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

CMR model can be framed as a state–space model (Gimenez et al. 2007). Although these quantities are unknown, their role in the model is to structure the manner in which data values are non-independent, just like block effects in a randomized complete block analysis of variance (ANOVA) design. In theory, a more realistic model structure will lead to better estimation and inference. Other modeling categories or synonyms that have this general feature include generalized linear mixed models, latent variable models, hidden state or hidden population models, and more. A useful umbrella term for all of these cases, which emphasizes their commonality, is hierarchical models (Royle and Dorazio 2008; Cressie et al. 2009).

It should be noted that in the Lapwing example used below, the abundance data are really an abundance index, so "true abundance" really means "true abundance index". The use of an abundance index raises all of the potential issues of how to relate raw survey data, estimated abundance indices, and actual population sizes (Buckland et al. 2001; Williams et al. 2002), but these issues are not addressed in this article.

The goal of this paper is to review, illustrate, and discuss methods for frequentist analysis of hierarchical models of population dynamics and demographic studies. Of particular interest are so-called *integrated population models*, which combine a model for a time-series of abundance data, such as a state–space model, with models for individual demographic information, such as a model for ring-recovery data (Besbeas et al. 2002). In bird population studies, frequentist analysis of integrated population models has been limited to linear, Gaussian approximations using the Kalman filter (Besbeas et al. 2002; Gauthier et al. 2007; Tavecchia et al. 2009), while Bayesian analysis has allowed nonlinear relationships and/or non-Gaussian variation (Brooks et al. 2004; Schaub et al. 2007; King et al. 2008b). Methods for frequentist analysis allowing nonlinear and/or non-Gaussian state–space models have been developed in other areas of statistics and ecology, but they have not been used in integrated population models. The Lapwing example below provides the first use of such methods for an integrated population model.

Estimating parameter values and their uncertainty from hierarchical models is fundamentally difficult because the fit provided by any candidate parameters is not a simple calculation (Robert and Casella 1999; Durbin and Koopman 2001). For example, the likelihood of a state–space model for some candidate parameters involves a sum (or integral) over all of the possible true population trajectories that might have produced the data. Here, current approaches to this problem are reviewed with an attempt to highlight their pros and cons in order to stimulate their application to ecological problems. While the emphasis is on state–space models and integrated population models,

the methods discussed here can be useful for other hierarchical models.

Many of the methods for maximum-likelihood estimation of hierarchical models do not calculate the full likelihood (de Valpine 2004). Instead, they omit a constant factor known as a normalizing constant. Since it is a constant, it does not affect maximization of the likelihood, but it is needed for values such as the Akaike information criterion (AIC) or likelihood ratios (Harvey 1991; Durbin and Koopman 2001; de Valpine 2008; Ponciano et al. 2009). Therefore, methods for calculating normalizing constants need to be part of the toolkit for frequentist analysis of hierarchical models, and they are also reviewed here. Mathematically, this problem is similar to the problem of calculating Bayes factors or marginal likelihoods in Bayesian analysis, which involve the normalizing constants that are omitted in Markov chain Monte Carlo (MCMC) posterior samplers (Han and Carlin 2001).

To illustrate frequentist analysis of a non-Gaussian state–space model, I use the well-studied British Lapwing data, on which both Kalman filter and Bayesian methods have been demonstrated (Besbeas et al. 2002; Brooks et al. 2004). I present maximum-likelihood estimates and profile likelihood confidence intervals for a model of Brooks et al. (2004), using Monte Carlo kernel likelihood (MCKL) (de Valpine 2004) for estimation and bridge sampling (Mira and Nicholls 2004; de Valpine 2008) for calculating likelihood values for the estimated parameters, all iterated many times to obtain profile likelihood confidence intervals. This combination of methods was first used for an insect host–parasitoid analysis by Karban and de Valpine (2010). The main point of the example is to illustrate the feasibility of these methods rather than to reconsider biological conclusions of these well-studied data.

The following section introduces state–space models, integrated population models, and the Lapwing example. "Bayesian and frequentist estimation" explains the challenges of each estimation approach for these models. "Why a frequentist analysis?" discusses why frequentist results may be useful even in this age of great advances in Bayesian methods due to MCMC. "Methods for maximum-likelihood estimation" reviews major methods, including those for calculation of likelihood values given estimated parameters. "Lapwing results", demonstrates MCKL and bridge sampling, and the "Discussion" summarizes the status of computational methods for frequentist state–space modeling.

State–space models and integrated population models

A state–space model is a combination of two models, one for the population dynamics and one for the data sampling (or "observation" or "measurement"). State–space models

for population dynamics date back to a now-obscure book chapter by Brillinger et al. (1980) that was ahead of its time. Much development has since taken place in fisheries ecology (reviewed in de Valpine 2002), and the framework has been proposed as a general one in a wildlife ecology context (Borchers et al. 2002; Buckland et al. 2004). Introductions and reviews have been written by de Valpine and Hastings (2002), Calder et al. (2003), Clark and Bjornstad (2004), Thomas et al. (2005), and Newman et al. (2009), among others.

