*Ecology Letters*, (2007) **10**: 551–563 doi: 10.1111/j.1461-0248.2007.01047.x

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.

## Keywords

Bayesian statistics, density dependence, Fisher information, frequentist statistics, generalized linear mixed models, hierarchical models, Markov chain Monte Carlo, state-space models, stochastic population models.

*Ecology Letters* (2007) 10: 551–563

## 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 hierarchical 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

**Table 1** Sample of recent ecological publications featuring hierarchical statistical models

*Capture-recapture:* George & Robert (1992); Feinberg *et al.* (1999); Brooks *et al.* (2000); Basu & Ebrahimi (2001); Rivot & Prévost (2002)

*Fisheries stock assessment:* Meyer & Millar (1999a,b); Millar & Meyer (2000a,b)

*Reaction-diffusion models of spatial spread:* Clark *et al.* (2003); Wikle (2003)

*Geospatial models of species and habitats:* Gelfand *et al.* (2005)

*Estimating trends in spatially distributed populations:* Wikle *et al.* (1998); Link & Sauer (2002); Sauer & Link (2002); Thogmartin *et al.* (2004); Link *et al.* (2006)

*Forest growth and yield modelling:* Green *et al.* (1999); Radtke *et al.* (2002)

*Modelling structured, density-dependent populations:* Clark (2003); Buckland *et al.* (2004); Clark *et al.* (2005); Newman *et al.* (2006)

*Estimating different sources of variability in population time series:* Clark & Bjørnstad (2004)

*Analysing species abundance distributions in biodiversity*: Etienne & Olff (2005)

to different values of the parameters. Finally, interpretation of the credible intervals when non-informative or objective priors are used is still controversial in statistics (Barnett 1999). Indeed, in ecology, the increased use of Bayesian inference has been partly pragmatic, because the Bayesian/MCMC approach to date has provided the only practical solution for fitting various complex hierarchical models.

In comparison with Bayesian analysis, likelihood-based statistical inference for hierarchical models can be extremely difficult. Computing the likelihood function involves high-dimensional integration over the unobserved variables (McCulloch & Searle 2001; De Valpine & Hastings 2002; Dennis *et al.* 2006; Lele 2006). Likelihood-based approaches to inference for hierarchical models have involved either approximations or computer-intensive simulation algorithms. McCulloch (1997) reviewed various likelihood approaches for generalized linear models with mixed effects, and Robert & Casella (2004) provided an excellent review of simulation approaches to statistical inference for hierarchical models. Under the Bayesian setup with uninformative priors, the mode or mean of the posterior distribution can be a reasonable large-sample approximation to the maximum likelihood (ML) estimator (Karim & Zeger 1992). Approximation methods such as penalized quasi-likelihood (Breslow & Clayton 1993) and composite likelihood (Heagerty & Lele 1998; Lele 2006) can produce useful, approximate estimates for limited classes of models. More recently, Laplace approximation (Breslow & Lin 1995) has been combined with automatic differentiation to produce approximate ML estimates for hierarchical models (Skaug & Fournier 2006), but statistical investigators have warned that the approximation can be poor for some nonlinear models (Breslow & Lin 1995; McCulloch 1997; Carlin & Louis 2000). Kitagawa (1987) introduced a discretization approach to simulating the likelihood in non-Gaussian state-space models. Geyer & Thompson (1992, 1995) showed how MCMC simulations could be used directly for estimating likelihood ratios, from which ML estimates can be obtained. George & Thompson (2003), following Bennett (1976), suggested using a prior in Bayesian MCMC calculations and then dividing the posterior by the prior, yielding the likelihood function up to an unknown constant. De Valpine (2003, 2004) used priors in Bayesian MCMC calculations, approximating the likelihood surface up to an unknown constant with weighted kernel density estimation of the posteriors. The methods of path sampling and bridge sampling can also be used in conjunction with MCMC simulations to estimate likelihood ratios (Gelman & Meng 1998). Computer-intensive likelihood simulations have been featured in some ecological studies (De Valpine & Hastings 2002; Ponciano *et al.* 2007).

When the likelihood function must be simulated, computational approaches to ML estimation are difficult to implement. The ML calculations involve the computationally challenging task of maximizing a noisy function, which requires proper use of stochastic optimization routines (Spall 2003). The ratios of random variables in the likelihood ratios tend to amplify the noisiness of the function making the task of locating the maxima of the random functions tricky. The estimates are slow to converge and can require considerable hands-on attention, restarts and patience. Finally, most existing methods are not convenient to implement using current statistical software packages.

In this study, we introduce a simple method for calculating ML estimates for hierarchical models using the MCMC algorithms. The method is an adaptation, for hierarchical models, of a computational ML approach developed by Robert (1993) and is related to simulated annealing (Brooks & Morgan 1995). The method adopts the full Bayesian setup for MCMC statistical calculations, but uses the framework only as a device for likelihood calculations. In contrast to the Bayesian inferences, these inferences are completely invariant to the choice of the prior distributions. Standard software packages for Bayesian MCMC calculations, such as the WINBUGS software (Spiegelhalter *et al.* 2004), can be used to produce ML estimates after a simple adjustment of the inputs. The calculations require merely computing sample mean values and variances and not numerical maximization or differentiation of a noisy function. Theoretically, the method will provide the location of the global maximum and not just a local maximum, although the asymptotic conditions of theory cannot always be approximated in practice. The global maximization property is especially important in that hierarchical models can have multimodal likelihood functions (Dennis *et al.* 2006). We provide for ecologists an accessible explanation of the data cloning method with the goal of making it immediately available for ecological applications. A proof of why the method works, based on the concept of iterative maps, appears in the Appendix. We illustrate the technique by analysing two ecological examples for which ML estimation has been problematic.

## THE DATA CLONING METHOD

The method we propose is based on an idea we call data cloning. The idea is simple: construct a full Bayesian model of the problem, complete with fully specified, proper prior distributions for unknown parameters, but instead of using the likelihood for the observed data, use the likelihood corresponding to $k$ copies (clones) of the data, where $k$ is large and the copies are assumed to be independent of each other. The posterior is then calculated with the usual MCMC approach. The mean of the resulting posterior distribution equals the ML estimate, and $k$ times the variance of the posterior equals the asymptotic variance of the ML estimate.

In the following description, we assume some familiarity with the use of likelihood functions (Dennis & Taper 1994; Dennis *et al.* 1995; Hilborn & Mangel 1997) and Bayesian MCMC methods (Casella & George 1992; Chib & Greenberg 1995; Meyer & Millar 1999b; Clark & Bjørnstad 2004) in ecological applications.

Suppose observations $\underline{Y} = (Y_1, Y_2, \ldots, Y_n)$ arise from the following hierarchical statistical model:

$$\underline{Y} \sim f(\underline{y}|\underline{X}, \varphi),$$

$$\underline{X} \sim g(\underline{x}|\theta),$$

where $f$ and $g$ are joint probability density functions (pdfs), $\underline{X}$ is a vector of random quantities or processes affecting the observations, $\varphi = (\varphi_1, \varphi_2, \ldots, \varphi_q)$ is a vector of unknown fixed parameters affecting the observations, and $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$ is a vector of unknown fixed parameters related to the process $\underline{X}$.

For example, in a state-space model of population abundance, $\underline{X}$ is a vector containing a time series of unobserved population abundances governed by a stochastic growth model with joint pdf $g$, and $\underline{Y}$ is a vector containing the observed or estimated population abundances, with the joint pdf $f$ being a model of how $\underline{Y}$ arises as a measurement-error-corrupted version of $\underline{X}$. As another example, in a mixed effects analysis of variance model, $\underline{Y}$ is the vector of response variables, and $\underline{X}$ is a vector of random mean values. A third example is a closed population capture-recapture model, in which $\underline{Y}$ is a matrix of capture histories (rows of 0s and 1s), and $\underline{X}$ is a vector of random catch probabilities. In the hierarchical models literature, the random quantities in $\underline{X}$ are variously called random effects, latent variables or system states.

