Hierarchical Models are not Bayesian

A Review and Synthesis

Brian Dennis
Department of Fish and Wildlife Sciences
and
Department of Statistical Science
University of Idaho

Outline

1. Growth of hierarchical models in ecology

2. Bayesian and frequentist inference

3. What is an hierarchical model

4. Inference for hierarchical models

5. The problem of estimability

6. Examples

   non-hierarchical model (nonlinear regression)

   hierarchical model (Poisson-gamma)

7. Concluding remarks

1. Growth of hierarchical models in ecology

Expositions of HMs in have burgeoned in the ecological literature

    almost always Bayesian

    fit models of unheard-of complexity in the face of limited data

A Bayesian hierarchical model of Antarctic fur seal foraging and pup growth related to sea ice and prey abundance

Lisa M. Hiruki-Raring,[1] Jay M. Ver Hoef, Peter L. Boveng, and John L. Bengtson

*National Marine Mammal Laboratory, Alaska Fisheries Science Center, National Marine Fisheries Service, NOAA,
7600 Sand Point Way NE, Seattle, Washington 98115-6349 USA*

Abstract. We created a Bayesian hierarchical model (BHM) to investigate ecosystem relationships between the physical ecosystem (sea ice extent), a prey measure (krill density), predator behaviors (diving and foraging effort of female Antarctic fur seals, *Arctocephalus gazella*, with pups) and predator characteristics (mass of maternal fur seals and pups). We collected data on Antarctic fur seals from 1987/1988 to 1994/1995 at Seal Island, Antarctica. The BHM allowed us to link together predators and prey into a model that uses all the data efficiently and accounts for major sources of uncertainty. Based on the literature, we made hypotheses about the relationships in the model, which we compared with the model outcome after fitting the BHM. For each BHM parameter, we calculated the mean of the posterior density and the 95% credible interval. Our model confirmed others' findings that increased sea ice was related to increased krill density. Higher krill density led to reduced dive intensity of maternal fur seals, as measured by dive depth and duration, and to less time spent foraging by maternal fur seals. Heavier maternal fur seals and lower maternal foraging effort resulted in heavier pups at 22 d. No relationship was found between krill density and maternal mass, or between maternal mass and foraging effort on pup growth rates between 22 and 85 days of age. Maternal mass may have reflected environmental conditions prior to the pup provisioning season, rather than summer prey densities. Maternal mass and foraging effort were not related to pup growth rates between 22 and 85 d, possibly indicating that food was not limiting, food sources other than krill were being used, or differences occurred before pups reached age 22 d.

Key words: Antarctic fur seal; Arctocephalus gazella; *Bayesian hierarchical model; diving; ecosystem; ecosystem model;* Euphausia superba; *foraging; krill; pup growth; sea ice.*

Introduction

The western Antarctic Peninsula area has been the focus of attention in recent years, as the marine ecosystem in the Antarctic Peninsula area is one of the most rapidly warming regions on Earth (Clarke et al. 2007, Ducklow et al. 2007). Several large-scale ecosystem analyses have been done around the Antarctic Peninsula and South Georgia areas (e.g., Ducklow et al. 2007, Murphy et al. 2007*a*) in addition to studies focusing primarily on predator response to climate change and prey availability (Reid and Croxall 2001, Fraser and Hofmann 2003, Reid et al. 2005, Clarke et al. 2007, Trathan et al. 2007), and physical process studies (Murphy et al. 2007*b*, Stammerjohn et al. 2008). There is an urgent need to assess the response of marine populations to the warming trend, but relationships between environmental measures, prey, and predators are not always straightforward, and studies based on few species over a short period of time may not show the complexity of the ecosystem response (Clarke et al. 2007).

Ecosystem models aim to depict the main components and processes of ecosystems, both to gain a better understanding of interactions between ecosystem elements, and to predict how they change in response to inputs to the model. Ecosystem models often address two aspects: estimating parameters, and predicting missing observations. An estimate of a parameter is a mathematical construct, where the data are related mathematically to give a value (e.g., the slope of a regression line) that gives a better understanding of the relationship between the data points. A prediction takes measurable data (such as time series data) and predicts values for years in which data were not collected.

The value of ecosystem models is in their holistic approach to the relationships within the model. Rather than single relationships driving an analysis, each association in turn affects the other interactions within the model. With the emphasis in the past 25 years on ecosystem management (Christensen et al. 1996) rather than management of single species, a single unified model relating several trophic levels and predator–prey relationships provides information both on how the independent elements relate to one another and on making useful predictions from the linked components of the model. Our goal was to explore ecosystem relationships using such a multilevel unified, but

OF *BUGS* AND BIRDS: MARKOV CHAIN MONTE CARLO FOR HIERARCHICAL MODELING IN WILDLIFE RESEARCH

WILLIAM A. LINK,[1] USGS Patuxent Wildlife Research Center, 11510 American Holly Drive, Laurel, MD 20708, USA
EMMANUELLE CAM,[2] USGS Patuxent Wildlife Research Center, 11510 American Holly Drive, Laurel, MD 20708, USA
JAMES D. NICHOLS, USGS Patuxent Wildlife Research Center, 11510 American Holly Drive, Laurel, MD 20708, USA
EVAN G. COOCH, Department of Natural Resources, Fernow Hall, Cornell University, Ithaca, NY 14853, USA

Abstract: Markov chain Monte Carlo (MCMC) is a statistical innovation methodology that allows researchers to fit far more complex models to data than is feasible using conventional methods. Despite its widespread use in a variety of scientific fields, MCMC appears to be underutilized in wildlife applications. This may be due to a misconception that MCMC requires the adoption of a subjective Bayesian analysis, or perhaps simply to its lack of familiarity among wildlife researchers. We introduce the basic ideas of MCMC and software *BUGS* (Bayesian inference using Gibbs sampling), stressing that a simple and satisfactory intuition for MCMC does not require extraordinary mathematical sophistication. We illustrate the use of MCMC with an analysis of the association between latent factors governing individual heterogeneity in breeding and survival rates of kittiwakes (*Rissa tridactyla*). We conclude with a discussion of the importance of individual heterogeneity for understanding population dynamics and designing management plans.

Key words: Bayesian analysis, kittiwake, Markov chain Monte Carlo, *Rissa tridactyla*, statistical methods.

This paper is an introduction to Markov chain Monte Carlo (MCMC), a powerful statistical tool that is used to analyze large, complicated data sets, especially those with complex hierarchical structures. The basic ideas of MCMC were introduced almost 50 years ago (Metropolis et al. 1953) and gained popularity during the 1980s in image processing (Geman and Geman 1984). A growing appreciation of the usefulness of MCMC has led to an explosion of publications in the statistical literature (Gilks et al. 1996). However, relatively few examples are found in wildlife-related applications.

We suggest several reasons why MCMC has not been more widely used in wildlife applications. First, MCMC has a decidedly Bayesian flavor, which may not appeal to data analysts with a classical (Frequentist) training. However, MCMC can be used as a tool to obtain the maximum likelihood estimates used by Frequentists, even for models with complexity that defies conventional analysis. Markov chain Monte Carlo also can be used in Objective Bayes analyses, the results of which are similar to those of Frequentist analyses. We begin with a brief review of Bayesian modeling, contrasting it with the Frequentist approach. This review lays the foundation for description of MCMC and describes Objective Bayes methods.

Another reason MCMC is not yet widely used among wildlife biologists may be a lack of familiarity. Markov chain Monte Carlo involves some complex mathematical ideas. A need exists for a clear accounting of what can be done with it and how it works. We describe the need for MCMC and its basic ideas and mechanisms. A simple and satisfactory intuition for MCMC does not require extraordinary mathematical sophistication.

We illustrate the usefulness of MCMC by analyzing the association between latent factors governing individual heterogeneity in breeding and survival rates of kittiwakes. The data set consists of survival and breeding records for 845 birds, collected over 13 years. The question of interest was whether there are trade-offs between components of fitness. For example, is it true that better breeders tend to have lower survival rates? The models we fit to these data are quite complex—impossible to fit using conventional methods—but are satisfactorily estimated using MCMC. We illustrate the application of MCMC to these data, using software *BUGS* (Spiegelhalter et al. 1995), available for free download (http://www.mrc-bsu.cam.ac.uk/bugs/).

FREQUENTIST, BAYES, AND OBJECTIVE BAYES MODELS

The primary distinction between Bayesian and Frequentist analyses is in the interpretation of model parameters. In both types of analysis, data (Y) are regarded as sampled from a sampling distrib-

---
[1] E-mail: william_link@usgs.gov
[2] Present address: Laboratoire de Biologie, Université de Bretagne Occidentale, 29285 Brest Cedex, France.

A future for models and data in environmental science

James S. Clark[1] and Alan E. Gelfand[2]

[1]Nicholas School of the Environment and Department of Biology, Duke University, Durham, NC 27708, USA
[2]Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708, USA

Together, graphical models and the Bayesian paradigm provide powerful new tools that promise to change the way that environmental science is done. The capacity to merge theory with mechanistic understanding and empirical evidence, to assimilate diverse sources of information and to accommodate complexity will transform the collection and interpretation of data. As we discuss here, we specifically expect a shift from a focus on simple experiments with inflexible design and selection among models that embrace parts of processes to a synthesis of integrated process models. With this potential come new challenges, including some that are specific and technical and others that are general and will involve reexamination of the role of inference and prediction.