To keep the ideas specific, the British Lapwing model will be used. The data consist of annual abundance indices calculated from British Trust for Ornithology survey data from 1965 to 1998 and ring recovery data from individuals marked as chicks from 1963 to 1997. One must assume that the sampling process and other aspects of the relationship between population size and abundance index obtained this way do not themselves have unknown systematic trends. Covariates include year and number of days in each year below freezing. See Catchpole et al. (1999), Besbeas et al. (2002), and Brooks et al. (2004), from which the data were taken, for details.

The Lapwing data are modeled with two stage classes, 1-year-old, N_1, and adults, N_a. Here, N_1 and N_a are used as the abundance index values, and explicit modeling of the survey sampling process and the relationship between abundance indices and population size is not considered. The true (unknown) abundance index at time t is defined as vector $X(t) = (N_1(t), N_a(t))$. This is called the "state" variable at time t. Next, the data at time t are defined as a survey-based estimate of $N_a(t)$, labeled as $y(t)$. The two parts of the state–space model correspond to two aspects of how any two pairs of abundances, $X(t)$ and $X(t + 1)$, respectively, should be related. First, the $N_1(t + 1)$ and $N_a(t + 1)$ should not be too far off from what would be predicted based on $N_1(t)$ and $N_a(t)$, i.e., by the population dynamics. Second, $N_a(t)$ and $N_a(t + 1)$ should not be too far off from $y(t)$ and $y(t + 1)$, respectively, according to the data sampling distribution.

To write models for the population dynamics and data sampling, it is helpful to think of each relationship in terms of the average, or *expected value*, and the distribution around that average. The expected value of the population state given its previous state is:

$$E[N_1(t + 1)|N_1(t), N_a(t)] = \rho(t)\phi_1(t)N_a(t) \qquad (1)$$

$$E[N_a(t + 1)|N_1(t), N_a(t)] = \phi_a(t)(N_1(t) + N_a(t)) \qquad (2)$$

Equation (1) says recruits, $N_1(t + 1)$, are the product of adults, $N_a(t)$, fecundity, $\rho(t)$, and first-year survival, $\phi_1(t)$. Equation (2) says adults are the product of adult survival, $\phi_a(t)$, and the sum of surviving first-years, $N_1(t)$, and previous adults, $N_a(t)$.

Variation in the relationship between $X(t)$ and $X(t + 1)$ is represented by allowing $X(t)$ to follow a distribution around its expected value. Brooks et al. (2004) consider two options. First, following Besbeas et al. (2002), they use additive normal (Gaussian) noise:

$$N_1(t + 1)|X(t) \sim N(E[N_1(t + 1)|N_1(t), N_a(t)], \sigma_1^2(t)), \qquad (3)$$

$$N_a(t + 1)|X(t) \sim N(E[N_a(t + 1)|N_1(t), N_a(t)], \sigma_a^2(t)). \qquad (4)$$

These equations say that $N_1(t + 1)$ and $N_a(t + 1)$ are normally distributed around their expected values with variances $\sigma_1^2(t)$ and $\sigma_a^2(t)$, respectively. In some situations one chooses to estimate the variances, although Knape (2008) and de Valpine and Hilborn (2005) have illustrated how imprecise such estimates can be, and Dennis et al. (2010) have shown the potential benefits of replicated sampling. Instead, Besbeas et al. (2002) and Brooks et al. (2004) assumed the variances to match those that would arise from demographic stochasticity. For births, they assume stochasticity is a Poisson process, so $\sigma_1^2(t) = \rho_1(t)\phi_1(t)N_1(t)$ because the variance is equal to the mean for a Poisson distribution. For adult survival, they assume death is a binomial process, so $\sigma_a^2(t) = \phi_a(t)(1 - \phi_a(t))(N_1(t) + N_a(t))$, the variance for a binomial distribution. Aspects of these assumptions will be discussed more below.

Even for this simple model, which might appear linear and Gaussian (normal) at first glance, simple considerations render it nonlinear and/or non-Gaussian. First, even as specified, it will lead to non-Gaussian distributions of population states. The importance of Gaussian distributions is that they can be handled analytically—as part of the Kalman filter summarized below—while non-Gaussian distributions require more heavily computational methods. To see how non-Gaussian distributions arise from this model, notice that even if the distribution of unknown states at one time is Gaussian, the future distribution will be non-Gaussian because the variance depends on the true state. Therefore, to approximate the distribution as Gaussian one must use a fixed state value to calculate the variances (Besbeas et al. 2002). Second, in many settings it is realistic to add environmental stochasticity as well as the demographic stochasticity represented by Poisson births and binomial survival. Environmental stochasticity is usually modeled by multiplying by a log-normal random variable, i.e., adding a normal random variable on a log scale. However, if a sum such as in Eq. 2 is involved, then one faces the difficulty that a sum of log-normally distributed variables is not itself log-normally distributed, and indeed it is not analytically tractable.