The likelihood function for the general hierarchical model described above is given by

$$L(\theta, \varphi; \underline{y}) = \int f(\underline{y}|\underline{X}, \varphi) g(\underline{X}|\theta) d\underline{X},$$

where $\underline{y}$ is the observed data (the realized outcome of the random variable $\underline{Y}$). The ML estimates of the parameters, which we denote by $(\hat{\theta}, \hat{\varphi}) = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p; \hat{\varphi}_1, \hat{\varphi}_2, \ldots, \hat{\varphi}_q)$, are the values of $(\theta, \varphi) = (\theta_1, \theta_2, \ldots, \theta_p; \varphi_1, \varphi_2, \ldots, \varphi_q)$ that jointly maximize the likelihood function. Clearly, the calculation of the likelihood function and ML estimates involves the computationally daunting task of high-dimensional integration.

The Bayesian approach completely circumvents the problem of high-dimensional integration. The Bayesian approach begins by assuming that, instead of being fixed but unknown quantities, the parameters $(\theta, \varphi)$ are random variables. The joint statistical distribution corresponding to these random variables is called a prior distribution. The prior distribution quantifies the investigator's pre-data beliefs about the different values of the parameters. The prior distribution is then mixed with the likelihood function, by using Bayes' rule, to form the joint posterior distribution representing the investigator's post-data beliefs about the parameters. Notice that, in the hierarchical model setup, the variables $\underline{X}$ are also unknown and random. Let us denote the joint prior distribution on the parameters $(\theta, \varphi)$ by $\pi(\theta, \varphi)$. According to Bayes' rule, the joint posterior distribution on the unknown quantities $(\theta, \varphi, \underline{X})$ conditional on the observed data $\underline{y}$ is given by

$$h(\theta, \varphi, \underline{X}|\underline{y}) = \frac{f(\underline{y}|\underline{X}, \varphi) g(\underline{X}|\theta) \pi(\theta, \varphi)}{\int f(\underline{y}|\underline{X}, \varphi) g(\underline{X}|\theta) \pi(\theta, \varphi) d\underline{X} d\theta d\varphi}.$$

The marginal posterior distribution for the parameters, denoted $\pi(\theta, \varphi \lfloor \underline{y})$, is simply obtained by integrating the posterior $h(\theta, \varphi, \underline{X} \lfloor \underline{y})$ over $\underline{X}$. It may appear that we are replacing a problem of high-dimensional integration for likelihood calculations by an even higher dimensional integration problem in the denominator of the posterior distribution $h(\theta, \varphi, \underline{X} \lfloor \underline{y})$ and another high-dimensional integral to obtain the marginal posterior distribution $\pi(\theta, \varphi \lfloor \underline{y})$. However, that is not the case as we explain below.

Markov chain Monte Carlo algorithms are computational tools that allow generation of random numbers from the posterior distribution $h(\theta, \varphi, \underline{X} \lfloor \underline{y})$ using only the numerator of the expression without ever calculating the integral in the denominator. We refer the reader to standard sources (Casella & George 1992; Chib & Greenberg 1995; Gilks *et al.* 1996; Robert & Casella 2004) for details of implementation. Notice that the numerator involves no integration. Let us denote the MCMC-generated random numbers by $(\theta, \varphi, \underline{X})_j$, $j = 1, 2, \ldots, B$. Here $B$, the number of observations generated from $h(\theta, \varphi, \underline{X}|\underline{y})$, is large enough (say, at least 10 000) to provide a good estimate of $h(\theta, \varphi, \underline{X}|\underline{y})$. With $h(\theta, \varphi, \underline{X}|\underline{y})$ obtained, computing the posterior distribution $\pi(\theta, \varphi|\underline{y})$ might seem to require integration over the variables $\underline{X}$ in $h(\theta, \varphi, \underline{X}|\underline{y})$. Fortunately, such integration is unnecessary. The marginal posterior distribution of $(\theta, \varphi)$ is found by simply discarding the $\underline{X}$ component of the random numbers $(\theta, \varphi, \underline{X})_j$, leaving $(\theta, \varphi)_j$, $j = 1, 2, \ldots, B$. Similarly, the mean values and variances of $\pi(\theta, \varphi|\underline{y})$ are simply the sample mean values and sample variances of the random numbers $(\theta, \varphi)_j$, $j = 1, 2, \ldots, B$. The process of simulating the marginal posterior distribution thus involves no integration.

Now we explain the data cloning algorithm heuristically. Imagine that an individual performs the statistical experiment underlying the observations $\underline{y}$ not just once but rather $k$ times simultaneously and independently. Suppose in addition that each of the $k$ experimental replicates produces, by happenstance, exactly the *same* result $\underline{y}$. The new likelihood function for the $k$ data 'clones' is the original likelihood, $L(\theta, \varphi; \underline{y})$, raised to the $k$th power: $[L(\theta, \varphi; \underline{y})]^k$. Note that the cloned data likelihood has the same location of the maximum, namely the ML estimates $(\hat{\theta}, \hat{\varphi})$, as that of the original likelihood. Now suppose the investigator obtains a Bayesian posterior, say $h^{(k)}(\theta, \varphi, \underline{X}|\underline{y})$, along with a marginal posterior $\pi^{(k)}(\theta, \varphi|\underline{y})$, using a prior distribution $\pi(\theta, \varphi)$ and the cloned data likelihood. It turns out that if $k$ is large, the marginal posterior $\pi^{(k)}(\theta, \varphi|\underline{y})$ will be concentrated around the ML estimates $(\hat{\theta}, \hat{\varphi})$ (proof in Appendix).

In fact, such a cloned data posterior provides not only ML estimates, but also their asymptotic standard errors as well. A well-known result from statistical theory states that as the sample size in a likelihood function increases, the posterior distribution converges to a multivariate normal distribution centred at the ML estimates (Walker 1969). Walker's theorems are the basis of the often-mentioned assertion that frequentist and Bayesian inferences become similar for large sample sizes. By modifying Walker's theorems to cover the 'deterministic' nature of cloned data likelihoods, S.R. Lele (unpublished work) proved the following result: as $k$ becomes large, $\pi^{(k)}(\theta, \varphi|\underline{y})$ converges to a multivariate normal distribution with mean equal to the ML estimate $(\hat{\theta}, \hat{\varphi})$ and variance equal to $\frac{1}{k} I^{-1}(\hat{\theta}, \hat{\varphi})$, where $I(\hat{\theta}, \hat{\varphi})$ is the Fisher information matrix corresponding to the original likelihood function. The Fisher information matrix is related to the Hessian of the log-likelihood function and represents the average curvature of the log-likelihood near its maximum. The inverse of the Fisher information matrix contains the asymptotic variances and covariances of the ML estimates (see Stuart & Ord 1987). Thus, the scaled variances and covariances of the posterior distribution $\pi^{(k)}(\theta, \varphi|\underline{y})$ can be used as estimates of the asymptotic variances and covariances.

Of course, in reality we do not have $k$-independent replications of the same experiment yielding exactly the same data. However, the thought experiment can be mimicked using computers as described in the algorithm below.

### Step 1

Create a $k$-cloned data set $\underline{y}^{(k)} = (\underline{y}, \underline{y}, \ldots, \underline{y})$ where the observed data vector is repeated $k$ times.

### Step 2

Using an MCMC algorithm, generate random numbers from the posterior distribution that is based on a prior $\pi(\theta, \varphi)$, the appropriate hierarchical model structure, and the cloned data vector $\underline{y}^{(k)} = (\underline{y}, \underline{y}, \ldots, \underline{y})$ where the $k$ copies of $\underline{y}$ are assumed to be independent of each other. Virtually any proper prior distribution can be used. The calculations can be performed with WINBUGS (Spiegelhalter *et al.* 2004).