Introduction

Be it climate change, exotic invasions, extinction risk, or emerging diseases, the contrast between the simple models and experiments used to learn about ecosystems versus actual ecosystem behavior has never been more apparent. With a growing appreciation of ecosystem complexity has come scrutiny and criticism of traditional models, appeals for synthesis, and frustration with statistical methods used to understand and predict nature [1–3]. Ecosystems are increasingly seen as the product of huge numbers of interacting forces [4,5], food webs being among the few examples where ecologists have attempted to enumerate the complex interactions. Even here, efforts to define a seemingly obvious relationship (whether a species eats another) have been viewed as highly subjective [6]. Influences vary with setting and scale, nonlinearities abound and ecosystem properties emerge from interacting local and global influences.

For many pressing environmental challenges, ecologists do not feel qualified to anticipate ecosystem change [7] and extrapolation from small-scale experiments to relevant settings can be an act of faith. The challenges faced can be expressed in several ways: How do we combine observations that derive from many sources? Can we join what is learnt about parts of a process in isolation (e.g. controlled experiments) in ways that integrate their interactions in nature? How do we connect observations that are specific to location, time and setting with

understanding that comes from a diverse body of nonspecific theory? Can we accommodate the uncountable interactions among unseen forces that collectively determine ecosystem behavior? Given that learning requires models, and modeling demands simplification, how is simplicity achieved from such overwhelming complexity?

The long-standing dichotomy of 'statistical' or 'empirical' models versus 'theoretical' or 'mechanistic' models contributes to the challenge of synthesis [8]. 'Statistical modeling' is applied to experiments for purposes of inference. Few ecologists make predictions from such models, appreciating the specific settings and scales from which they derive [1,9] and their 'non-mechanistic' nature. Rather, experimental results are extrapolated to nature in informal ways, as demonstrated by the many examples included in debates on carbon cycling and increasing atmospheric CO_2 concentrations [10].

'Theoretical' or 'mechanistic' modeling is used for understanding and prediction. Simple models with few parameters are analyzed to discover general features of model behavior. Complex models with many parameters are explored with simulation. Parameter values, rather than data, are the inputs, which might be scavenged from the literature, as they are rarely available for the specific contexts in which the models are applied. For example, ecological and conservation journals include lively debate on topics such as predicted time to extinction for endangered species, based on models and/or parameter values that might not apply to the future or to the specific setting [11,12].

The divide between statistical versus theoretical models is partly a reaction to the complexity challenge. Classical statistical practice demands simplicity in terms of strict experimental design and control. Efforts to shoehorn relevant observations into a rigid framework have contributed to debates on classical hypothesis testing [13,14], the importance of controlled experiments [15], pseudoreplication [16] and the expanding application of Bayesian inference [17]. Uncertainty comes with observations [18–20], model parameters [11,21,22], the specification of processes [23,24] and scenarios developed for the exploration of possible futures [7,25]. An abiding consensus cutting across the debates is that proper statistical practice must be the objective arbiter of evidence. But synthesis of general theory, mechanistic understanding

Corresponding author: Clark, J.S. (jimclark@duke.edu).
Available online 19 April 2006

A Bayesian hierarchical model for monitoring harbor seal changes in Prince William Sound, Alaska

JAY M. VER HOEF and KATHRYN J. FROST

Alaska Department of Fish and Game, 1300 College Road, Fairbanks, Alaska 99701, U.S.A.
E-mail: ffjmv@uaf.edu

Bayesian hierarchical models were used to assess trends of harbor seals, *Phoca vitulina richardsi*, in Prince William Sound, Alaska, following the 1989 *Exxon Valdez* oil spill. Data consisted of 4–10 replicate observations per year at 25 sites over 10 years. We had multiple objectives, including estimating the effects of covariates on seal counts, and estimating trend and abundance, both per site and overall. We considered a Bayesian hierarchical model to meet our objectives. The model consists of a Poisson regression model for each site. For each observation the logarithm of the mean of the Poisson distribution was a linear model with the following factors: (1) intercept for each site and year, (2) time of year, (3) time of day, (4) time relative to low tide, and (5) tide height. The intercept for each site was then given a linear trend model for year. As part of the hierarchical model, parameters for each site were given a prior distribution to summarize overall effects. Results showed that at most sites, (1) trend is down; counts decreased yearly, (2) counts decrease throughout August, (3) counts decrease throughout the day, (4) counts are at a maximum very near to low tide, and (5) counts decrease as the height of the low tide increases; however, there was considerable variation among sites. To get overall trend we used a weighted average of the trend at each site, where the weights depended on the overall abundance of a site. Results indicate a 3.3% decrease per year over the time period.

Keywords: trend analysis, abundance estimation, population monitoring, Markov Chain Monte Carlo, Poisson regression, aerial surveys, *Exxon Valdez* oil spill, harbor seal, *Phoca vitulina richardsi*, Prince William Sound

1. Introduction

Monitoring programs to track long-term changes in population size are important for applied ecological studies. Such monitoring programs often have multiple objectives that include monitoring trends, estimating abundance, and estimating the effects of covariates, both for large areas and smaller areas that comprise the larger area. In this paper we develop a Bayesian hierarchical model for analyzing trend, abundance, and the effects of covariates for monitoring programs of multiple sites, and we apply it to counts of harbor seals following the *Exxon/Valdez* oil spill of 1989 in the Prince William Sound, Alaska.

Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling

NOEL CRESSIE,[1,5] CATHERINE A. CALDER,[1] JAMES S. CLARK,[2] JAY M. VER HOEF,[3] AND CHRISTOPHER K. WIKLE[4]

[1]*Program in Spatial Statistics and Environmental Statistics, Department of Statistics, Ohio State University, Columbus, Ohio 43210 USA*
[2]*Nicholas School of the Environment, Duke University, Durham, North Carolina 27708 USA*
[3]*NOAA National Marine Mammal Laboratory, NMFS Alaska Fisheries Science Center, Fairbanks, Alaska 99775-7345 USA*
[4]*Department of Statistics, University of Missouri, Columbia, Missouri 65201 USA*

Abstract. Analyses of ecological data should account for the uncertainty in the process(es) that generated the data. However, accounting for these uncertainties is a difficult task, since ecology is known for its complexity. Measurement and/or process errors are often the only sources of uncertainty modeled when addressing complex ecological problems, yet analyses should also account for uncertainty in sampling design, in model specification, in parameters governing the specified model, and in initial and boundary conditions. Only then can we be confident in the scientific inferences and forecasts made from an analysis. Probability and statistics provide a framework that accounts for multiple sources of uncertainty. Given the complexities of ecological studies, the hierarchical statistical model is an invaluable tool. This approach is not new in ecology, and there are many examples (both Bayesian and non-Bayesian) in the literature illustrating the benefits of this approach. In this article, we provide a baseline for concepts, notation, and methods, from which discussion on hierarchical statistical modeling in ecology can proceed. We have also planted some seeds for discussion and tried to show where the practical difficulties lie. Our thesis is that hierarchical statistical modeling is a powerful way of approaching ecological analysis in the presence of inevitable but quantifiable uncertainties, even if practical issues sometimes require pragmatic compromises.

Key words: Bayesian modeling; data model; design; empirical Bayes; harbor seals; MCMC; prior; process model; spatial process; spatiotemporal process.

INTRODUCTION

The field of ecology is becoming increasingly aware of the importance of accurately accounting for multiple sources of uncertainty when modeling ecological phenomena and making inferences. This development is motivated in part by the desire to provide an accurate picture of the state of knowledge of ecosystems and to be able to assess the quality of predictions of local and global change (Hilborn and Mangel 1997, Daszak et al. 2000, Clark et al. 2001, Beckage and Platt 2003, Clark 2005, Ibáñez et al. 2006, Sacks et al. 2007). However, accounting for various sources of uncertainty is by no means a simple task.

Ecological data are almost always observed incompletely with large and unknown amounts of measurement error or data uncertainty, and often the expense of data collection prohibits collecting as much data as might be desirable. How much and where to sample are

important design questions (e.g., Stevens and Olsen 2004). In addition, most ecological phenomena of interest can only be studied by combining various sources of data; aligning these data properly presents interesting statistical challenges. While data play a large role in most ecological analyses, incorporating scientific knowledge through substantive modeling of ecological processes is essential. Often such process modeling is based on competing scientific theories and simplifications of reality. This results in an additional source of uncertainty, termed model or process uncertainty. Furthermore, substantive models should acknowledge parameter uncertainty. Parameter uncertainty can be handled either by estimating the unknown parameters (empirical-Bayesian analysis) or by expressing that uncertainty via a prior probability distribution (Bayesian analysis); see, for example, Ver Hoef (1996), Carlin and Louis (2000), and Gelman and Hill (2006), where the two approaches are presented. An empirical-Bayesian analysis looks for plug-in estimates and may avoid more realistic and flexible specifications that can include variation over space and time. The Bayesian analysis can use such variation to help with the choice of the prior distribution.

