It is reasonable to assume that 'a_s', the growth rate, is different at each spatial location. But, then we have too many parameters as compared to the data. This is especially true if we have large number of spatial locations but shorter time series at each location.

A way to get around this difficult situation is by assuming that the growth rates in nearby locations are similar. For example, we assume that $\underline{a} \sim MN(a\underline{1}, V)$.

On top of this, we also have the same kind of sampling

variability issues to deal with. For example, the population abundances of insects are often estimated using pheromone traps or light traps. Similar to Gause's data, these lead to Poisson sampling error.

Writing this model systematically:

Hierarchy 1: $\underline{a} \sim MN(a\underline{1}, V)$

Hierarchy 2: $\log \underline{N}_{t+1} \mid \log \underline{N}_t; \underline{a}, b, \sigma^2 \sim MN(\log \underline{N}_t + a + b\underline{N}_t, \sigma^2 I)$

Hierarchy 3: $\hat{\underline{N}}_t \mid \underline{N}_t \sim Poisson(\underline{N}_t)$

If we have 200 spatial locations and 10 time points, computation of the likelihood function would require 200x10 dimensional integration.

Unfortunately for the scientists, such situations are ubiquitous. One of my students did a quick search on the web of science. In the last five years, he found out that there were nearly 4000 papers that used such models. These were papers only in the ecology related journals.

Likelihood based statistical inference has a number of attractive properties but simplicity is clearly not one of them, especially for such hierarchical models.

So what should we do?

Statistical inference paradigms: A brief introduction

1) The frequentist paradigm:

- The parameters are considered '*fixed but unknown*' quantities.
- Given the data, we apply inductive inference to infer about the value of these unknown quantities.
- The uncertainty of our knowledge is defined in terms of 'replicability' of the inferential statements. If someone else conducts another experiment, applies the method, how different would the conclusions be? The quantities such as standard errors, confidence intervals quantify the replicability of the inferential statements.
- The missing data, latent variables are considered random variables that need to be integrated over.

2) *The Bayesian paradigm:*

- All unknown quantities (parameters, missing data, latent variables etc.) are considered random variables.

- The probability distribution for these unknown quantities quantifies the uncertainty of our knowledge about them. This is called the 'prior distribution'. This is the *belief* that the researcher has about the plausibility of different values that θ takes before any data are observed. For the notational simplicity, and for the time being, let us denote all unknown quantities by θ. We denote the prior distribution by $\pi(\theta)$.
- The data are denoted by $\underline{y}$. These data, in order to be informative about the unknown quantities θ, should be related to them in some fashion. It is assumed that the distribution of $\underline{y}$ depends on the value of θ. This is denoted by $f(\underline{y} \mid \underline{\theta})$.
- In the light of the data, how should we change our prior beliefs? The rule is simple:

$$\pi(\theta \mid \underline{y}) = \frac{f(\underline{y} \mid \theta)\pi(\theta)}{\int f(\underline{y} \mid \theta)\pi(\theta)d\theta}$$

This is called the 'posterior distribution' and is the conditional distribution of θ given the data $\underline{y}$. This distribution quantifies the post-data belief about the parameter values.

- The posterior distribution is guaranteed to have smaller variation than the prior distribution. Thus, observing the data decreases the uncertainty about the 'unknown' quantities.

Of course, if we start with perfect certainty about the parameter values (GOD created this earth 6000 years ago with probability

one), no amount of data (evidence) is going to change such prior beliefs. To change strongly held beliefs, one will need substantial amount of data (evidence).

The posterior distribution, thus, quantifies the *change in the belief* and does not answer the questions: What do the data say about the unknown quantities? How replicable are the inferential statements?

Saving grace? (pun intended)

As the sample size increases, even the strongly held beliefs (except those with degenerate priors) are swamped by the data. Eventually, even the Bayesians learn!

$$\pi(\theta \mid \underline{y}) \sim N(\hat{\theta}_{MLE}, \frac{1}{n} I^{-1}(\hat{\theta}_{MLE}))$$

How would the Bayesian paradigm deal with hierarchical models?