Specifically, a simple Metropolis-Hastings scheme (Hastings 1970; McCulloch 1997; Robert & Casella 2004) to accomplish the MCMC calculations is as follows (proof in Appendix). (a) Generate $(\theta, \varphi)^*$ from $\pi(\theta, \varphi)$, and set these as the initial parameter values: $(\theta, \varphi)_1 = (\theta, \varphi)^*$. (b) Generate $k$ values of $\underline{X}$, say $\underline{X}^{(1)}, \underline{X}^{(2)}, \ldots, \underline{X}^{(k)}$, from $g(\underline{X}|\theta^*)$. (c) Calculate the product

$$q* = f\left(\underline{y}|\underline{X}^{(1)}, \varphi*\right) f\left(\underline{y}|\underline{X}^{(2)}, \varphi*\right) \ldots f\left(\underline{y}|\underline{X}^{(k)}, \varphi*\right),$$

and set this as the initial $q$ value: $q_1 = q^*$. (d) Repeat the simulations of steps (a) and (b), obtaining new values $(\theta, \varphi)^{\#}$ and $q^{\#}$. (e) Generate a uniform(0,1) random variable $U$, and calculate $p = \min[1, (q^{\#}/q_j)]$, where the initial value of $j$ will be 1. If $U > p$, set $(\theta, \varphi)_{j+1} = (\theta, \varphi)_j$; otherwise set $(\theta, \varphi)_{j+1} = (\theta, \varphi)^{\#}$. (f) Repeat (d) and (e), many times. According to theory (Hastings 1970), the resulting values $(\theta, \varphi)_j$, $j = 1, 2, \ldots, B$, commencing after a 'burn-in' period (typically 1000 or more), have been generated from the marginal posterior distribution $\pi(\theta, \varphi|\underline{y})$. The $k$ sets of latent variables generated each step from $g(\underline{X}|\theta)$ are just discarded. We chose the Metropolis-Hastings algorithm to describe in detail here for its simplicity; other Bayesian MCMC algorithms besides the Metropolis-Hastings are known to give faster convergence to the posterior distribution and are available in WINBUGS.

### Step 3

Compute the sample mean values and sample variances of the values $(\theta, \varphi)_j$, $j = 1,2, \ldots, B$ generated from the marginal posterior. The ML estimates of $(\theta, \varphi)$ correspond to the posterior mean values and the approximate variances of the ML estimates correspond to $k$ times the posterior variances.

In the algorithm, the key facts are that the cloned data likelihood (a high-dimensional integral) does not have to be evaluated, and that no numerical maximization is required. The main function to be evaluated is just the conditional pdf given by $f(\underline{y}|\underline{X},\varphi)$. The ML estimates are just averages of large numbers of computer-generated random variables.

For implementing the algorithm, number of clones $k$ can be taken as large as necessary for good approximation of the ML estimates. While data cloning in theory produces the global maximum as $k$ becomes infinite, in practice $k$ is finite, and one should take measures to reduce the possibility that the algorithm becomes 'stuck' in a persistent local maximum. As a strategy, we suggest rerunning the algorithm with several different starting prior distributions, and with increasing values of $k$, until the results from the different starting priors are in agreement. The posterior mean values should converge to stable, common values for different priors when $k$ is large enough. We note that the free R language for statistical computing (R Core Development Team 2006) has a shell (R2WINBUGS: Sturtz *et al.* 2005) for running WINBUGS, which might serve to automate the increasing $k$ values via looping. Also, we find that using fairly informative or even disinformative prior distributions, instead of flat ones, helps speed convergence of the posterior mean values (Examples, below).

We emphasize that the estimates resulting from data cloning are full ML estimates, resulting from maximizing the full likelihood function in which the random effects have been 'integrated out'. Recall that Bayesian MCMC methods integrate out the random effects by simulating random variates from the joint distribution of the parameters and the random effects, and then just discarding the random effects. Data cloning is just a Bayesian MCMC method, using a different (cloned) likelihood. Note that in each step of the above-described Metropolis-Hastings algorithm, $k$ sets of the latent variables $(\underline{X}^{(1)}, \underline{X}^{(2)}, \ldots, \underline{X}^{(k)})$ are generated. The resulting posterior is the joint distribution of the parameters as well as $k$ sets of latent variables, and of course the generated latent variables are just thrown out. The random effects *are* integrated out, just as in Bayesian MCMC methods.

Note that increasing the number of clones only improves the *numerical accuracy* of the approximation to the ML estimates and not the statistical accuracy. As well, the length of the MCMC run only improves the numerical accuracy. Statistical accuracy of the estimator is a function of the amount of information in the data vector $\underline{y}$ and depends on factors such as sample size and model quality. Also, the standard errors and confidence intervals provided by data cloning are large-sample approximations, and whether they have correct nominal coverage properties depends on the sample size and not on the number of clones. Data cloning does not make up for lack of data. Yet, there is nothing irregular or unscientific about cloning the data in this algorithm. It is simply a calculation trick for obtaining ML estimates.

## EXAMPLES

The two examples we develop are state-space population models, in which the time series observations are influenced by both environmental process noise and observation or estimation error. The first example is an initial test case with a known likelihood function; the second example, featuring an intractable likelihood function and missing data, would pose a tough challenge to earlier numerical ML methods.

### Gompertz state-space model

The Gompertz state-space (GSS) model is a stochastic, density-dependent model for time series observations of population abundances (Dennis *et al.* 2006). Let $N_t$ denote the true population abundance at time $t$ and let $X_t = \log(N_t)$ be the logarithm of the unobserved population abundance at time $t$ $(t = 0, 1, \ldots, q)$. In practice, the true population abundances are usually not available, but estimates of these true population abundances, based on some sampling scheme, are available. Let $Y_t$ denote an estimate of $X_t$ obtained by the investigator. A stochastic Gompertz model for the underlying true population abundance is represented by

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

where $a$ and $c$ are constants, $E_t$ has a normal$(0,\sigma^2)$ distribution. The parameter $a$ is the intrinsic growth rate and the parameter $c$ is the density dependence parameter. The variance parameter $\sigma^2$ measures the intensity of environmental variability (process noise) in the system. The observations $Y_t$ are related to the true population abundances $X_t$ by the model

$$Y_t = X_t + F_t,$$

where $F_t$ has a normal$(0,\tau^2)$ distribution. The variance parameter $\tau^2$ quantifies the amount of measurement error or observation error. The unknown parameters in the model are $a$, $c$, $\sigma$ and $\tau$. The likelihood function is a multivariate normal distribution, and is identical to that of a mixed effects analysis of variance model with repeated measures on one subject (Dennis *et al.* 2006).

The GSS model serves as an excellent initial test case for the data cloning method. Although the likelihood function for this model can be written analytically, ML estimation nonetheless has problematic aspects, the main one being that the likelihood function routinely has multiple peaks. In the following, we employ the GSS model in a reanalysis of a population data set studied by Dennis *et al.* (2006: the data values appear in the legend of their Table 1). The data consist of time series abundances of American Redstart (*Setophaga ruticilla*) at a particular location, from the North American Breeding Bird Survey (BBS).

The GSS likelihood function for the BBS data set provides a graphical image of how the data cloning method works. The likelihood in question is multimodal and ridge-shaped; a profile likelihood calculated as a function of $\sigma$ appears in Fig. 1 ($k = 1$). The shape reflects the weak identifiability of model parameters; the lesser mode corresponds to a model with no observation error ($\tau = 0$). The data-cloned GSS likelihood is just a multivariate normal pdf raised to the $k$th power; additional profile likelihoods for increasing $k$ appear in Fig. 1. Note that the data-cloned likelihoods have the same peak locations as the original. Through the multiplications, data cloning stretches the likelihood and magnifies the highest peak. When used in a Bayesian analysis, such a data-cloned likelihood completely swamps the prior and concentrates the posterior around the location of the highest peak of the likelihood. The data cloning method thus represents a global maximization method.