[5] E-mail: ncressie@stat.asu.edu

2. Bayesian and frequentist inference

Frequentist inference

 based on *sample space probability* (hypothetical repeated sampling)

 x data, θ parameter(s), $f(x \,|\, \theta)$ model or likelihood function

 ex. μ is mean beak length in a population of birds, population (of beak lengths is represented by a normal (μ, σ^2) distribution, data (random sample) $x = (x_1, x_2, ..., x_n)$ are collected, likelihood is

$$f(x \,|\, \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left[-\tfrac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right]$$

 inferences

 point estimation of θ (ML, etc)   ex. $\widehat{\mu} = \overline{x}$, $\widehat{\sigma}^2 = \frac{n-1}{n} s^2$

 interval estimation of θ (CIs)

statistical hypothesis testing (Neyman-Pearson, Fisher/goodness of fit)

information-based model selection (AIC, etc)

prediction

Bayesian inference

based on *parameter space probability* (personal beliefs about the possible values of θ)

x data, θ parameter(s), $f(x \,|\, \theta)$ model or likelihood function

$g(\theta)$ prior distribution (distribution on parameter space reflecting beliefs about θ before the advent of data)

inferences about θ are based on the posterior distribution (distribution of beliefs modified by data)

$$g(\theta \,|\, x) = C f(x \,|\, \theta) g(\theta)$$

$$C^{-1} = \int f(x \,|\, \theta) g(\theta) d\theta$$

ex. product normal likelihood, $g(\mu)$ pdf of a normal(ϕ, τ^2) distribution, then $g(\mu \mid x)$ becomes the pdf of a normal distribution with mean given by

$$\mu_{\mathsf{B}} \;=\; \frac{(\sigma^2/\tau^2)}{(\sigma^2/\tau^2)+n}\phi \;+\; \frac{n}{(\sigma^2/\tau^2)+n}\overline{x}$$

point estimation (mean of the posterior, ex. μ_{B})

interval estimation (HPR of the posterior, etc)

statistical hypothesis testing (posterior probabilities of θ values)

model selection (multiple posterior probabilities)

prediction (using posterior to find distribution of future data given x)

example: Dugong length-age data (Ratkowsky 1983, Carlin and Gelfand 1991)
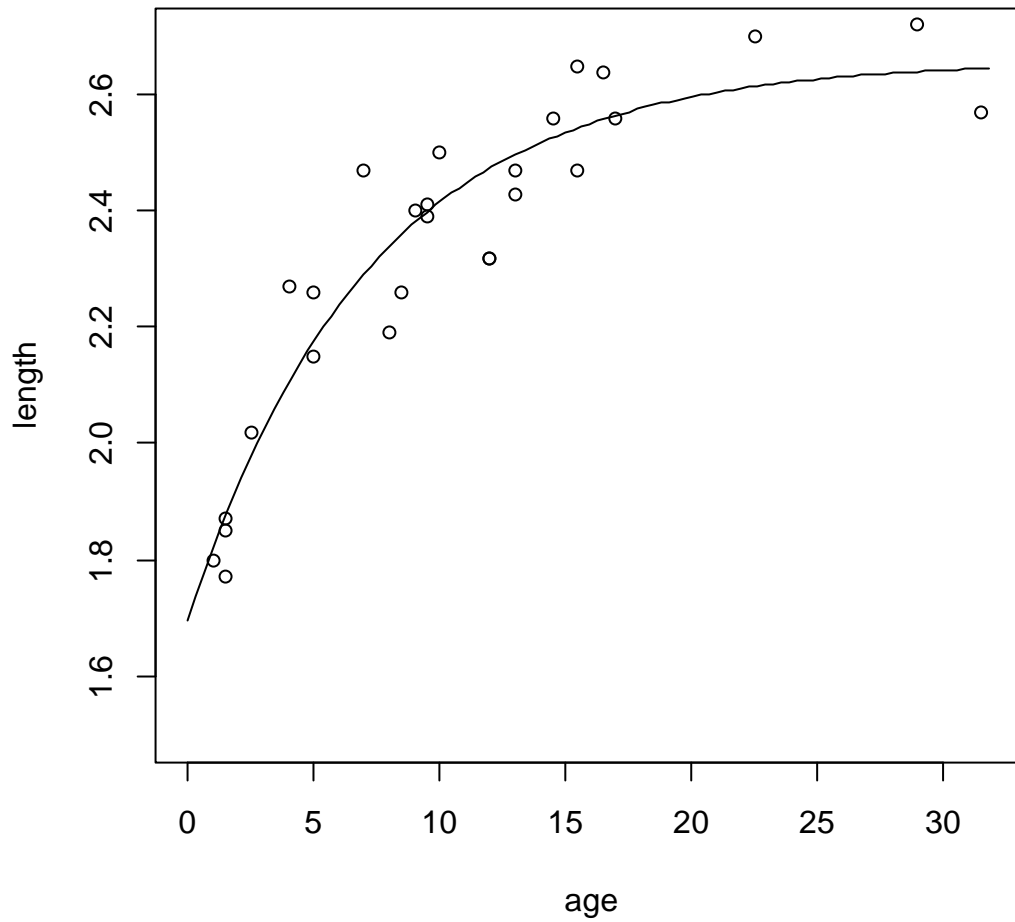
Data: $x_1, x_2, ..., x_n$ ages
        $y_1, y_2, ..., y_n$ lengths

Model: length $= \alpha(1 - \beta\gamma^{\text{age}})$ growth curve

$$Y_i \sim \text{normal}(\alpha(1 - \beta\gamma^{x_i}), \sigma^2)$$

Likelihood function $f(y \,|\, x, \alpha, \beta, \gamma, \sigma^2)$ is a product of normal pdfs

Frequentist estimation:

ML estimates of α, β, γ minimize sum of squared departures of y_i from
$\alpha(1 - \beta\gamma^{x_i})$ (nonlinear regression)

Bayesian estimation:

specify prior distributions $g_1(\alpha)$, $g_2(\beta)$, $g_3(\gamma)$, $g_4(\sigma^2)$
(actually must specify $g(\alpha, \beta, \gamma, \sigma^2)$)

calculate $g(\alpha, \beta, \gamma, \sigma^2 \,|\, y) = C f(y \,|\, x, \alpha, \beta, \gamma, \sigma^2) g_1(\alpha) g_2(\beta) g_3(\gamma) g_4(\sigma^2)$

technical problem up until around 1990: could not calculate the posteriors
for any but the simplest models (whereas numerical maximization of
likelihood was feasible)

3. What is an hierarchical model

Hierarchical model: physical *heterogeneity* (usually unrecorded) in one or more elements of θ modeled by a probability distribution, $h(\theta)$

   the distribution typically has its own unknown parameters, $h(\theta) = h(\theta \,|\, \psi)$

   random effects (ex. plant yield μ varies from field to field)

   missing data (ex. x are observed population sizes, $\theta = (v, \theta_1)$ where v is
      vector of actual population sizes or unrecorded covariates)

   heterogeneity/overdispersion/aggregation (capture probabilities, population
      densities, vary)

   likekihood $f_{\mathsf{h}}(x \,|\, \psi)$ is $f(x \,|\, \theta) h(\theta \,|\, \psi) d\theta$ integrated over all possible values of θ

Note: nothing Bayesian yet

ex. The famous Yule model of accident-proneness (& statistical ecology model of aggregation)

$$X_i | \lambda \sim \text{Poisson}(\lambda t_i)$$

$$\Lambda \sim \text{gamma}(\alpha, \beta)$$

$$X_i \sim \text{negative binomial}\left(\alpha, \left(\frac{t_i}{\beta + t_i}\right)\right)$$

4. Inference for hierarchical models

Frequentist estimation: in principle can simply maximize $f_h(x \mid \psi)$ for $\widehat{\psi}$

> technical problem up until mid-2000s: could not calculate the likelihood $f_h(x \mid \psi)$ for any but the simplest models (normal, beta-binomial, etc)
>
> therefore, frequentist inferences for hierarchical models was mostly not feasible

Bayesian estimation: priors on all unknown parameters; calculate posterior

$$g(\psi \mid x) = C f(x \mid \theta) h(\theta \mid \psi) g(\psi)$$

Until 1990, Bayesian estimation for HMs was mostly not feasible as well

In approx 1990, a huge breakthrough occurred. It was discovered that the MCMC algorithms for random variable generation (especially Gibbs sampling) could be used to simulate observations from the posterior distribution for an HM, without having to calculate the constant C (or do any numerical integrations)

Furthermore, any number of hierarchies could be thrown into the Bayesian MCMC calculations, opening the way for a vast leap in complexity of models that could be fitted:

$$g(\psi \,|\, x) =$$

$$C f(x \,|\, \theta_1, \psi_1) h_1(\theta_1 \,|\, \theta_2, \psi_2) h_2(\theta_2 \,|\, \theta_3, \psi_3) \cdots h_m(\theta_m \,|\, \psi_m) g_1(\psi_1) g_2(\psi_2) \cdots g_m(\psi_m)$$

Frequentist inference was left stranded at the side of the road

For the trend parameters, we will also group the site-specific covariate parameters and give them a distribution; $f(\tau_{qi}|\eta_q, \gamma_q^2) = N(\eta_q, \gamma_q^2)$, where jointly,

$$f(\boldsymbol{\tau}|\boldsymbol{\eta}, \boldsymbol{\gamma}) = \prod_{q=0}^{1} \prod_{i=1}^{25} f(\tau_{qi}|\eta_q, \gamma_q^2).$$

We also group the overdispersion parameters and give them a distribution; $f(\varepsilon_{ijk}|\mathbf{0}, \xi_i^2) = N(0, \xi_i^2)$, where jointly,

$$f(\boldsymbol{\varepsilon}|\mathbf{0}, \boldsymbol{\xi}) = \prod_{i=1}^{25} \prod_{j=-4}^{5} \prod_{k=1}^{n_{ij}} f(\varepsilon_{ijk}|0, \xi_i^2),$$

and $f(\xi_i|\nu_a, \nu_b) = GAM(\nu_a, \nu_b)$, where $GAM(a, b)$ is a gamma distribution with parameters a and b, where jointly,

$$f(\boldsymbol{\xi}|\nu_a, \nu_b) = \prod_{i=1}^{25} f(\xi_i|\nu_a, \nu_b).$$

In the fourth and final level of the hierarchy, we give diffuse prior distributions, $f(\mu_p)$ and $f(\eta_q)$ are $N(0, 1{,}000{,}000)$; and $f(\delta^2), f(\sigma_p^2), f(\gamma_q^2), f(\nu_a)$ and $f(\nu_b)$ are $GAM(0.001, 0.001)$. Jointly,

$$f(\boldsymbol{\mu}) = \prod_{p=1}^{7} f(\mu_p), \, f(\boldsymbol{\eta}) = \prod_{q=0}^{1} f(\eta_q), \, f(\boldsymbol{\sigma}) = \prod_{p=1}^{7} f(\sigma_p^2), \text{ and } f(\boldsymbol{\gamma}) = \prod_{q=0}^{1} f(\gamma_q^2).$$