Hierarchical models

Hierarchy 1: (Parameters)

$$\theta \sim \pi(\theta)$$

Hierarchy 2: (Missing data, latent variables, unobserved states, random effects etc.)

$$X \mid \theta \sim f(x \mid \theta)$$

Hierarchy 3: (Observed data)

$$Y \mid X \sim g(y \mid x)$$

Compute the posterior distribution

$$\pi(\theta, x \mid y) = \frac{g(y \mid x) f(x \mid \theta)\pi(\theta)}{\int g(y \mid x) f(x \mid \theta)\pi(\theta)dxd\theta}$$

The problem is solved!

BUT, BUT, BUT …

1) The integral in the denominator is even higher dimensional than for the likelihood computation! We have added dimensions corresponding to θ.

2) If we want to compute marginal posterior distribution of θ, we will need to compute

$$\pi(\theta \mid y) = \int \pi(\theta, x \mid y)dx$$

It seems that the Bayesian paradigm leads to even harder problem than the computation of the likelihood function. But there is a very clever solution to this problem. (Bayesians are technically smart, just misguided)

Markov Chain Monte Carlo (MCMC) method:

- Metropolis et al. (1954?, Journal of Chemical Physics) and Hastings (1970, Biometrika) are the two breakthrough papers in the development of MCMC. These papers provide a computationally simple method to generate random numbers from the posterior distribution

$$\pi(\theta, x \mid y) = \frac{g(y \mid x) f(x \mid \theta)\pi(\theta)}{\int g(y \mid x) f(x \mid \theta)\pi(\theta)\, dx d\theta}$$

without computing the integral in the denominator!

- Given random numbers $(\theta_i, x_i), i = 1,2,\ldots.,B$, under this distribution, we can easily obtain the random numbers from $\pi(\theta \mid y)$, by simply dropping the ‘x’ component and considering $\theta_i, i = 1,2,\ldots.,B$.
- Given these random numbers, we can compute mean, variance, quantiles and any such relevant information about the posterior beliefs about the parameters.

Thus, it may seem that we have achieved the holy grail of making inferential statements about the unknown parameters in the hierarchical models without ever having to integrate!

Unfortunately, for those who like to do 'evidence based science' as against 'belief based science', while appreciating the cleverness and technical beauty of this solution, it still remains an unsatisfactory solution. Can we trick this Bayesian solution to give likelihood-based inference?

The method

number______________________

expiration____________________

just $199.95 (plus shipping and handling)

We call it: **data cloning**

Imaginary sequence of experiments:

We will describe the case where there are no random effects. The paper has a proof that is general.

Experimenter 1: First person to do the experiment, obtains data $\underline{y}$. He puts his prior $\pi(\theta)$ and conducts the Bayesian inference. The posterior is given by

$$\pi^{(1)}(\theta \mid \underline{y}) = \frac{f(y \mid \theta)\pi(\theta)}{\int f(y \mid \theta)\pi(\theta) d\theta}.$$

Experimenter 2: Second person does the same experiment, and by golly, he obtains exactly the same data $\underline{y}$ as the first experimenter (it can happen). Being a good scientist, he uses $\pi^{(1)}(\theta \mid \underline{y})$ as his prior and conducts the Bayesian inference. The posterior is

$$\pi^{(2)}(\theta \mid \underline{y}) = \frac{f(\underline{y} \mid \theta)\pi^{(1)}(\theta \mid \underline{y})}{\int f(\underline{y} \mid \theta)\pi^{(1)}(\theta \mid \underline{y})d\theta}$$

$$= \frac{f(\underline{y} \mid \theta) f(y \mid \theta)\pi(\theta)}{\int f(\underline{y} \mid \theta) f(y \mid \theta)\pi(\theta)d\theta}$$

$$= \frac{\{f(\underline{y} \mid \theta)\}^2 \pi(\theta)}{\int \{f(\underline{y} \mid \theta)\}^2 \pi(\theta)d\theta}$$

Continuing in this fashion …

Experimenter K: K-th researcher does the same experiment, and by golly, he obtains exactly the same data $\underline{y}$ as the first experimenter (it can happen). Being a good scientist, he uses $\pi^{(K-1)}(\theta \mid \underline{y})$ as his prior and conducts the Bayesian inference. The posterior is given by

$$\pi^{(K)}(\theta \mid \underline{y}) = \frac{f(\underline{y} \mid \theta)\pi^{(K-1)}(\theta \mid \underline{y})}{\int f(\underline{y} \mid \theta)\pi^{(K-1)}(\theta \mid \underline{y})d\theta}$$

$$= \frac{\{f(\underline{y} \mid \theta)\}^K \pi(\theta)}{\int \{f(\underline{y} \mid \theta)\}^K \pi(\theta)d\theta}$$