For the BBS data we calculated the exact (to four digits) ML estimates using numerical maximization of the analytical likelihood function. For comparison, we used the data cloning method to calculate three sets of ML estimates with three widely different sets of priors. With just 240 clones of the original data, the data cloning results are nearly equal to the exact ML results (Table 2). The standard errors reported with the exact ML estimates in the first column of Table 2 are based on the inverse of the Fisher information matrix (Stuart & Ord 1987). The (observed) Fisher information matrix was obtained by numerically computing the matrix of second derivatives of the log-likelihood function at the ML values. The data-cloned ML estimates and their standard errors in the next three columns were obtained as the mean and scaled variance of the posterior distribution. The posterior distributions were obtained with the WINBUGS software (Spiegelhalter *et al.* 2004). Among the three different sets of prior distributions used, some were non-informative and some were quite disinformative (highly informative but wrong; Table 2). For each set of prior distributions, a burn-in period of 1000 MCMC steps was used to equilibrate the Markov chain, and then 10 000 values were generated from the posterior distribution. Each run required *c.* 20 min on a 2.4 Ghz Pentium 4 processor. All three sets of priors yielded ML estimates close to the exact values. The ML estimates were similar among the

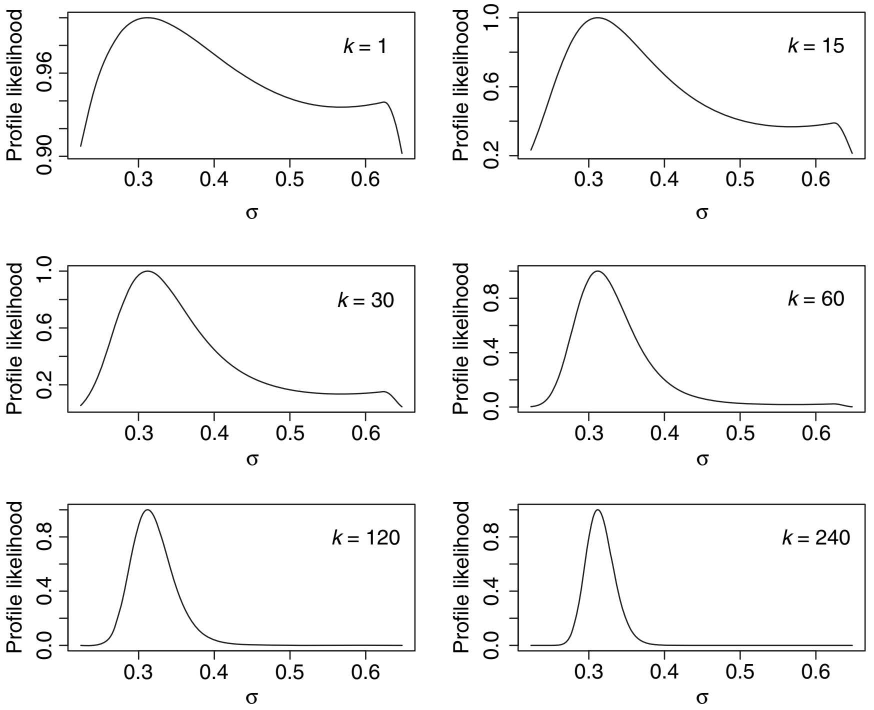

**Figure 1** Profile likelihood, divided by its maximum value, for the Gompertz state-space model as a function of the process noise parameter $\sigma$, plotted for increasing number $k$ of data clones. Data are yearly American Redstart (*Setophaga ruticilla*) counts at a location in the North American Breeding Bird Survey. Data and likelihood function formula are listed by Dennis *et al.* (2006).

**Table 2** Maximum likelihood estimates (and standard errors) calculated for the parameters $a$, $c$, $\sigma$ and $\tau$ 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) |
| $\sigma$ | 0.3119 (0.2784) | 0.3132 (0.2751) | 0.3217 (0.2262) | 0.3207 (0.2934) |
| $\tau$ | 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).

three different sets of priors. The priors with smaller variances produced posteriors that were closer to the actual ML estimates, even if such priors were substantially biased. As well, the standard errors obtained with the data cloning method were quite close to the estimates of the ML standard errors arising from the Fisher information matrix (Table 2).

## Stochastic Ricker model with Poisson errors

Gause's (1934) laboratory experiments on the population growth of two *Paramecium* species (*P. aurelia*, *P. caudatum*) are the iconic illustrations of sigmoidal growth curves in ecology textbooks. Although Gause and textbooks alike plotted mean abundance across replicate cultures, the individual replicate cultures display considerable stochastic variability (Fig. 2). The variability is a combination of stochasticity in the process itself as well as observation error in the data. Earlier analyses have used either process noise or observation error, but not both (Pascual & Kareiva 1996). Gause sampled the microbe populations by counting the number of cells in a small volume (0.5 cm$^3$) of growth media removed from well-mixed cultures. The sampling mechanism can be reasonably modelled with a Poisson distribution, with mean equal to the concentration of cells per volume sampled in the culture.

We analysed Gause's experimental data (species growing separately) with a Ricker-Poisson state-space model. The underlying population growth process in the Ricker-Poisson is a stochastic version of the Ricker model (Dennis & Taper 1994), and the sampling error model is Poisson. The Ricker-Poisson state-space model is given by

$$N_t = N_{t-1} \exp(a + bN_{t-1} + E_t),$$

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

Here, $N_t$ is population abundance (cells per volume) of a culture at time $t$ (days), $O_t$ is the cells per volume in the sample at time $t$, and the process noise $E_t$ has a normal(0,$\sigma^2$) distribution. For this model, the parameter $a$ (not the coefficient parameter $b$) measures the strength of density dependence, because it is related to the eigenvalue of the deterministic one-dimensional map near equilibrium (May & Oster 1976). The parameter $b$ serves to scale the level of the equilibrium population size. We define one unit of volume to be the volume of a sample, 0.5 cm$^3$. The initial cell concentration in the cultures was set experimentally and is therefore treated as a known parameter in the model. All cultures were started with exactly two cells per unit volume.

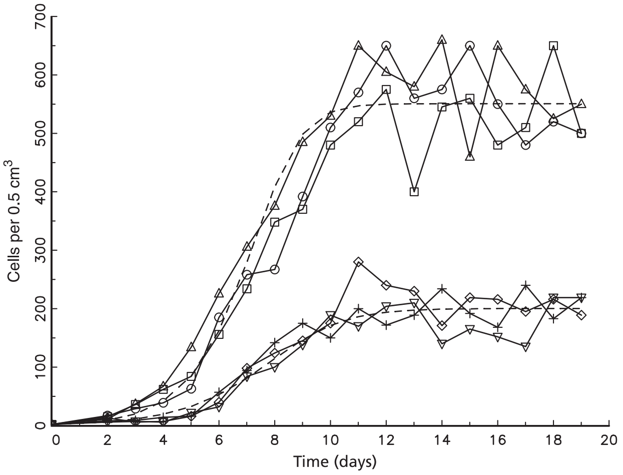

**Figure 2** Population abundances of two *Paramecium* species, three replicate cultures each (solid lines), from Gause (1934: Appendix I, Table 3), plotted with solution trajectories from deterministic Ricker population growth model (dashed lines). Upper three time series: *P. aurelia*. Lower three time series: *P. caudatum*. Ricker solution trajectories use maximum likelihood parameter estimates from the Ricker-Poisson state-space model, computed with data cloning for the combined replicates (Table 3).

The likelihood function for time series observations arising from this stochastic Ricker-Poisson model cannot be written down in an analytical form. To complicate matters, Gause did not record data for any of the cultures at time

$t = 1$. However, with the data cloning method, fitting the Ricker-Poisson model to Gause's data by ML estimation becomes straightforward. We used $k = 10$ clones. The missing data were treated as latent variables (as in Clark & Bjørnstad 2004). For each model fitting, two parallel MCMC chains were generated, each for total of 5000 iterates. The two parallel runs help insure that the Markov chain does not get 'stuck' in a long-term transient state. The burn-in period to equilibrate each Markov chain was 2000, and the next 3000 iterates were used to generate parameter values from the posterior distribution. The 3000 parameter values from each chain were combined to calculate the ML estimates. The sample mean values and standard deviations of the 6000 generated parameter values formed the data-cloned ML estimates and their standard errors. Analyses were rerun with different prior distributions and $k$ values in order to insure that the results were invariant to the choice of priors and that the number of clones was sufficiently large.

We fitted the model separately to each species. Within each species, we analysed separately the three replicate populations that were cultured for 19 days, omitting from the analysis a fourth replicate of *P. caudatum* that was not maintained for as long a period. For each species, we also fitted the model based on the three replicates together. For this, the three replicates were assumed to be independent of each other. Although the likelihood cannot be written down in closed form, it turned out to be well behaved and the calculations converged quickly. The combined replicates estimation required *c.* 22 s running time (10 clones, 2.4 Ghz Pentium 4 processor). For comparison, we fitted the stochastic Ricker model without observation error starting at time $t = 2$ using conditional least squares (Dennis & Taper 1994).