The hierarchical model is shown diagrammatically in Fig. 2. Using the hierarchical setup, Bayes theorem allows us to write the posterior distribution:

$$f(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\varepsilon}, \delta^2, \xi, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \nu_a, \nu_b|\mathbf{z}) \propto$$

$$f(\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\theta}) f(\boldsymbol{\theta}|\boldsymbol{\tau}, \delta^2) f(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\sigma}) f(\boldsymbol{\varepsilon}|\mathbf{0}, \boldsymbol{\xi}) f(\boldsymbol{\tau}|\boldsymbol{\eta}, \boldsymbol{\gamma}) f(\delta^2) f(\boldsymbol{\xi}|\nu_a, \nu_b) f(\boldsymbol{\mu}) f(\boldsymbol{\sigma}) f(\boldsymbol{\eta}) f(\boldsymbol{\gamma}) f(\nu_a) f(\nu_b).$$
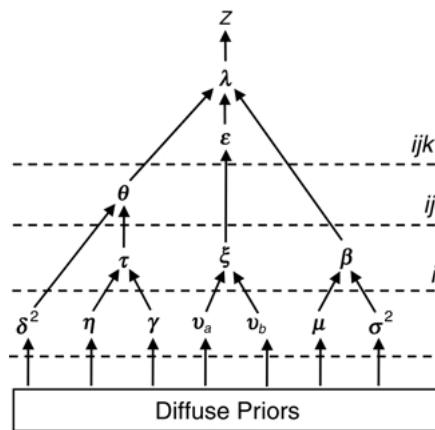
$$(2)$$

Figure 2. Diagrammatic scheme of hierarchical model.

By 2007, frequentist estimation was back on the road and almost caught up

Various algorithms had been devised to calculate ML estimates for HMs

One, called data cloning, is easy to use:

1. Set up a Bayesian MCMC simulation for the HM and get it to run.

2. Perform the calculations for 2 copies (clones) of the data. Calculate the sample mean (vector) of the posterior. Perform the calculations for 4 clones of the data. Calculate the mean of the posterior. Continue the process for, say, 2, 4, 8, 16, 32, ... clones of the data, say, until the posterior means converge to the same value(s).

3. The mean (vector) of the posterior distribution is then the ML estimate(s)

4. If K is the number of data clones used, and V_B is the variance-covariance matrix of the posterior distribution, then $K \times V_B$ is the estimated variance-covariance matrix for the ML estimates.

5. The problem of estimability

Incorporating heterogeneity in the form of a probability distribution for one or more parameters adds variability to parameter estimates, sometimes quite substantial (estimation for negative binomial, beta-binomial, etc)

Sometimes, parameters in HMs become entangled: likelihoods become ridge-shaped, and/or multimodal

The entanglement can be structural (built-in) or by accident of a bad data draw.

ex. $\qquad X_i|\theta_i \sim$ normal(θ_i, σ^2)

$\qquad \Theta_i \sim$ normal(μ, τ^2)

$\qquad X_i \sim$ normal$(\mu, \sigma^2 + \tau^2)$

no amount of data x_1, x_2, ..., x_n can estimate σ^2 and τ^2 separately (if you are a frequentist)

"From a Bayesian perspective, as long as proper prior distributions are assigned to all model parameters, all of them are technically identifiable."
(Cressie et al. 2009)

The example, of course, is extreme. No one in their right mind would try to estimate σ^2 and τ^2 separately in a Bayesian MCMC (we hope).

However, in the big, complex HMs being built, it is not at all clear whether the parameters being estimated are in fact estimable (i.e. data are contributing meaningfully to their estimation)

Problem is widely acknowledged in the Bayesian HM literature.

Bayesian diagnostics concerning estimability (or lack thereof) are currently neither adequate nor convincing

     ex. compare priors to posteriors: usually done only with univariate
      distributions, risks being mislead by marginalization paradox

ML estimation with data cloning provides a simple diagnostic tool for estimability:

As the number of clones K becomes larger, the posterior distribution should converge to a degenerate distribution if everything is ok

For each value of K, calculate the largest eigenvalue λ_K of the posterior variance-covariance matrix V_K

Plot the ratios λ_K/λ_1 as a function of K. The ratios should decrease at a rate $1/K$

6. Examples

Dugong length-age data & growth model

$$Y_i \sim \text{normal}(\alpha(1 - \beta\gamma^{x_i}), \sigma^2)$$

y_i	x_i		
1.80	1.0	2.50	10.0
1.85	1.5	2.32	12.0
1.87	1.5	2.32	12.0
1.77	1.5	2.43	13.0
2.02	2.5	2.47	13.0
2.27	4.0	2.56	14.5
2.15	5.0	2.65	15.5
2.26	5.0	2.47	15.5
2.47	7.0	2.64	16.5
2.19	8.0	2.56	17.0
2.26	8.5	2.70	22.5
2.40	9.0	2.72	29.0
2.39	9.5	2.57	31.5
2.41	9.5		

WinBUGS code

```
model {
    for( i in 1 : N ) {
        y[i] ~ dnorm(mu[i], tau)
        mu[i] <- alpha - beta * pow(gamma,x[i])
    }
    alpha ~ dnorm(0.0, 1.0E-6)
    beta ~ dnorm(0.0, 1.0E-6)
    gamma ~ dunif(0.0, 1.0)
    tau ~ dgamma(0.01, 0.01)
}

DATA

list(N=27)
y[] x[]
1.80 1.0
1.85 1.5
1.87 1.5
1.77 1.5
2.02 2.5
2.27 4.0
```

```
2.15 5.0
2.26 5.0
2.47 7.0
2.19 8.0
2.26 8.5
2.40 9.0
2.39 9.5
2.41 9.5
2.50 10.0
2.32 12.0
2.32 12.0
2.43 13.0
2.47 13.0
2.56 14.5
2.65 15.5
2.47 15.5
2.64 16.5
2.56 17.0
2.70 22.5
2.72 29.0
2.57 31.5
END
```

```
FIRST INITIAL VALUES
list(alpha=1,beta=1,tau=1,gamma=0.9)

SECOND INITIAL VALUES ARE RANDOMLY DRAWN FROM
THE PRIOR DISTRIBUTIONS
```

	alpha	beta	gamma	tau (=1/sigma^2)
ML:	2.658	0.9635	0.8715	124.0
K = 40	2.658	0.9636	0.8713	123.4
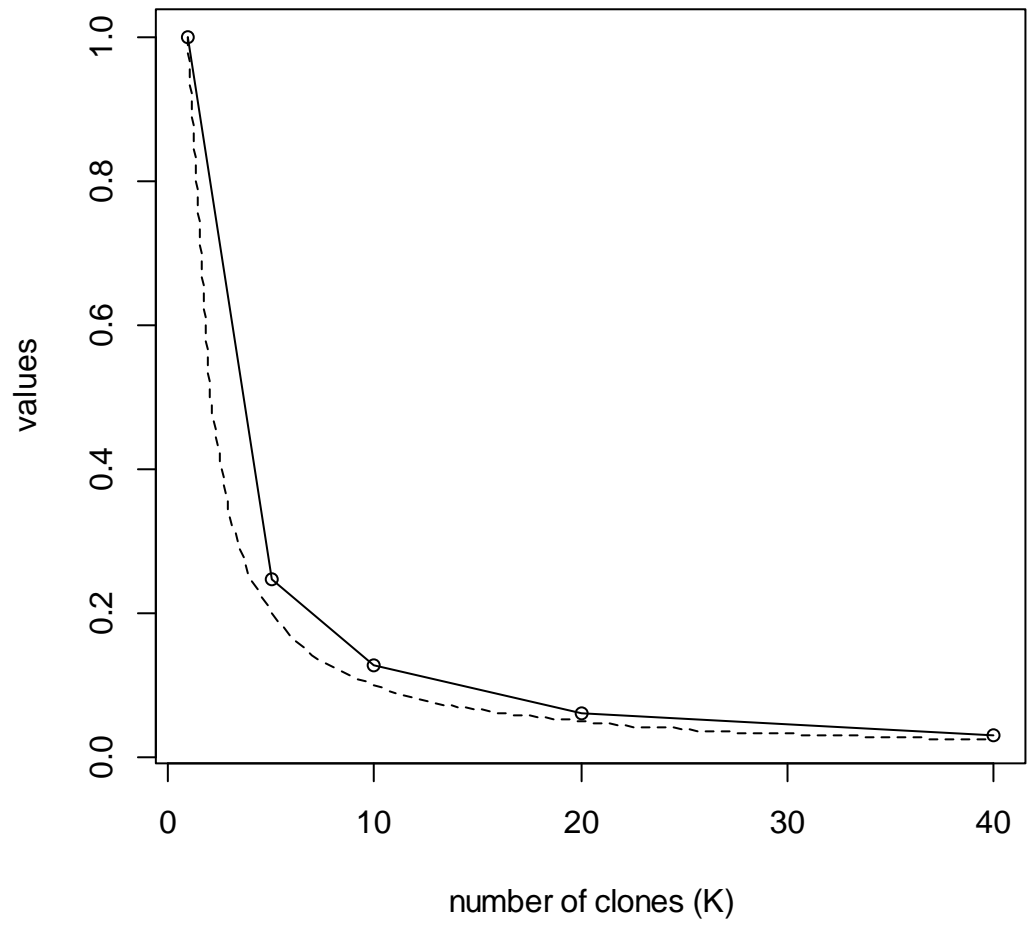
Pump data with Poisson-gamma model