Now let us see what happens this posterior as K increases. Let $\hat{\theta}$ denote the MLE, that is,

$$f(y \mid \hat{\theta}) > f(y \mid \theta) \text{ for all } \theta \neq \hat{\theta}$$

It is easy to see that:

1) $\dfrac{\pi^{(K)}(\theta \mid y)}{\pi^{(K)}(\hat{\theta} \mid y)} = \dfrac{\{f(y \mid \theta)\}^K}{\{f(y \mid \hat{\theta})\}^K} \dfrac{\pi(\theta)}{\pi(\hat{\theta})} \to 0 \text{ if } \theta \neq \hat{\theta}$

2) $\dfrac{\pi^{(K)}(\hat{\theta} \mid y)}{\pi^{(K)}(\hat{\theta} \mid y)} = \dfrac{\{f(y \mid \hat{\theta})\}^K}{\{f(y \mid \hat{\theta})\}^K} \dfrac{\pi(\hat{\theta})}{\pi(\hat{\theta})} = 1$

Hence, as we increase K, the posterior distribution becomes degenerate at the MLE.

In fact, more can be shown:

- $E_{\pi^{(K)}}(\theta \mid y) \to \hat{\theta}$
- $K * Var_{\pi^{(K)}}(\theta \mid y) \to I^{-1}(\hat{\theta})$ where $I(\hat{\theta})$ is the Fisher information, inverse of it is the asymptotic variance of the MLE.

Data cloning in a nutshell:

1) Select some prior for the parameters.
2) Clone the data K times.
3) Apply MCMC to compute the posterior.
4) The mean of the posterior is the MLE and variance of the posterior is the asymptotic variance of the MLE.
5) We have managed to optimize the likelihood without ever evaluating the likelihood function or differentiating it. We have obtained the second derivative (Fisher information) without ever differentiating the function as well.

6) This method is a generalization of the simulated annealing method to random effects case. Hence, it is a general optimization method and is not restricted to likelihood optimization.

Example 1

Gompertz state space model of population growth: a test case with known likelihood function

X_t: log-population abundance (unobserved) at time t

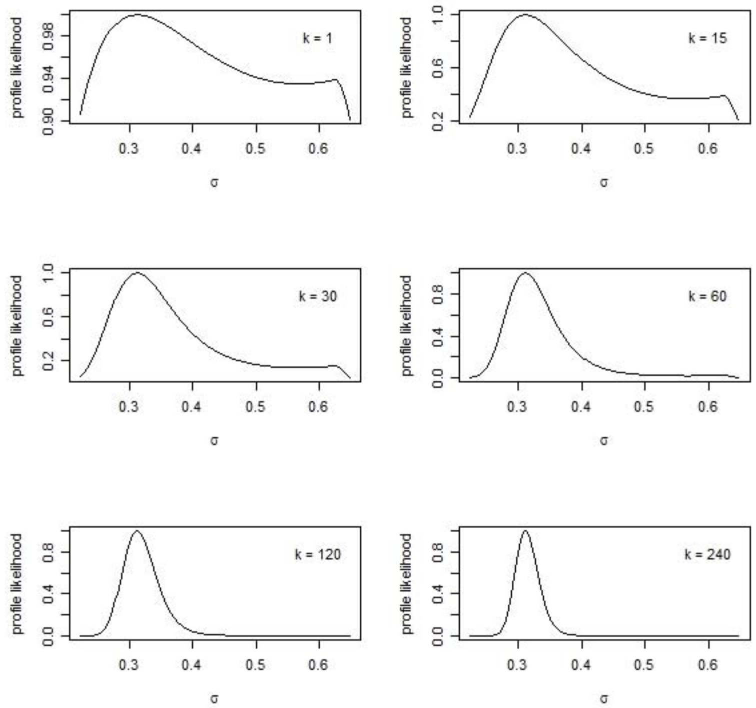
Y_t: estimated value of X_t (observed)

$$X_t = a + cX_{t-1} + E_t$$

$$Y_t = X_t + F_t$$

$E_t \sim$ normal(0, σ^2), $F_t \sim$ normal(0, τ^2)