Within each species, the ML parameter estimates for the state-space model among the replicates are quite similar (Table 3). However, substantial differences between the state-space model estimates and the process-error-only model estimates can be discerned. The estimates of the parameter $a$ are consistently larger under the state-space model, and the estimates of $\sigma$ are consistently smaller. The state-space model estimates stronger density dependence and lower process variability than the model with only process error. A plot of the data for the individual replicates together with the Ricker map trajectory estimated from the pooled replicates forms a more contemporary image of the contrasts between deterministic and stochastic forces in population growth (Fig. 2).

## DISCUSSION

We have shown how the Bayesian formulation and MCMC algorithms can be redirected by the data cloning method to produce ML parameter estimates, their standard errors and confidence intervals for complex ecological models. With data cloning, analysis of hierarchical models is no longer Bayesian by default. The choice of the Bayesian or

**Table 3** Maximum likelihood estimates (and standard errors) for the parameters $a$, $b$ and $\sigma$ in the stochastic Ricker-Poisson state-space model and in the stochastic Ricker model with no observation error

| | Replicate 1 | | Replicate 2 | | Replicate 3 | |
|---|---|---|---|---|---|---|
| | Ricker | Ricker-Poisson | Ricker | Ricker-Poisson | Ricker | Ricker-Poisson |
| *Paramecium aurelia* | | | | | | |
| $a$ | 0.595 | 0.735 (0.053) | 0.667 | 0.771 (0.059) | 0.822 | 0.830 (0.051) |
| $b$ | −0.0010 | −0.0013 (0.0001) | −0.0013 | −0.0015 (0.0002) | −0.0015 | −0.0015 (0.00012) |
| $\sigma$ | 0.193 | 0.136 (0.035) | 0.158 | 0.145 (0.034) | 0.167 | 0.121 (0.025) |
| *Paramecium caudatum* | | | | | | |
| $a$ | 0.585 | 0.607 (0.0534) | 0.450 | 0.579 (0.066) | 0.576 | 0.562 (0.068) |
| $b$ | −0.0030 | −0.0030 (0.0004) | −0.0021 | −0.0027 (0.0004) | −0.0029 | −0.0030 (0.0005) |
| $\sigma$ | 0.283 | 0.131 (0.0411) | 0.406 | 0.171 (0.0475) | 0.346 | 0.172 (0.044) |

| | *P. aurelia* | | *P. caudatum* | |
|---|---|---|---|---|
| Combined replicates | 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) |
| $\sigma$ | 0.174 | 0.139 (0.031) | 0.339 | 0.162 (0.044) |

Models were fitted to data on two species of *Paramecium* (Appendix I, Table 3 of Gause 1934) using data cloning. Data are plotted in Fig. 2. Priors on the parameters were normal(1,1), normal(−1,1) and uniform(0,1), respectively.

frequentist inference for hierarchical models now boils down to whether or not prior distributions are relevant for the scientific inferences. The main difference between the Bayesian and the likelihood-based inferences resulting from data cloning is that the Bayesian inferences depend on the specification and choice of the prior distribution, whereas the likelihood-based inferences are completely invariant to the choice of the prior distributions.

Of course, if the data fundamentally do not contain information about the parameters in question, the likelihood as well as the Bayesian inferences could be ill-behaved. For instance, it is all too easy to build complex statistical models with non-identifiable or nearly non-identifiable parameters, i.e. models in which a wide range of parameter values and combinations could give rise to the same data with equal probability. A simple example of non-identifiable parameters is when independent, identically distributed observations are drawn from a normal distribution with a mean of $\mu$ and variance of $\sigma^2 + \tau^2$: such data cannot inform about the separate values of $\sigma^2$ and $\tau^2$, but only about their sum. The data cloning technique will not remedy over- or ill-parameterized models. Incidentally, the Bayesian approach, by assigning prior distributions, assumes the parameters are identifiable, and thereby risks providing 'answers' when none are justified by the data. Such results could be misleading for scientific inferences and resultant policy decisions. The problem of how to assess identifiability in complex models remains a difficult challenge for the Bayesian and frequentist approaches alike.

The data cloning algorithm bears some similarity to the simulated annealing algorithm for optimization (Kirkpatrick *et al.* 1983; Brooks & Morgan 1995; Geyer & Thompson 1995; Brooks *et al.* 2003). Applying simulated annealing to hierarchical models, however, requires evaluating or calculating a likelihood; data cloning avoids the need for such evaluation. A simulated annealing algorithm applied to ML estimation uses an MCMC method to generate observations from an equilibrium distribution with pdf $b_T(\theta)$ proportional to $[L(\theta;\underline{y})]^{1/T}$, where $L(\theta;\underline{y})$ is a likelihood function (Brooks *et al.* 2003). In simulated annealing the 'temperature' $T$ is slowly reduced until $b_T(\theta)$ becomes concentrated at the global maximum of $L(\theta;\underline{y})$. Although $1/T$ in simulated annealing has ostensibly the same role as the number of clones $k$ in data cloning, the simulated annealing algorithm does not accommodate latent variables or random effects, nor does it automatically provide standard errors.

Data cloning is closely related to the 'prior feedback' method developed by Robert (1993) for ML estimation in a particular class of statistical distributions for which the likelihood could be evaluated. Like data cloning, prior feedback combines simulated annealing and Bayesian MCMC simulations, with the likelihood raised to a power to concentrate the posterior around the ML estimates. Robert & Titterington (1998) adapted the prior feedback method for estimation in hidden Markov models, which are a special type of latent variable model. Data cloning as we have described here extends ML estimation, along with the calculation of standard errors, to a wide variety of hierarchical models in general.

The approximate standard errors obtained for the ML estimates under data cloning are those arising from the Fisher information matrix under large-sample ML theory. The standard errors can be used for approximate confidence intervals based on the asymptotic normal distribution of the ML estimates. The standard errors and confidence intervals should be treated with the same cautions as with any conclusions based on asymptotic ML theory. Improved confidence intervals could presumably be obtained with bootstrapping (for instance, Manly 2006). Although the required refitting of the model via data cloning to thousands of simulated data sets is technically feasible, current software packages for Bayesian MCMC calculations do not yet contain such 'looping' facility. Additional research is needed towards improving the estimates of the variance/covariance structure of ML estimates for hierarchical models.

The data cloning method at present does not provide the subsidiary information necessary for model selection with the Akaike Information Criterion (AIC; see Burnham & Anderson 2002) and its relatives. The method yields only the location of the ML estimate but not the actual value of the likelihood function at its maximum. Additional calculations, of the types cited earlier for simulating likelihoods, are required to obtain the value of the likelihood function. One technique, direct Monte Carlo integration, is to repeatedly simulate values of $\underline{X}$ from $g(\underline{X}|\hat{\theta})$, until the mean of the values $f(\underline{y}|\underline{X}^{(1)}, \hat{\varphi}), f(\underline{y}|\underline{X}^{(2)}, \hat{\varphi}), \ldots$ stabilizes. We caution that the problem of model selection for hierarchical models is not well understood. Some modification to the AIC is needed because it is unclear how the likelihood penalty term should reflect the number of unobserved state variables or the number of random effects.

Bayesian methods for fitting complex ecological models have often been advocated because the methods can provide reasonable solutions to difficult problems, problems in which ML estimation has previously been impracticable. We believe the data cloning algorithm partly removes the relevance of such a justification. Data cloning is a potentially important and useful tool for those scientists who prefer to use the frequentist approach for conducting statistical inferences with hierarchical models.

## ACKNOWLEDGEMENTS

The authors thank editor Marti Anderson and four anonymous referees for numerous helpful suggestions for improving the paper. Work by Subhash Lele and Frithjof Lutscher was supported by a grant from the National Science and Engineering Research Council of Canada. Work by Brian Dennis was supported in part by National Science Foundation grant DEB 0315289 to D. R. Strong and A. Hastings and by Montana Fish, Wildlife, and Parks contract #060327 to M. L. Taper. This work originated during the discussions at the 'Graduate Short Course on Analysis of Ecological Systems Under Uncertainty' which was conducted at the Kananaskis Field Station of the University of Calgary and was supported by a grant from Pacific Institute for the Mathematical Sciences to Frithjof Lutscher and Ed McCauley.