X_i is number of failures of pump i in time t_i

Pump	t_i	x_i
1	94.5	5
2	15.7	1
3	62.9	5
4	126	14
5	5.24	3
6	31.4	19
7	1.05	1
8	1.05	1
9	2.1	4
10	10.5	22

$$X_i|\theta_i \sim \text{Poisson}(\theta_i t_i) \qquad \Theta_i \sim \text{gamma}(\alpha, \beta)$$

$$X_i \sim \text{negative binomial}\left(\alpha, \left(\frac{t_i}{\beta+t_i}\right)\right)$$

WinBUGS code

```
model {
    for (i in 1 : N) {
        theta[i] ~ dgamma(alpha, beta)
        lambda[i] <- theta[i] * t[i]
        x[i] ~ dpois(lambda[i])
    }
    alpha ~ dexp(1)
    beta ~ dgamma(0.1, 1.0)
}
```

DATA

```
list(N=10)
t[] x[]
94.3 5
15.7 1
62.9 5
126 14
5.24 3
31.4 19
1.05 1
```

```
1.05 1
2.1 4
10.5 22
END

INITS
list(alpha = 1, beta = 1)
```

	alpha	beta
ML:	0.8230	1.262
$K = 160$:	0.8222	1.259
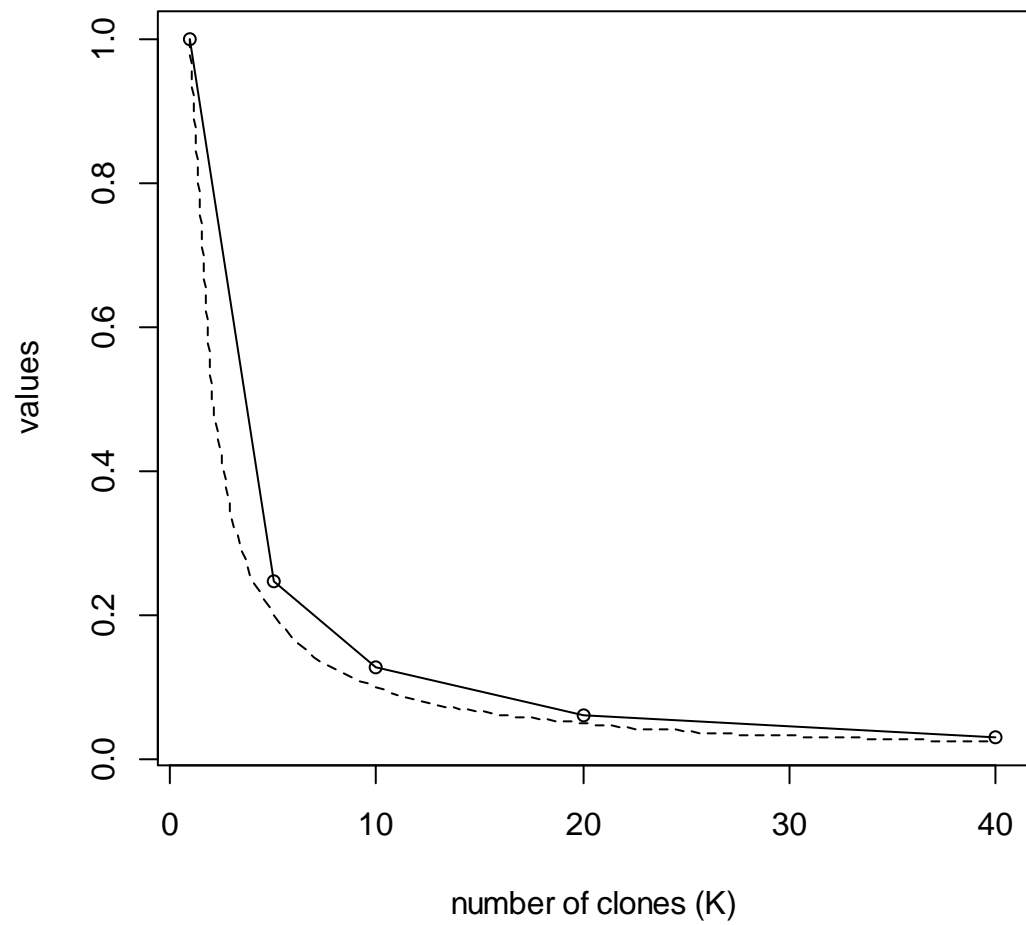
State space model of population growth

$$X_t = a + cX_{t-1} + E_t \qquad (X_t: \text{log-population abundance})$$
$$Y_t = X_t + F_t \qquad\qquad\qquad (Y_t: \text{observed or estimated value of } X_t)$$

$$E_t \sim \text{normal}(0, \sigma^2), \quad F_t \sim \text{normal}(0, \tau^2)$$

data: time series $y_0, y_1, y_2, ..., y_q$

parameters: a, c, σ^2, τ^2

It is known that likelihoods for this model can sometimes be ridge-shaped and/or multimodal

Estimability of parameters in time series population abundance models assessed using data cloning
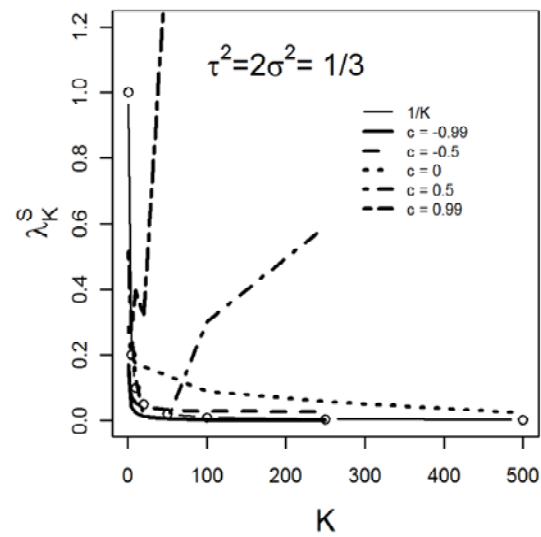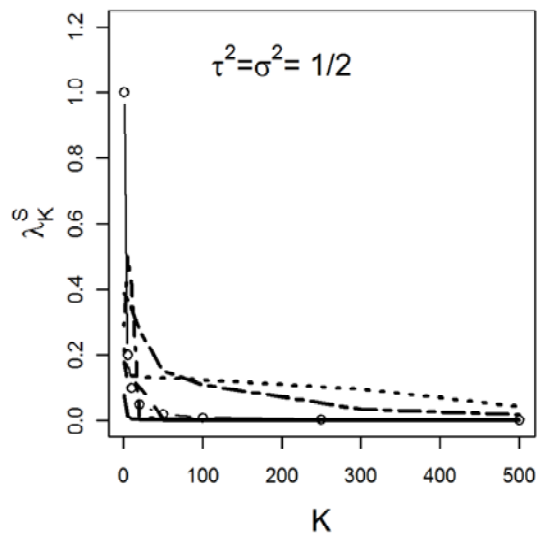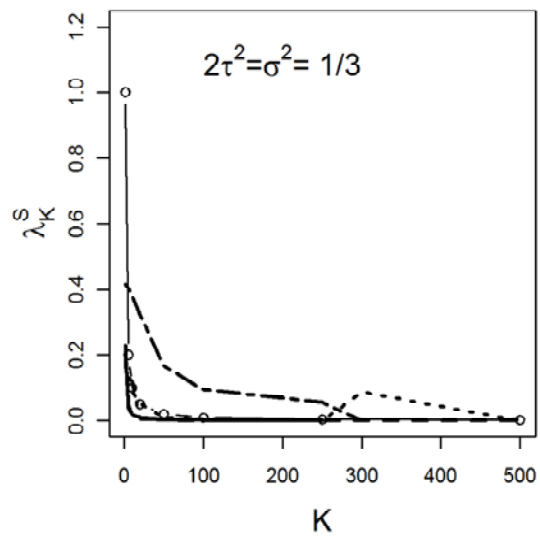
A thesis presented in partial fulfillment of the requirements for the
Master of Science Degree with a major in Statistical Science
in the College of Science
University of Idaho

