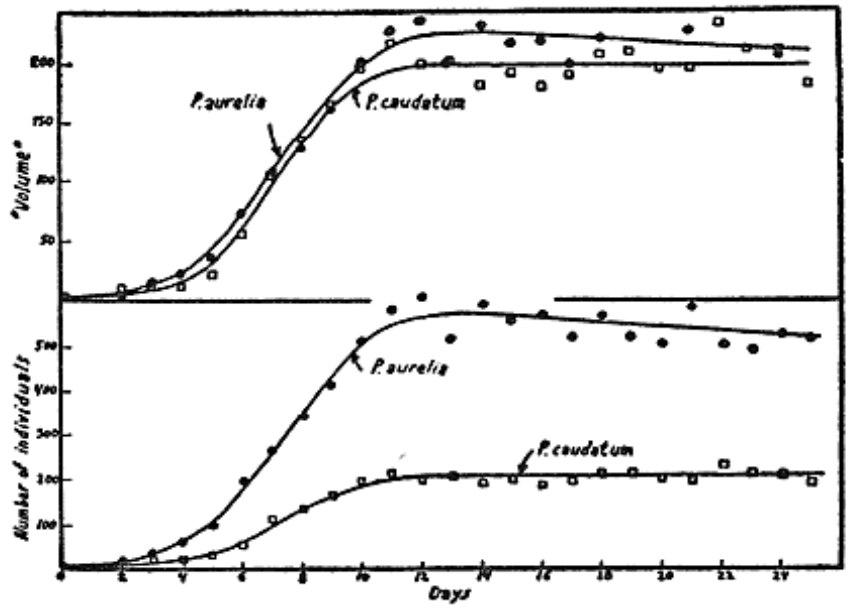
Data cloning
a.k.a.
How to trick Bayesians into giving frequentist answers

Subhash R. Lele
Department of Mathematical and Statistical Sciences
University of Alberta
slele@ualberta.ca
(Joint work with Brian Dennis and Fritjof Lutscher)
Ecology Letters, 2007

Why bother?

- Gause's data on population growth of two species of paramacia

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



Features of Gause's data

- Gause's figure plots *means* at each time of three replicate cultures!
- 0.5 cc of well-stirred culture media *sampled* each unit of time
- Intrinsic stochastic process noise in the cultures
- sampling error
- Some missing data (populations at time point 1 not sampled)

Mathematical Models for population growth:

Density dependent Logistic growth model: Continuous time

$$\frac{1}{N_t}\frac{dN_t}{dt} = a + bN_t$$

Discrete time version: Ricker model

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

Discrete time stochastic Ricker model

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

where $\varepsilon_t \sim N(0, \sigma^2)$ independent errors indicating environmental variation.

An alternative to Ricker model is Gompertz model:

$$\log N_{t+1} - \log N_t = a + b\log N_t + \varepsilon_t$$

where $\varepsilon_t \sim N(0, \sigma^2)$ independent errors indicating environmental variation.

This is simply an AR(1) model.

The likelihood function:

If the number of individuals in the population are observed without any sampling error, the likelihood function for this process is very easy to write:

$$L(a,b,\sigma^2;\underline{N}) = \prod_{t=0}^{T-1} f(\log N_{t+1} \mid \log N_t; a,b,\sigma^2)$$

This is easy for the Gompertz model (AR(1) process Likelihood). For the Ricker model, recall that:

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

This is quite easy. BUT REALITY BITES …

- Gause did not count all the individuals
- He took a small portion of the medium and counted the number of individuals. He then multiplied the number to scale it to the test tube.
- This is an estimate (indeed a good one) of total number of individuals in the test tube. As such there is some sampling error associated with this estimate. One needs to take that into account when doing the estimation.

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

$$\{N_0, N_1, ..., N_T\}$$

But it is: $\{N_0, \hat{N}_1, ..., \hat{N}_T\}$

The likelihood function should be written as:

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

- **This involves T (T=20 for Gause data) dimensional integration.**
- **This integral has no analytical form.**

Possible solutions:

1) Numerical integration: Difficult, if not impossible
2) EM algorithm: Expectation step is not easy.
3) MCEM or MCNR: Computationally difficult
4) Particle filter: Okay for first order
5) Kitagawa's algorithm: Okay for first order

COMPOSITE LIKELIHOOD?