## REFERENCES

Barnett, V. (1999). *Comparative Statistical Inference*. Wiley, New York.

Basu, S. & Ebrahimi, N. (2001). Bayesian capture-recapture methods for error detection and estimation of population size: heterogeneity and dependence. *Biometrika*, 88, 269–279.

Bennett, C.H. (1976). Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.*, 22, 245–268.

Breslow, N.E. & Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, 82, 81–91.

Breslow, N.E. & Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.*, 88, 9–25.

Brooks, S.P. & Morgan, B.J.T. (1995). Optimization using simulated annealing. *Statistician*, 44, 241–257.

Brooks, S.P., Catchpole, E.A. & Morgan, B.J.T. (2000). Bayesian animal survival estimation. *Statist. Sci.*, 15, 357–376.

Brooks, S.P., Friel, N. & King, R. (2003). Classical model selection via simulated annealing. *J. Roy. Stat. Soc. B*, 65, 505–520.

Buckland, S.T., Newman, K.B., Thomas, L. & Koesters, N.B. (2004). State-space models for the dynamics of wild animal populations. *Ecol. Model.*, 171, 157–175.

Burnham, K.P. & Anderson, D.R. (2002). *Model Selection and Multimodel Inference: a Practical Information-Theoretic Approach*, 2nd edn. Springer, New York.

Carlin, B.P. & Louis, T.A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd edn. Chapman and Hall/CRC, Boca Raton.

Casella, G. & George, E.I. (1992). Explaining the Gibbs sampler. *Am. Stat.*, 46, 167–174.

Chib, S. & Greenberg, E. (1995). Understanding the Metropolis–Hastings algorithm. *Am. Stat.*, 49, 327–335.

Clark, J.S. (2003). Uncertainty in population growth rates calculated from demography: the hierarchical approach. *Ecology*, 84, 1370–1381.

Clark, J.S. (2005). Why environmental scientists are becoming Bayesians. *Ecol. Lett.*, 8, 2–14.

Clark, J.S. & Bjørnstad, O.N. (2004). Population time series: process variability, observation errors, missing values, lags, and hidden states. *Ecology*, 85, 3140–3150.

Clark, J.S. & Gelfand, A.E. (2006). A future for models and data in ecology. *Trends Ecol. Evol.*, 21, 375–380.

Clark, J.S., Lewis, M., McLachlan, J.S. & Hille Ris Lambers J. (2003). Estimating population spread: what can we forecast and how well? *Ecology*, 84, 1979–1988.

Clark, J.S., Ferraz, G., Oguge, N., Hays, H. & DiCostanzo, J. (2005). Hierarchical Bayes for structured and variable populations: from capture-recapture data to life-history prediction. *Ecology*, 86, 2232–2244.

De Valpine, P. (2003). Better inferences from population-dynamics experiments using Monte Carlo state-space likelihood methods. *Ecology*, 84, 3064–3077.

De Valpine, P. (2004). Monte Carlo state-space likelihoods by weighted posterior kernel density estimation. *J. Am. Stat. Assoc.*, 99, 523–536.

De Valpine, P. & Hastings, A. (2002). Fitting population models incorporating process noise and observation error. *Ecol. Monogr.*, 72, 57–76.

Dennis, B. & Taper, M.L. (1994). Density dependence in time series observations of natural populations: estimation and testing. *Ecol. Monogr.*, 64, 205–224.

Dennis, B., Desharnais, R.A., Cushing, J.M. & Costantino, R.F. (1995). Nonlinear demographic dynamics: mathematical models, statistical methods, and biological experiments. *Ecol. Monogr.*, 65, 261–281.

Dennis, B., Ponciano, J.M., Lele, S.R. & Taper, M.L. (2006). Estimating density dependence, process noise, and observation error. *Ecol. Monogr.*, 76, 323–341.

Etienne, R.S. & Olff, H. (2005). Confronting different models of community structure to species abundance data: a Bayesian model comparison. *Ecol. Lett.*, 8, 493–504.

Feinberg, S.E., Johnson, M.S. & Junker, B.W. (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *J. Roy. Stat. Soc. A*, 62, 383–406.

Gause, G.F. (1934). *The Struggle for Existence*. Williams & Wilkins, Baltimore.

Gelfand, A.E. & Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.*, 85, 398–409.

Gelfand, A.E., Schmidt, A.M., Wu, S., Silander, J.A. Jr, Latimer, A. & Rebelo, A.G. (2005). Explaining species diversity through species level hierarchical modeling. *Appl. Stat.*, 54, 1–20.

Gelman, A. & Meng, X.L. (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statist. Sci.*, 13, 163–185.

George, E.I. & Robert, C.P. (1992). Capture-recapture estimation via Gibbs sampling. *Biometrika*, 79, 677–683.

George, A.W. & Thompson, E.A. (2003). Discovering disease genes: multipoint linkage analysis via a new Markov chain Monte Carlo approach. *Statist. Sci.*, 18, 515–531.

Geyer, C.J. & Thompson, E.A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. Roy. Stat. Soc. B*, 54, 657–699.

Geyer, C.J. & Thompson, E.A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Am. Stat. Assoc.*, 90, 909–920.

Gilks, W.R., Richardson, S. & Spiegelhalter, D.J. (eds) (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

Green, E.J., MacFarlane, H.T., Valentine, H.T. & Strawderman, W.E. (1999). Assessing uncertainty in a stand growth model by Bayesian synthesis. *For. Sci.*, 45, 528–538.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97–109.

Heagerty, P. & Lele, S.R. (1998). A composite likelihood approach to binary data in space. *J. Am. Stat. Assoc.*, 93, 1099–1111.

Hilborn, R. & Mangel, M. (1997). *The Ecological Detective: Confronting Models With Data*. Princeton University Press, Princeton, New Jersey.

Karim, M.R. & Zeger, S.L. (1992). Generalized linear models with random effects: salamander mating revisited. *Biometrics*, 48, 631–644.

Kirkpatrick, S., Gelatt, C.D. & Vecchi, M.P. (1983). Optimisation using simulated annealing. *Science*, 220, 671–680.

Kitagawa, G. (1987). Non-Gaussian state-space modeling of non-stationary time series (with discussion). *J. Am. Stat. Assoc.*, 82, 1032–1063.

Lele, S.R. (2006). Sampling variability and estimates of density dependence: a composite-likelihood approach. *Ecology*, 87, 189–202.

Link, W.A. & Sauer, J.R. (2002). A hierarchical model for population change with application to Cerulean Warblers. *Ecology*, 83, 2832–2840.

Link, W.A., Cam, E., Nichols, J.D. & Cooch, E.G. (2002). Of bugs and birds: Markov chain Monte Carlo for hierarchical modeling in wildlife research. *J. Wildlife Manage.*, 66, 277–291.

Link, W.A., Sauer, J.R. & Niven, D.K. (2006). A hierarchical model for regional analysis of population change using Christmas Bird Count data, with application to the American Black Duck. *Condor*, 108, 13–24.

Manly, B.F.J. (2006). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 3rd edn. Chapman & Hall/CRC, New York.

May, R.M. & Oster, G.F. (1976). Bifurcations and dynamic complexity in simple ecological models. *Am. Nat.*, 110, 573–599.

McCulloch, C.E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Am. Stat. Assoc.*, 92, 162–170.

McCulloch, C.E. & Searle, S.R. (2001). *Generalized, Linear and Mixed Models*. Wiley, New York.

Meyer, R. & Millar, R.B. (1999a). Bayesian stock assessment using a state-space implementation of the delay difference model. *Can. J. Fish. Aquat. Sci.*, 56, 37–52.

Meyer, R. & Millar, R.B. (1999b). BUGS in Bayesian stock assessments. *Can. J. Fish. Aquat. Sci.*, 56, 1078–1087.