By
Kevin White

Spring 2012

Major Professor: Brian Dennis, Ph.D.
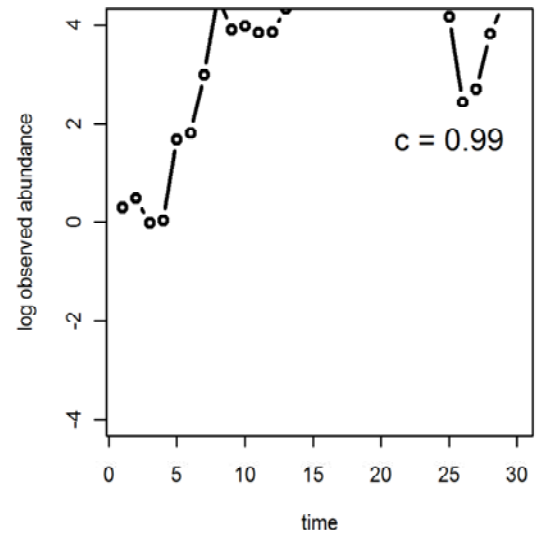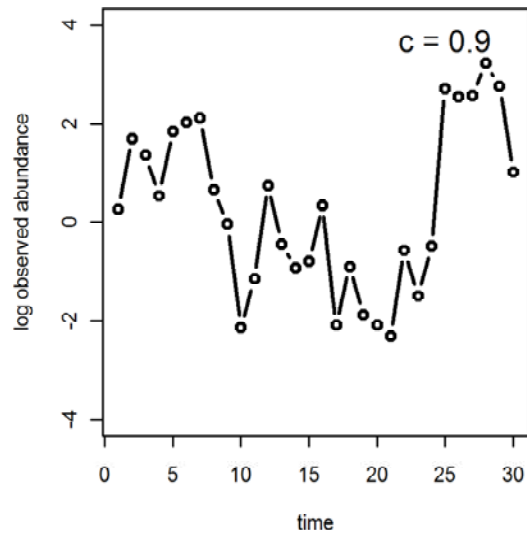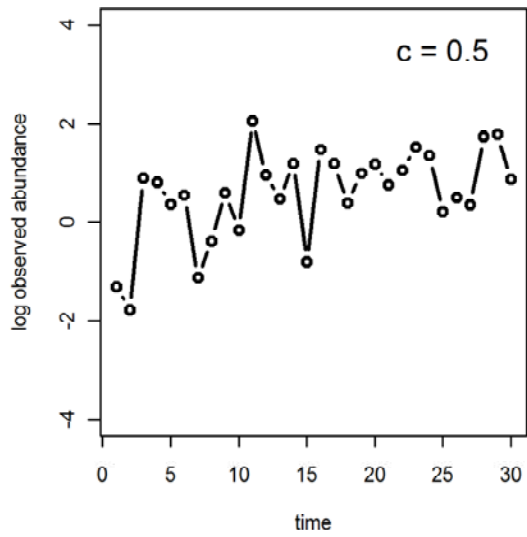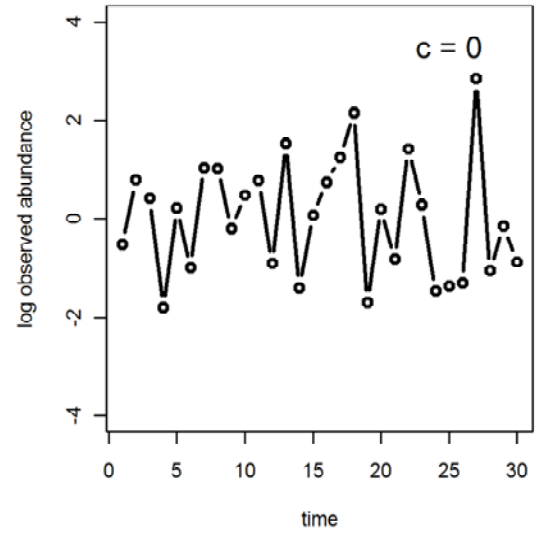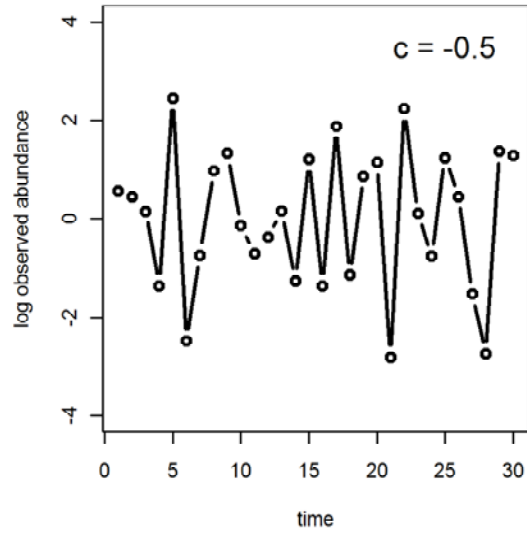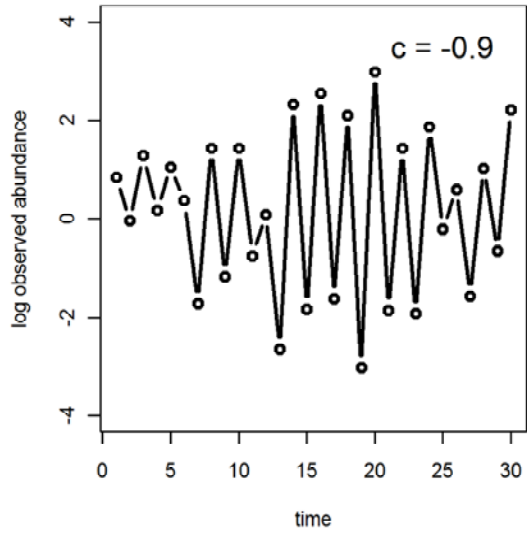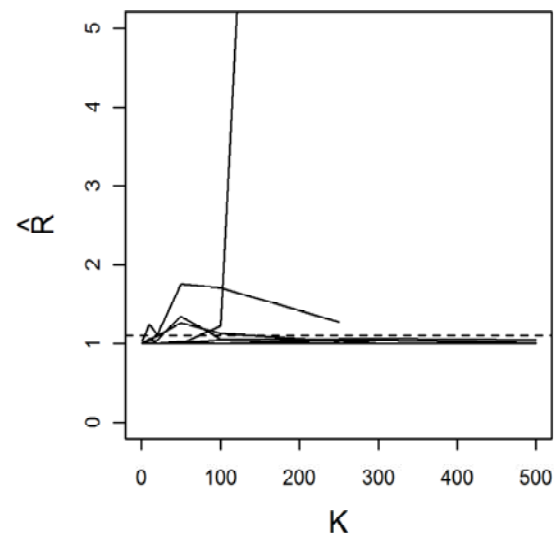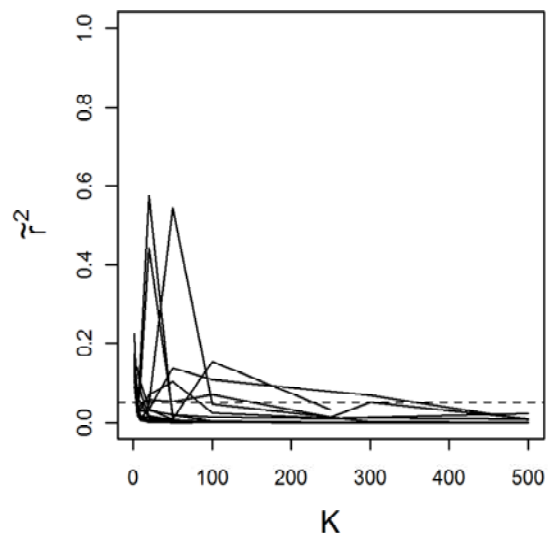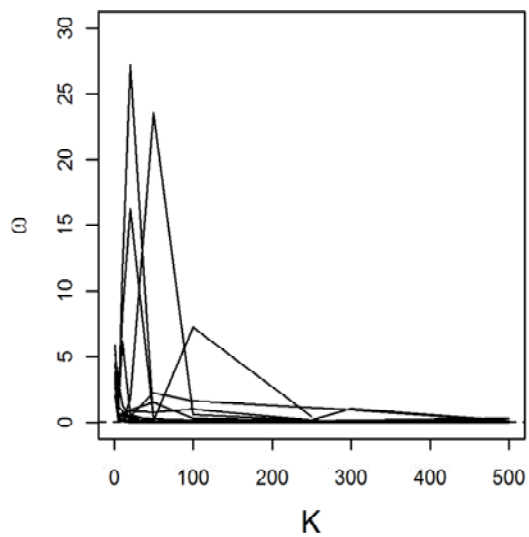
LETTER

Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods

Subhash R. Lele,[1] Brian Dennis[2] and Frithjof Lutscher[3]

[1]*Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB T6G2G1, Canada*
[2]*Department of Fish and Wildlife Resources and Department of Statistics, University of Idaho, Moscow, ID 83844-1136, USA*
[3]*Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON K1N6N5, Canada*
*Correspondence: E-mail: brian@uidaho.edu

Abstract

We introduce a new statistical computing method, called data cloning, to calculate maximum likelihood estimates and their standard errors for complex ecological models. Although the method uses the Bayesian framework and exploits the computational simplicity of the Markov chain Monte Carlo (MCMC) algorithms, it provides valid frequentist inferences such as the maximum likelihood estimates and their standard errors. The inferences are completely invariant to the choice of the prior distributions and therefore avoid the inherent subjectivity of the Bayesian approach. The data cloning method is easily implemented using standard MCMC software. Data cloning is particularly useful for analysing ecological situations in which hierarchical statistical models, such as state-space models and mixed effects models, are appropriate. We illustrate the method by fitting two nonlinear population dynamics models to data in the presence of process and observation noise.

INTRODUCTION

A sea-change in the scale and complexity of ecological data analysis occurred with the development in statistics of practical inference methods for hierarchical models. Hierarchical models are statistical models containing random components in addition to or instead of the usual fixed parameter values, and take such varied forms as generalized linear models with mixed random and fixed effects, structured population state-space models with observational and process variability and capture-recapture models with randomly varying capture probabilities. Applications of hierarchical models in ecology are expanding rapidly, due to the wealth of realistic model structures for describing ecological processes (Table 1).