Two features that make use of pairwise CL feasible:

(Gompertz model)

1) Two dimensional marginals for the state part of the model is easy: Gaussian distribution
2) Conditionally on the states, the observations are independent of each other.

Problems with CL: (Ricker model)

Gaussian state distribution does NOT hold for the Ricker model. Low dimensional marginal distributions are not easy to compute analytically.

Unfortunately, these situations are not uncommon in ecological and epidemiological models. The low dimensional marginals are not easy to compute either analytically or numerically.

The models for delayed density dependence are useful to model insect populations that are prone to host-parasite interactions. These higher order Markov processes for the states.

SIR and other compartmental models for spread of disease do not have nice low dimensional marginal distributions.

Generalized linear mixed models tend to assume the Gaussian random effects. In those cases, one can apply composite likelihood Ideas to simplify the computation.

However, there is no reason to assume that the random effects or latent variables are Gaussian. For example, in animal breeding or evolutionary genetics many researchers think that non-Gaussian random effects are quite likely.

If the hidden states or latent variables are such that the low dimensional marginal distributions are not analytical, applying the composite likelihood method for hierarchical models is difficult if not impossible.

Confession:

I am an inherently lazy person. So I searched for some method that will give me good answers in these tricky situations.

I hate to admit this but as much as the Bayesian Philosophy of science is a 'Faustian bargain', Bayesian *methods* are quite clever (and, great for a lazy and not so smart statistician).

Note to self:
Change to the pdf file

Generalized Linear Mixed Models:

1) Logistic-Normal model
2) Poisson-Log Normal model
3) Spatial count data model

1) **Logistic-Normal Mixed Model**: Crowder (1978, table 3) presented data on the proportion of seeds that germinated on each of 21 plates arranged according to a 2 x 2 factorial layout by seed variety and type of root extract. He noted that the within-group variation exceeded that predicted by binomial sampling theory. A natural way to account for extraneous plate-to-plate variability in this situation is by means of the following GLMM:

Hierarchy 1: $Y_i \mid p_i \sim Binomial\ (n_i, p_i)$ where

$$\log \frac{p_i}{1-p_i} = \alpha_0 + \alpha_{seed} + \alpha_{extract} + \alpha_{\text{int}\,eraction} + b_i$$

Hierarchy 2: $b_i \sim N(0, \sigma_b^2)$

Table 1: Maximum likelihood estimation for Logistic-Normal model for overdispersed binary data

Parameters	MLE - NI	MLE-P1	MLE-P2	MLE-P3
α_0	-0.546 (0.167)	-0.5484(0.168)	-0.548 (0.166)	-0.5472 (0.167)
α_1	0.097 (0.278)	0.0951(0.277)	0.09547(0.273)	0.0946 (0.283)
α_2	1.337 (0.237)	1.338(0.240)	1.336 (0.239)	1.335 (0.238)
α_{12}	-0.811 (0.385)	-0.8109 (0.382)	-0.8090 (0.379)	-0.8069 (0.393)
σ	0.236(0.110)	0.2392 (0.107)	0.2396 (0.109)	0.2408 (0.111)

2) **Longitudinal data:** Thall and Vail (1990, table 2) presented data from a clinical trial of 59 epileptics who were randomized to a new drug (Trt=1) or a placebo (Trt=0) as an adjuvant to the standard chemotherapy. Baseline data available at entry into the trial included the number of epileptic seizures recorded in the preceding 8-week period and age in years. The logarithm of the fourth of the number of baseline seizures (Base) and the logarithm of age (AGE) were treated as covariates in the analysis. A multivariate response variable consisted of the counts of seizures during the 2-weeks before each of four clinic visits (Visit, coded -3,-1,1,and 3). Preliminary analysis indicated that the counts were substantially lower during the fourth visit and a binary variable (V4 =1 for fourth visit, 0 otherwise) was constructed to model such effects. Breslow and Clayton (1993) use the following GLMM for modeling these data.