Millar, R.B. & Meyer, R. (2000a). Non-linear state space modeling of fisheries biomass dynamics using the Metropolis-Hastings within-Gibbs sampling. *Appl. Stat.*, 49, 327–342.

Millar, R.B. & Meyer, R. (2000b). Bayesian state-space modeling of age-structured data: fitting a model is just the beginning. *Can. J. Fish. Aquat. Sci.*, 57, 43–50.

Newman, K.B., Buckland, S.T., Lindley, S.T., Thomas, L. & Fernández, C. (2006). Hidden process models for animal population dynamics. *Ecol. Appl.*, 16, 74–86.

Pascual, M.A. & Kareiva, P. (1996). Predicting the outcome of competition using experimental data: maximum likelihood and Bayesian approaches. *Ecology*, 77, 337–349.

Ponciano, J.M., De Gelder, L. & Top, E.M. (2007). The population biology of bacterial plasmids: a hidden Markov model approach. *Genetics* (in press).

Press, S.J. (2003). *Subjective and Objective Bayesian Statistics*. Wiley, New York.

R Core Development Team (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Radtke, P.J., Burk, T.E. & Bolstad, P.V. (2002). Bayesian melding of a forest ecosystem model with correlated inputs. *For. Sci.*, 48, 701–711.

Rivot, E. & Prévost, E. (2002). Hierarchical Bayesian analysis of capture-mark-capture data. *Can. J. Fish. Aquat. Sci.*, 59, 1768–1784.

Robert, C.P. (1993). Prior feedback: Bayesian tools for maximum likelihood estimation. *J. Comput. Stat.*, 8, 279–294.

Robert, C.P. & Casella, G. (2004). *Monte Carlo Statistical Methods*, 2nd edn. Springer, New York.

Robert, C.P. & Titterington, D.M. (1998). Reparameterization strategies for hidden Markov models and Bayesian approches to maximum likelihood estimation. *Stat. Comput.*, 8, 145–158.

Sauer, J.R. & Link, W.A. (2002). Hierarchical modeling of population stability and species group attributes using Markov chain Monte Carlo methods. *Ecology*, 83, 1743–1751.

Skaug, H.G. & Fournier, D.A. (2006). Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. *Comput. Stat. Data An.*, 51, 699–709.

Spall, J.C. (2003). *Introduction to Stochastic Search and Optimization*. Wiley, New York.

Spiegelhalter, D., Thomas, A. & Best, N. (2004). *WinBUGS Version 1.4 User Manual. MRC Biostatistics Unit*. Institute of Public Health, London.

Stuart, A. & Ord, J.K. (1987). *Kendall's Advanced Theory of Statistics. Vol. 2. Classical Inference and Relationship*, 5th edn. Oxford University Press, Oxford.

Sturtz, S., Ligges, U. & Gelman, A. (2005). *R2WinBUGS: a Package for Running WinBUGS from R. J. Statist. Software, 12.* Available at: http://www.jstatsoft.org/ and http://www.cran.r-project.org/, last accessed in April 2007.

Thogmartin, W.E., Sauer, J.R. & Knutson, M.G. (2004). A hierarchical spatial model of avian abundance with application to Cerulean Warblers. *Ecol. Appl.*, 14, 1766–1779.

Walker, A.M. (1969). On the asymptotic behaviour of posterior distributions. *J. Roy. Stat. Soc. B*, 31, 80–88.

Wikle, C.K. (2003). Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology*, 84, 1382–1394.

Wikle, C.K., Berliner, L.M. & Cressie, N. (1998). Hierarchical Bayesian space–time models. *Environ. Ecol. Stat.*, 5, 117–154.

## APPENDIX

In this appendix, we prove that the estimates obtained with the data cloning algorithm (i.e. as the mean values of the data-cloned posterior distribution of the parameters) converge to the ML estimates as the number of clones *k* increases. The derivation interprets Bayes' rule as an iterated map on the space of probability distributions. First, for simplicity of description we suppress the latent variables and write the proof instead using a somewhat simpler model form. Second, we provide the proof that explicitly includes latent variables in the model. Finally, we show that the Metropolis-Hastings algorithm generates random variables from the data-cloned posterior distribution of the parameters.

A complete proof of the convergence of the scaled variances and covariances of the data-cloned posterior distribution to the inverse of the Fisher information is provided by S. R. Lele (unpublished work).

## Fixed effects case

Let $\underline{y} \sim f(\underline{y}\,|\,\theta)$ be the statistical model for the data. Let $\Theta$ denote the parameter space, the set of values the parameter $\theta$ can possibly take. Let $\pi(\theta)$ be the prior distribution on the parameter space $\Theta$. We assume that the prior distribution is positive over the entire parameter space $\Theta$, i.e. it does not preclude *a priori* any values of the parameter space $\Theta$. Then, the posterior distribution is given by

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

Now suppose we substitute this distribution as prior back again then we obtain

$$\pi^{(2)}(\theta|\underline{y}) = \frac{f(\underline{y}|\theta)\pi(\theta|\underline{y})}{\int f(\underline{y}|\theta)\pi(\theta|\underline{y})d\theta} = \frac{[f(\underline{y}|\theta)]^2\pi(\theta)}{\int [f(\underline{y}|\theta)]^2\pi(\theta)d\theta}.$$

By induction, it follows that

$$\pi^{(k)}(\theta|\underline{y}) = \frac{[f(\underline{y}|\theta)]^k\pi(\theta)}{\int [f(\underline{y}|\theta)]^k\pi(\theta)d\theta}.$$

Note that $\pi^{(k)}(\theta\,|\underline{y})$ is the posterior distribution resulting from a cloned data likelihood, $[f\,(\underline{y}\,|\,\theta)]^k$. The posterior can therefore be looked upon as an iterated map, $\pi^{(1)} = F(\pi, f)$, $\pi^{(2)} = F[\pi^{(1)}, f]$, …, $\pi^{(k)} = F[\pi^{(k-1)}, f]$. We can now study if this iterated map has a fixed point (a distribution, in this case) and if it is independent of the initial distribution. To establish the existence of such a fixed point, recall that

$$\pi^{(k)}(\theta|\underline{y}) = \frac{[f(\underline{y}|\theta)]^k\pi(\theta)}{\int [f(\underline{y}|\theta)]^k\pi(\theta)d\theta}.$$

Let $\hat{\theta}$ be a point in $\Theta$ such that $f(y|\hat{\theta}) > f(y|\theta)$ for all $\theta \in \Theta$. By definition, this is the MLE of $\theta$. As $\pi$ is positive everywhere on the parameter space $\Theta$, it follows that

$$\frac{\pi^{(k)}(\theta|\underline{y})}{\pi^{(k)}(\hat{\theta}|\underline{y})} = \frac{[f(\underline{y}|\theta)]^k\pi(\theta)}{[f(\underline{y}|\hat{\theta})]^k\pi(\hat{\theta})} \to 0 \qquad \text{if } \theta \neq \hat{\theta}$$

and

$$\frac{\pi^{(k)}(\theta|\underline{y})}{\pi^{(k)}(\hat{\theta}|\underline{y})} = \frac{[f(\underline{y}|\theta)]^k\pi(\theta)}{[f(\underline{y}|\hat{\theta})]^k\pi(\hat{\theta})} \to 1 \qquad \text{if } \theta = \hat{\theta}.$$

In other words, the fixed point for the iterated map is a degenerate distribution, degenerate at $\hat{\theta}$. The degenerate distribution is also independent of the initial distribution $\pi$. Because the mean of a degenerate distribution is the point at which it is degenerate, the mean of the posterior distribution for large enough $k$ approaches the MLE of $\theta$.