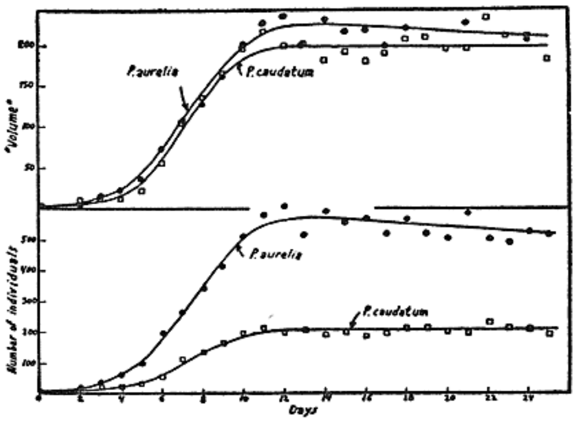
Likelihood function for the unknown parameters a, c, σ^2, τ^2 is a multivariate normal distribution; it can be decomposed (into a product of univariate normals) with a set of recursion equations known as the Kalman filter (see Dennis et al. 2006 *Ecol. Monogr.*)



Gompertz state-space model fitted to American Redstart time series (BBS): ML, data cloning (three different sets of prior distributions).

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)

Example 2: Gause's *Paramecia*: two species cultured separately (& together in competition)



The iconic, mandatory "S-shaped growth curve" data, plotted in every ecology textbook

Features of Gause's data

Gause's figure plots *means* at each time of three replicate cultures! (Original data in appendix to his book)

0.5 cc of well-stirred culture media *sampled* each unit of time

Intrinsic stochastic process noise in the cultures as well as sampling error

Some missing data (populations at $t = 1$ not sampled)

Ricker-Poisson state space model for Gause's data

N_t: concentration of unobserved population (cells per 0.5 cc)
O_t: sample concentration (cells per 0.5 cc)
t: time (days)

$$N_t = N_{t-1}\exp(a + bN_{t-1} + E_t)$$
$$E_t \sim \text{normal}(0, \sigma^2)$$

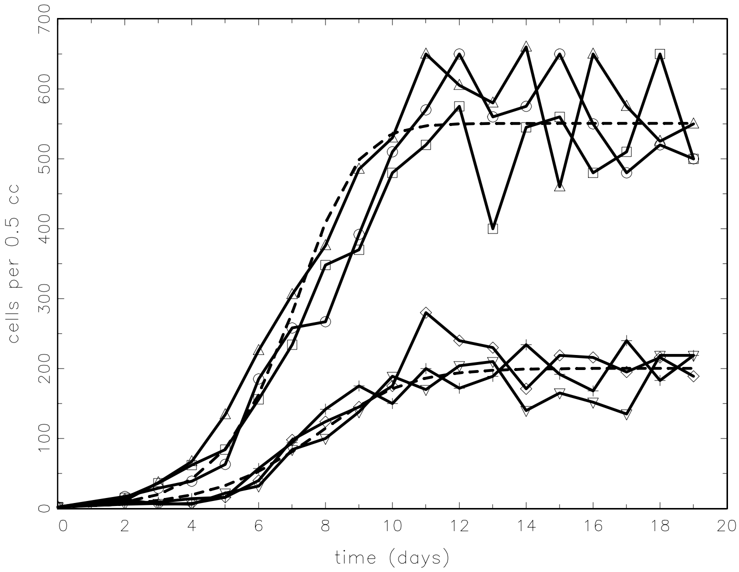
(a stochastic Ricker model)