Hierarchy 1: $Y_i \mid \mu_i \sim Poisson(\mu_i)$ where

$$\log \mu_{jk} = \alpha_0 + \alpha_{AGE} AGE + \alpha_{BASE} BASE + \alpha_{Trt} Trt + \alpha_{BT} (BASE * Trt) + \alpha_{V4} V4 + b_j + b_{jk}$$

Hierarchy 2: $b_j \sim N(0, \sigma_b^2)$ and $b_{jk} \sim N(0, \sigma_{b1}^2)$

Table 2: Maximum likelihood estimation for Poisson-Normal model for repeated counts data

Parameters	PQL	MLE-P1	MLE-P2	MLE-P3
α_{AGE}	0.47 (0.35)	0.477 (0.342)	0.481 (0.352)	0.479 (0.367)
α_{BT}	0.34 (0.21)	0.346 (0.180)	0.343 (0.203)	0.346 (0.223)
α_{BASE}	0.86 (0.13)	0.876 (0.128)	0.881 (0.136)	0.880 (0.140)
α_{Trt}	-0.93 (0.40)	-0.941(0.364)	-0.933 (0.400)	-0.941 (0.433)
α_{V4}	-0.10 (0.09)	-0.102 (0.086)	-0.102 (0.087)	-0.100 (0.086)
α_0	-1.27 (1.2)	-1.366 (1.180)	-1.386 (1.210)	-1.380 (1.177)
σ_b	0.36 (0.04)	0.366 (0.043)	0.360 (0.044)	0.365(0.043)
σ_{b1}	0.48 (0.06)	0.471 (0.063)	0.465 (0.063)	0.469 (0.063)

Hierarchy 1: $Y_i \mid \mu_i \sim Poisson(\mu_i)$

Hierarchy 2: $\log \mu_i = \log e_i + \alpha_0 + \alpha_1 \frac{x_i}{10} + b_i$ where $e_i = \text{expected count}$ and

$x_i =$ % of workforce employed in agriculture, fishing and forestry.

Hierarchy 3: $\underline{b} \sim MVN(\underline{0}, V)$ where $V = \sigma^2 (I - \gamma C)^{-1} M$, $M_{ij} = 1/e_i$, the inverse of the expected count in the i-th area and $C_{ij} = \left(e_i / e_j\right)^{1/2}$. The spatial association parameter $\gamma \in (\gamma_{\min}, \gamma_{\max})$ where $\gamma_{\min}^{-1}$ and $\gamma_{\max}^{-1}$ are the smallest and largest eigenvalues of $M^{-1/2} C M^{1/2}$. This assures that the distribution of the random effects is a proper distribution.

Table 3: Maximum likelihood estimation of spatial Generalized Linear Mixed Models

Parameters	MLE-P1	MLE-P2	MLE-P3
α_0	-0.4466 (0.00248)	-0.4590 (0.0031)	-0.4515(0.0027)
α_1	0.6043 (0.001164)	0.6170 (0.0012)	0.6121(0.0012)
γ	0.1750 (0.000017)	0.1745 (0.000022)	0.1749 (0.000019)
σ	1.302 (0.0044)	1.303 (0.0044)	1.303 (0.0045)

When do we need composite likelihood?

1) Very large datasets: Standard programs such as WinBUGS fail or are too slow if the dataset is large (which will become even larger with cloning). One possibility is to split the data in manageable chunks and analyze them separately using data cloning. One can combine the estimates using their Fisher information (although covariance will be tricky). This is justifiable using the CL principle.

2) Model robustness: The marginal distributions are specified (whether joint distribution exists or not) then one can use CL based estimator.

Summary

- Data cloning seems like a lazy man's way to analyze hierarchical models.
- Theoretically, no loss of efficiency
- Off the shelf programs can be used
- MCMC can be dangerous but less so if proper priors are used.
- Data cloning can be used to predict the random effects.
- Perhaps the EM algorithm for CL will also achieve the same. What are the properties of EMCL or data cloning based predictions?