## Latent variables case

Let us now explicitly include latent variables in the model. Following the notation in the paper, suppose observations $\underline{Y} = (Y_1, Y_2, \ldots, Y_n)$ arise from the following hierarchical statistical model:

$$\underline{Y} \sim f(\underline{y}|\underline{X}, \varphi),$$

$$\underline{X} \sim g(\underline{x}|\theta),$$

where $f$ and $g$ are joint pdfs, $\underline{X}$ is a vector of random quantities or processes affecting the observations, $\varphi = (\varphi_1, \varphi_2, \ldots, \varphi_q)$ is a vector of unknown fixed parameters affecting the observations, and $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$ is a vector of unknown fixed parameters related to the process $\underline{X}$. Let $\pi(\theta, \varphi)$ be the prior distribution on the parameter space $\Theta$. We assume that the prior distribution is positive over the entire parameter space $\Theta$, i.e. it does not preclude *a priori* any values of the parameter space $\Theta$. The posterior distribution of $(\theta, \varphi)$ is given by

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

where $h(\underline{y}) = \int f(\underline{y}\,|\,\underline{X}, \varphi)g(\underline{X}\,|\,\theta)\pi(\theta, \varphi)d\underline{X}d\theta d\varphi$. As before, we now substitute this posterior distribution back again as the prior distribution to obtain

$$\pi^{(2)}(\theta, \varphi|\underline{y}) = \frac{\left\{\int f(\underline{y}|\underline{X}, \varphi)g(\underline{X}|\theta)d\underline{X}\right\}\pi^{(1)}(\theta, \varphi)}{h^{(2)}(\underline{y})}$$

$$= \frac{\left\{\int f(\underline{y}|\underline{X}, \varphi)g(\underline{X}|\theta)d\underline{X}\right\}^2\pi(\theta, \varphi)}{h^{(2)}(\underline{y})}$$

$$= \frac{\left\{L(\theta, \varphi; \underline{y})\right\}^2\pi(\theta, \varphi)}{h^{(2)}(\underline{y})}$$

Continuing in this fashion, we obtain

$$\pi^{(k)}(\theta, \varphi|\underline{y}) = \frac{\left\{L(\theta, \varphi; \underline{y})\right\}^k\pi(\theta, \varphi)}{h^{(k)}(\underline{y})}.$$

We point out that adding more 'layers' of random effects in the original hierarchical model still produces a posterior that is proportional to the $k$th power of the (integrated) likelihood. Let $(\hat{\theta}, \hat{\varphi})$ be such that $L(\hat{\theta}, \hat{\varphi}; \underline{y}) > L(\theta, \varphi; \underline{y})$ for all $(\theta, \varphi)$. By definition, this is the MLE of $(\theta, \varphi)$. As $\pi(\theta, \varphi)$ is positive everywhere on the parameter space, it follows that

$$\frac{\pi^{(k)}(\theta,\varphi|\underline{y})}{\pi^{(k)}(\hat{\theta},\hat{\varphi}\,|\underline{y})}=\frac{[L(\theta,\varphi;\underline{y})]^{k}}{[L(\hat{\theta},\hat{\varphi};\underline{y})]^{k}}\to 0 \qquad \text{if } (\theta,\varphi)\neq(\hat{\theta},\hat{\varphi})$$

and

$$\frac{\pi^{(k)}(\theta,\varphi|\underline{y})}{\pi^{(k)}(\hat{\theta},\hat{\varphi}\,|\underline{y})}=\frac{[L(\theta,\varphi;\underline{y})]^{k}}{[L(\hat{\theta},\hat{\varphi};\underline{y})]^{k}}\to 1 \qquad \text{if } (\theta,\varphi)=(\hat{\theta},\hat{\varphi}).$$

In other words, the fixed point for the iterated map is a degenerate distribution, degenerate at $(\hat{\theta},\hat{\varphi})$. The degenerate distribution is also independent of the initial distribution $\pi$. Because the mean of a degenerate distribution is the point at which it is degenerate, the mean of the posterior distribution for large enough $k$ approaches the MLE of $(\theta, \varphi)$.

### Metropolis-Hastings algorithm

Further, we show that the Metropolis-Hastings algorithm described below generates random variables from the posterior distribution $\pi^{(k)}(\theta, \varphi\,|\,\underline{y})$.

(a) Generate $(\theta^*,\varphi^*)$ from $\pi(\theta, \varphi)$.

(b) Generate $k$ values of $\underline{X}$, say $\underline{X}^{*(1)}, \underline{X}^{*(2)}, \ldots, \underline{X}^{*(k)}$, from $g(\underline{X}\,|\,\theta^*)$.

(c) Calculate the product

$$q* = f\left(\underline{y}|\underline{X}^{*(1)},\varphi*\right) f\left(\underline{y}|\underline{X}^{*(2)},\varphi*\right)\ \ldots f\left(\underline{y}|\underline{X}^{*(k)},\varphi*\right).$$

(d) Repeat (a) and (b), obtaining new values $(\theta^{\#}, \varphi^{\#}, X^{\#(j)}, j = 1, 2, \ldots, k)$ and $q^{\#}$.

(e) Generate a uniform(0,1) random variable $U$, and calculate $p = \min[1,(q^{\#}/q^*)]$. If $U > p$, set $\theta,\varphi,\underline{X}^{(j)}, j = 1, 2, .., k_l = [\theta^*, \varphi^*, \underline{X}^*(j), j = 1, 2, .., k]$; otherwise set $[\theta, \varphi, X^{(j)}, j{=}1, 2, \ldots, k]_l = [\theta^{\#}, \varphi^{\#}, X^{\#(j)}, j =1, 2, \ldots, k]$.

(f) Repeat (d) and (e), many times.

Hastings (1970) proves that this algorithm defines a Markov chain with stationary distribution $\pi^{(k)}[\theta, \varphi, \underline{X}^{(j)}, j = 1,2, \ldots, k\,|\,\underline{y}]$ that is proportional to $f(\underline{y}\,|\,\underline{X}^{(1)},\varphi)f(\underline{y}\,|\,\underline{X}^{(2)},\varphi)\ \ldots\ f(\underline{y}\,|\,\underline{X}^{(k)}, \varphi)g(\underline{X}^{(1)}\,|\,\theta)g(\underline{X}^{(2)}\,|\,\theta)\ \ldots\ g(\underline{X}^{(k)}\,|\,\theta)\pi(\theta, \varphi)$. Thus, after sufficient number of steps in the Markov chain, the above algorithm generates random variates, albeit dependent, from this stationary distribution. To obtain random variates from the marginal posterior distribution of $(\theta, \varphi)$, we simply pick the $(\theta, \varphi)$ component of the random variates generated from the Markov chain defined above. This needs no integration over the $\underline{X}$ space.

We now prove that the marginal distribution of $(\theta, \varphi)$ for the above stationary distribution is equal to $\pi^{k}(\theta, \varphi\,|\,\underline{y})$. Consider, with the understanding that the integrals in the equation are multiple integrals (although only a single integral sign is used for notational simplicity), the steps listed below for integrating out the latent variables:

$$\begin{aligned}
&\int \pi^{(k)}(\theta,\varphi,\underline{X}^{(1)},\underline{X}^{(2)},\underline{X}^{(3)},\ldots,\underline{X}^{(k)}|\underline{y})d\underline{X}^{(1)}d\underline{X}^{(2)}\ldots d\underline{X}^{(k)}\\
&=\frac{\left\{\int f(\underline{y}|\underline{X}^{(1)},\varphi)f(\underline{y}|\underline{X}^{(2)},\varphi)\ldots f(\underline{y}|\underline{X}^{(k)},\varphi)g(\underline{X}^{(1)}|\theta)g(\underline{X}^{(2)}|\theta)\ldots g(\underline{X}^{(k)}|\theta)d\underline{X}^{(1)}d\underline{X}^{(2)}\ldots d\underline{X}^{(k)}\right\}\pi(\theta,\varphi)}{h^{(k)}(\underline{y})}\\
&=\frac{\left\{\int f(\underline{y}|\underline{X},\varphi)g(\underline{X}|\theta)d\underline{X}\right\}^{k}\pi(\theta,\varphi)}{h^{(k)}(\underline{y})}\\
&=\frac{\left\{L(\theta,\varphi;\underline{y})\right\}^{k}\pi(\theta,\varphi)}{h^{(k)}(\underline{y})}\\
&=\pi^{(k)}(\theta,\varphi|\underline{y}).
\end{aligned}$$

Editor, Marti Anderson
Manuscript received 8 December 2006
First decision made 11 January 2007
Second decision made 7 March 2007
Manuscript accepted 3 April 2007