$$O_t \sim \text{Poisson}(N_t)$$

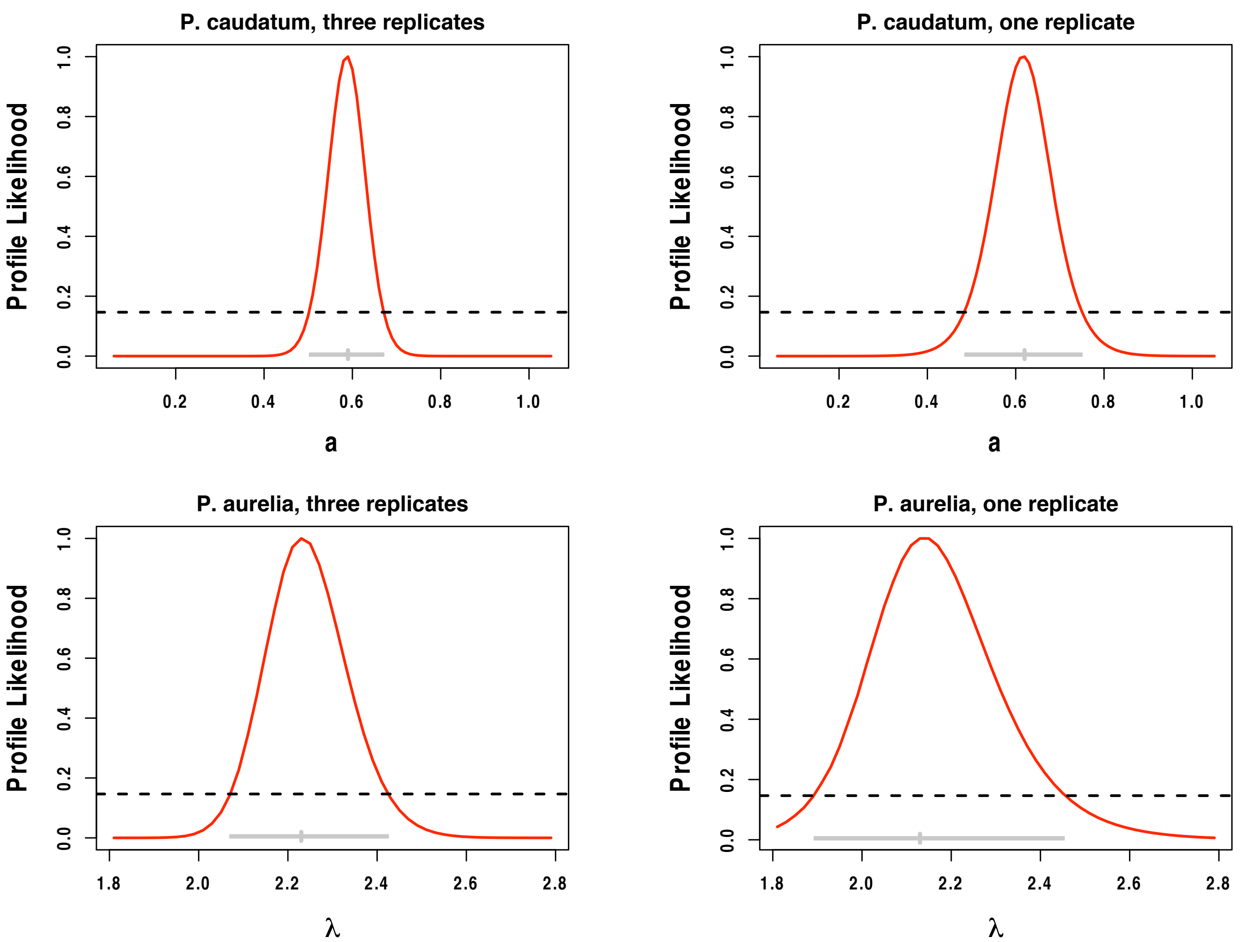
The likelihood function for data arising from this model is not available in closed form

Ricker and Ricker-Poisson models fitted to Gause's *Paramecium* data, combined replicates.

	P. aurelia		*P. caudatum*	
	Ricker	Ricker-Poisson	Ricker	Ricker-Poisson
a	0.686	0.771 (0.057)	0.529	0.581 (0.064)
b	-0.0013	-0.0014 (0.0001)	-0.0026	-0.0029 (0.0004)
σ	0.174	0.139 (0.031)	0.339	0.162 (0.044)



Population abundances of two *Paramecium* species, three replicate cultures each (solid lines), from Gause (1934: Appendix I, Table 3), plotted with solution trajectories from deterministic Ricker population growth model (dashed lines). Upper three time series: *P. aurelia*. Lower three time series: *P. caudatum*.



Lele, S. R., Dennis, B., and Lutscher, F. 2007. Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecology Letters* 10:551-563.

brian@uidaho.edu