The most commonly used approach for fitting hierarchical models to data is based on the Bayesian paradigm (Link *et al.* 2002; Clark 2005; Clark & Gelfand 2006). The prior distributions are chosen to be informative, if appropriate; otherwise non-informative priors are commonly used. Computing the Bayesian posterior distribution for hierar-

chical models became feasible with the advent of the Markov chain Monte Carlo (MCMC) algorithms. These algorithms are a collection of probabilistic simulation methods for generating observations from designated statistical distributions (Gelfand & Smith 1990; Casella & George 1992; Gilks *et al.* 1996; Robert & Casella 2004). Free software programs such as WINBUGS (Spiegelhalter *et al.* 2004) have made their application in ecology reasonably easy and straightforward. MCMC algorithms are especially useful when the target statistical distribution, such as the posterior distribution in the Bayesian formulation, contains a high-dimensional integral that cannot be simplified.

Although the Bayesian inferences are computationally feasible, their interpretation is problematic. First, the inferences depend on the choice of the prior distributions. Second, even among statisticians, there is a debate as to how one defines a non-informative or an objective prior (Press 2003, Chapter 5; Barnett 1999, Chapter 10). Third, the credible intervals produced in Bayesian inference have no meaning in terms of the replication of inferences by other studies, but rather represent the beliefs the analyst attaches

Hierarchical models in ecology: confidence intervals, hypothesis testing, and model selection using data cloning

José Miguel Ponciano,[1,5] Mark L. Taper,[2] Brian Dennis,[3] and Subhash R. Lele[4]

[1]*Centro de Investigación en Matemáticas, CIMAT A.C. Calle Jalisco s/n, Col. Valenciana, A.P. 402,*
C.P. 36240 Guanajuato, Guanajuato, México
[2]*Montana State University, Department of Ecology, 301 Lewis Hall, Bozeman, Montana 59717-3460 USA*
[3]*Department of Fish and Wildlife Resources and Department of Statistics, University of Idaho, Moscow, Idaho 83844-1136 USA*
[4]*Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta T6G 2G1 Canada*

Abstract. Hierarchical statistical models are increasingly being used to describe complex ecological processes. The data cloning (DC) method is a new general technique that uses Markov chain Monte Carlo (MCMC) algorithms to compute maximum likelihood (ML) estimates along with their asymptotic variance estimates for hierarchical models. Despite its generality, the method has two inferential limitations. First, it only provides Wald-type confidence intervals, known to be inaccurate in small samples. Second, it only yields ML parameter estimates, but not the maximized likelihood values used for profile likelihood intervals, likelihood ratio hypothesis tests, and information-theoretic model selection. Here we describe how to overcome these inferential limitations with a computationally efficient method for calculating likelihood ratios via data cloning. The ability to calculate likelihood ratios allows one to do hypothesis tests, construct accurate confidence intervals and undertake information-based model selection with hierarchical models in a frequentist context. To demonstrate the use of these tools with complex ecological models, we reanalyze part of Gause's classic *Paramecium* data with state–space population models containing both environmental noise and sampling error. The analysis results include improved confidence intervals for parameters, a hypothesis test of laboratory replication, and a comparison of the Beverton-Holt and the Ricker growth forms based on a model selection index.

Key words: AIC; Bayesian statistics; data cloning; frequentist statistics; hierarchical models; likelihood ratio; Markov chain Monte Carlo; maximum likelihood; model selection; profile likelihood; state–space models; stochastic population models.

Introduction

Reliable understanding of complex ecological data depends on the formulation of proper statistical models of the underlying processes. Hierarchical statistical models have proved highly useful for achieving such understanding in many ecological systems (see Table 1 in Lele et al. 2007). Such models allow researchers to incorporate variability in parameters that otherwise might be unrealistically treated as fixed. In addition, these models allow the incorporation of multiple layers of process and observation uncertainty. Stochastic population models with added observation error (De Valpine and Hastings 2002, Clark and Bjørnstad 2004, Staples et al. 2004, Dennis et al. 2006, Lele 2006, Newman et al. 2006, Sæther et al. 2007), stochastic models of species abundance distributions (Etienne and Olff 2005), and capture–recapture models with uncertain capture probabilities (George and Robert 1992) are just a few examples of this broad class of random effects models.

Until recently, computational difficulties rendered many frequentist statistical inferences for hierarchical models unfeasible. For all but the simplest models, computing the likelihood function needed for such inferences requires computing an intractable, high-dimensional integral. Inferences using computer intensive Bayesian methods side step this difficulty by simulating observations from a posterior distribution using one of the various Markov chain Monte Carlo (MCMC) algorithms (Robert and Casella 2005). Although other approaches are possible, the new data cloning (DC) algorithm by Lele et al. (2007) provides convenient tools to carry out frequentist estimation of the parameters in general hierarchical models. An often-repeated justification of the Bayesian approach is the fact that as sample size increases the Bayesian solution approaches the maximum likelihood solution (Walker 1969). The trick in data cloning is to apply a Bayesian methodology to a data set constructed by duplicating the original data set enough times that the Walker theorems apply.

Estimability and Likelihood Inference for Generalized Linear Mixed Models Using Data Cloning

Subhash R. LELE, Khurram NADEEM, and Byron SCHMULAND

Maximum likelihood estimation for Generalized Linear Mixed Models (GLMM), an important class of statistical models with substantial applications in epidemiology, medical statistics, and many other fields, poses significant computational difficulties. In this article, we use data cloning, a simple computational method that exploits advances in Bayesian computation, in particular the Markov Chain Monte Carlo method, to obtain maximum likelihood estimators of the parameters in these models. This method also leads to a simple estimator of the asymptotic variance of the maximum likelihood estimators. Determining estimability of the parameters in a mixed model is, in general, a very difficult problem. Data cloning provides a simple graphical test to not only check if the full set of parameters is estimable but also, and perhaps more importantly, if a specified function of the parameters is estimable. One of the goals of mixed models is to predict random effects. We suggest a frequentist method to obtain prediction intervals for random effects. We illustrate data cloning in the GLMM context by analyzing the Logistic–Normal model for over-dispersed binary data, and the Poisson–Normal model for repeated and spatial counts data. We consider Normal–Normal and Binary–Normal mixture models to show how data cloning can be used to study estimability of various parameters. We contend that whenever hierarchical models are used, estimability of the parameters should be checked before drawing scientific inferences or making management decisions. Data cloning facilitates such a check on hierarchical models.

KEY WORDS: Bayesian computation; Hierarchical models; Random effects.

1. INTRODUCTION

Linear mixed models (LMM) (Searle, Casella, and McCulloch 1992) and their extension to generalized linear models (GLMM) (McCulloch and Searle 2001) consist of some of the most useful models in statistics. They are widely used in various fields, for example, longitudinal data analysis (Diggle, Liang, and Zeger 1994), epidemiology (Clayton and Kaldor 1987) and ecology and environmental sciences (Clark and Gelfand 2006; Royle and Dorazio 2009). For theoretical discussion of LMM and GLMM, see McCulloch and Searle (2001). Most popular approaches to analyze these models are Bayesian, based on the Markov Chain Monte Carlo (MCMC) algorithm and noninformative priors. (Gilks, Richardson, and Spiegelhalter 1996; Spigelhalter et al. 2004). However, likelihood analysis for these models is difficult (McCulloch 1997; McCulloch and Searle 2001). Likelihood analysis, if used, is usually conducted using approximate likelihood (Breslow and Clayton 1993) or Monte Carlo estimation of the likelihood function (e.g., McCulloch 1997; deValpine 2004).

Recently, Lele, Dennis, and Lutscher (2007) reviewed the difficulties associated with Bayesian and likelihood based approaches and proposed an alternative approach, called data cloning, to compute maximum likelihood estimates and their standard errors for general hierarchical models. See also Doucet, Godsill, and Robert (2002), Kuk (2003), and Jacquier, Johannes, and Polson (2007) for methods similar to data cloning. This approach is based on Bayesian ideas, uses well-known MCMC methodology and can be easily implemented in standard software such as WinBUGS. Data cloning is applicable in most situations where the problem can be formulated as a Bayesian problem and where MCMC can be used to obtain random variates from the posterior distribution. Similar to the Bayesian methodology, data cloning avoids high-dimensional numerical integration and requires neither maximization nor differentiation of a function. It is based only on the computation of the means and the variances. Although data cloning uses a Bayesian formulation and computational techniques, the inferences are based on the classical frequentist paradigm. Unlike the Bayesian inference, these inferences do not depend on the choice of the prior distributions used in the implementation of the MCMC algorithm. The goals of this article are: (1) to use data cloning to analyze GLMM; (2) to provide a simple graphical procedure to determine an adequate number of clones; (3) to provide an algorithm to obtain prediction intervals for random effects; and, most importantly, (4) to provide a simple graphical procedure to determine estimability of the parameters in hierarchical models.

2. NOTATION AND STATISTICAL SET–UP

Let $\mathbf{y}_{(n)} = (y_1, y_2, \ldots, y_n)$ be the data vector where n denotes the sample size. We consider the following general hierarchical model set-up:

Hierarchy 1: $\mathbf{y}_{(n)}|\mathbf{X} = \mathbf{x} \sim h(\mathbf{y}_{(n)}; \mathbf{X} = \mathbf{x}, \boldsymbol{\theta}_1)$.
Hierarchy 2: $\mathbf{X} \sim g(\mathbf{x}; \boldsymbol{\theta}_2)$.

We observe $\mathbf{y}_{(n)}$ whereas $\mathbf{x}$ are unobserved. The parameters of interest are $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$.

The goal of the analysis is to estimate the parameters $\boldsymbol{\theta}$ and predict the unobserved states $\mathbf{x}$. The likelihood function for this hierarchical model set-up is $L(\boldsymbol{\theta}; \mathbf{y}_{(n)}) = \int h(\mathbf{y}_{(n)}|\mathbf{x}; \boldsymbol{\theta}_1)g(\mathbf{x}; \boldsymbol{\theta}_2) \, d\mathbf{x}$. The difficulties associated with using this function for statistical inference are mainly computational: (1) calculation of the likelihood function generally involves high-dimensional integration; (2) obtaining the location of the maximum using numerical search procedures is difficult because of the stochastic nature of the estimated likelihood; and (3) computing standard errors of the resultant estimators involves further difficulties in numerical computation of the second derivatives of the

REVIEW

Frequentist analysis of hierarchical models for population dynamics and demographic data

Perry de Valpine

Abstract Hierarchical models include random effects or latent state variables. This class of models includes state–space models for population dynamics, which incorporate process and sampling variation, and models with random individual or year effects in capture–mark–recapture models, for example. This paper reviews methods for frequentist analysis of hierarchical models and gives an example of a non-Gaussian, potentially nonlinear analysis of Lapwing data using the Monte Carlo kernel likelihood (MCKL) method for maximum-likelihood estimation and bridge sampling for calculation of likelihood values given estimated parameters. The Lapwing example uses the state–space model as part of an integrated population model, which combines survey data with ring-recovery demographic data. The methods reviewed include filtering methods, such as the Kalman filter and sequential Monte Carlo (or particle filtering) methods, Monte Carlo expectation maximization, data cloning, and MCKL. The latter methods estimate the maximum-likelihood parameters but omit a normalizing constant from the likelihood that is needed for model comparisons, such as the Akaike information criterion and likelihood ratio tests. The methods reviewed for normalizing constant calculation include filtering, importance sampling, likelihood ratios from importance sampling, and bridge sampling. For the Lapwing example, a novel combination of MCKL parameter estimation, bridge sampling likelihood calculation, and profile likelihood confidence intervals for an integrated population model is presented to illustrate the feasibility of these methods. A complementary view of Bayesian and frequentist analysis is taken.

Keywords Bridge sampling · Data cloning · Integrated population model · Monte Carlo expectation maximization · Monte Carlo kernel likelihood · Normalizing constant · Particle filter · State-space model · *Vanellus vanellus*

Introduction

Many types of ecological data can be statistically modeled by recognizing multiple sources of variation in the processes that led to the data, including both ecological and data-sampling variation (Clark 2007; Royle and Dorazio 2008; Cressie et al. 2009; King et al. 2009). For example, a *state–space model* for a time-series of abundance data includes unknown true abundances, stochastic relationships between true abundances at one time and the next, and stochastic relationships between true abundances and the data (Schnute 1994; de Valpine and Hastings 2002). Another example is *random effects models* for capture–mark–recapture (CMR), ring-recovery, or related data, in which year effects, between-individual variation, or other sources of variation may be modeled as following some distribution (Burnham and White 2002; Cam et al. 2002; Link et al. 2002; Royle and Link 2002; Barry et al. 2003; Gimenez and Choquet 2010).

What these models have in common is that they include statistical relationships between data and unknown quantities, such as true abundances in a state–space model or random effects values in a CMR model, that in turn have statistical relationships to model parameters. Indeed, a

P. de Valpine (✉)
Department of Environmental Science, Policy and Management, University of California, Berkeley, CA, USA
e-mail: pdevalpine@berkeley.edu

dclone: Data Cloning in R

by Péter Sólymos

Abstract   The **dclone** R package contains low level functions for implementing maximum likelihood estimating procedures for complex models using data cloning and Bayesian Markov chain Monte Carlo methods with support for JAGS, WinBUGS and OpenBUGS.

Introduction

Hierarchical models, including generalized linear models with mixed random and fixed effects, are increasingly popular. The rapid expansion of applications is largely due to the advancement of the Markov chain Monte Carlo (MCMC) algorithms and related software (Gelman et al., 2003; Gilks et al., 1996; Lunn et al., 2009). Data cloning is a statistical computing method introduced by Lele et al. (2007). It exploits the computational simplicity of the MCMC algorithms used in the Bayesian statistical framework, but it provides the maximum likelihood point estimates and their standard errors for complex hierarchical models. The use of the data cloning algorithm is especially valuable for complex models, where the number of unknowns increases with sample size (i.e. with latent variables), because inference and prediction procedures are often hard to implement in such situations.

The **dclone** R package (Sólymos, 2010) provides infrastructure for data cloning. Users who are familiar with the Bayesian methodology can instantly use the package for maximum likelihood inference and prediction. Developers of R packages can build on the low level functionality provided by the package to implement more specific higher level estimation procedures for users who are not familiar with the Bayesian methodology. This paper demonstrates the implementation of the data cloning algorithm, and presents a case study on how to write high level functions for specific modeling problems.

Theory of data cloning

Imagine a hypothetical situation where an experiment is repeated by k different observers, and all k experiments happen to result in exactly the same set of observations, $y^{(k)} = (y, y, \ldots, y)$. The likelihood function based on the combination of the data from these k experiments is $L(\theta, y^{(k)}) = [L(\theta, y)]^k$. The location of the maximum of $L(\theta, y^{(k)})$ is exactly equals the location of the maximum of the function $L(\theta, y)$, and the Fisher information matrix based on this likelihood is k times the Fisher information matrix based on $L(\theta, y)$.

One can use MCMC methods to calculate the posterior distribution of the model parameters (θ) conditional on the data. Under regularity conditions, if k is large, the posterior distribution corresponding to k clones of the observations is approximately Normal with mean $\hat{\theta}$ and variance $1/k$ times the inverse of the Fisher information matrix. When k is large, the mean of this posterior distribution is the maximum likelihood estimate and k times the posterior variance is the corresponding asymptotic variance of the maximum likelihood estimate if the parameter space is continuous. When some of the parameters are on the boundaries of their feasible space (Stram and Lee, 1994), point estimates can be correct, but currently the Fisher information cannot be estimated correctly by using data cloning. This is an area for further research, but such situations challenge other computing techniques as well.

Data cloning is a computational algorithm to compute maximum likelihood estimates and the inverse of the Fisher information matrix, and is related to simulated annealing (Brooks and Morgan, 1995). By using data cloning, the statistical accuracy of the estimator remains a function of the sample size and not of the number of cloned copies. Data cloning does not improve the statistical accuracy of the estimator by artificially increasing the sample size. The data cloning procedure avoids the analytical or numerical evaluation of high dimensional integral, numerical optimization of the likelihood function, and numerical computation of the curvature of the likelihood function. Interested readers should consult Lele et al. (2007, 2010) for more details and mathematical proofs for the data cloning algorithm.

The data cloning algorithm

Let us consider the following Poisson generalized linear mixed model (GLMM) with a random intercept for i.i.d. observations of Y_i counts from $i = 1, 2, \ldots, n$ localities:

$$
\begin{aligned}
\alpha_i &\sim \text{Normal}(0, \sigma^2) \\
\lambda_i &= \exp(\alpha_i + X_i^\mathsf{T} \beta) \\
(Y_i \mid \lambda_i) &\sim \text{Poisson}(\lambda_i).
\end{aligned}
$$

The corresponding code for the simulation looks like ($\beta = (1.8, -0.9)$, $\sigma = 0.2$, $x_i \sim \text{U}(0,1)$):

```
> library(dclone)
> set.seed(1234)
> n <- 50
> beta <- c(1.8, -0.9)
> sigma <- 0.2
> x <- runif(n, min = 0, max = 1)
> X <- model.matrix(~ x)
> alpha <- rnorm(n, mean = 0, sd = sigma)
```

Assessing parameter identifiability in phylogenetic models using Data Cloning

JOSÉ MIGUEL PONCIANO[1], J. GORDON BURLEIGH[1], EDWARD L. BRAUN[1], AND MARK L. TAPER[1]

[1]*Department of Biology, University of Florida, Gainesville, FL,32611, United States of America.*

Corresponding author: José Miguel Ponciano, Department of Biology, University of Florida, Gainesville, FL,32611, United States of America; E-mail: josemi@ufl.edu.

Abstract.—

The success of model-based methods in phylogenetics has motivated much research aimed at generating new, biologically informative models. This new computer-intensive approach to phylogenetics demands validation studies and sound measures of performance. To date there has been little practical guidance available as to when and why the parameters in a particular model can be identified reliably. Here, we illustrate how Data Cloning (DC), a recently developed methodology to compute the Maximum Likelihood estimates along with their asymptotic variance, can be used to diagnose structural parameter non-identifiability (NI) and distinguish it from other parameter estimability problems, including when parameters are structurally identifiable, but are not estimable in a given data set (INE), and when parameters are identifiable, and estimable, but only weakly so (WE).The application of the DC theorem uses well-known and widely used

7. Concluding remarks

Hierarchical models offer a quantum leap in the complexity of statistical modeling

Statistical inferences (fitting the models to data, model selection, etc) for HMs can be Bayesian or frequentist

> Bayesian inference uses the probability of personal beliefs as summarized through prior and posterior distributions on parameter spaces. Prior personal beliefs are mixed with the data to produce the inferences.

> Frequentist inference uses the long-run probability of data outcomes (probability on sample spaces).

Both Bayesian and frequentist inferences for HMs can be calculated with the MCMC intensive simulation algorithms

Frequentist ML estimation with the method of data cloning provides a simple tool for diagnosing problems with parameter estimability

Hierarchical models are not Bayesian

brian@uidaho.edu