The second option of Brooks et al. (2004) is to model the variation in $N_1(t + 1)$ and $N_a(t + 1)$ explicitly by Poisson and binomial distributions. With the approaches

here, this could also be done, but it is not included in the example.

One can generalize from the Lapwing example to the general concept of stochastic state dynamics. This is represented by

$$X(t + 1) = F(X(t), v(t)). \tag{5}$$

In this equation, $v(t)$ is any environmental or demographic stochasticity, and F is a general function relating population state at one time and stochasticity to population state at the next time. In other words, the frameworks considered here are very flexible due to the computational methods used, similar to the flexibility for Bayesian models afforded by MCMC.

The observation or sampling model, which relates the data, $y(t)$, to the states, $X(t)$, via some sampling distribution, can now be considered. For the Lapwing model,

$$E[y(t)] = N_a(t), \tag{6}$$

$$y(t) \sim N(E[y(t)], \sigma_y^2). \tag{7}$$

This says that the expected data value is the true index of adult abundance, and the distribution of data values is normal with standard deviation σ_y. Again, the procedures reviewed here could accommodate many reasonable sampling distributions.

Further details on the Lapwing model

The previous part of this section used the Lapwing model for a general introduction to state–space models. I now summarize further details of the model that will be needed to use it for an example later on.

Adult and 1-year-old survival follow separate logistic models as a function of annual number of days below freezing, $freeze(t)$, for year t:

$$\log\left(\frac{\phi_1(t)}{1 - \phi_1(t)}\right) = \alpha_1 + \beta_1 freeze(t), \tag{8}$$

$$\log\left(\frac{\phi_a(t)}{1 - \phi_a(t)}\right) = \alpha_a + \beta_a freeze(t). \tag{9}$$

Reproductive rate allows a trend with time on a log scale:

$$\log(\rho(t)) = \alpha_\rho + \beta_\rho t. \tag{10}$$

Probability of recovery of a dead, ringed bird in year t is $\lambda(t)$, which can take a trend with time on a logistic scale:

$$\log\left(\frac{\lambda(t)}{1 - \lambda(t)}\right) = \alpha_\lambda + \beta_\lambda t \tag{11}$$

Probabilities of the recovery data are calculated from multinomial distributions for the number of birds that are ringed in a given year and recovered in each subsequent year. These are standard calculations and are given

explicitly in Brooks et al. (2004). It should be noted that these add substantial calculations to the MCMC sampling for this model, and Besbeas et al. (2003) have pointed out that a Gaussian approximation of the ring-recovery likelihood would be a quite accurate proxy for the full calculations. On a separate issue, the index values y_t are the result of estimation from a complex spatial sampling protocol, and Besbeas and Freeman (2006) show how the full data and model for that protocol could be combined with a population model.

Bayesian and frequentist estimation

The two fundamental philosophies to statistical learning from data are Bayesian and frequentist. Bayesian analysis views parameters themselves as following probability distributions, where "probability" means "degree-of-belief" (O'Hagan 1994). Frequentist analysis views "probability" as the frequency of a random event among many possible realizations of the random process, and it does not view parameters as having occurred with some probability. In frequentist analysis, parameters are often estimated by maximum likelihood, and their uncertainty is typically characterized by confidence intervals, estimated, for example, by likelihood profiles, Wald approximations, or bootstrapping (Davison and Hinkley 1997; Severini 2001). Frequentist methods are sometimes called "classical" (e.g., Gelman et al. 2004). In this section I introduce the likelihood integral for state–space models and the reasons that Bayesian analysis has been so practical for these models.

The likelihood function says how well any candidate parameters fit the data, and it underlies both frequentist and Bayesian analysis. It is defined as the probability (or probability density) that the model, as a function of the parameters, would have generated the data. For discrete data values, one may talk of the "probability" of the data, while for continuous data values, one should say "probability density" of the data. In what follows I use simply "probability" for either.

There are two equivalent ways the likelihood is written in mathematical notation. Once the data are collected, we evaluate the likelihood for different parameters given fixed data, so the function is written as $L(\text{parameters} \mid \text{data})$; this is the notation typical of describing maximum-likelihood methods. However, since the likelihood is calculated as the probability of the data given the parameters, it is also written as $P(\text{data} \mid \text{parameters})$; this is the notation typically used in the numerator of Bayes' law for Bayesian analysis. These equivalent notations are two sides of the same coin.

What does the probability of the data mean for a state–space model that is defined using population states that we

do not know exactly? First, it will be convenient to write the vector of all observations from time 1 to the final time, T, as $Y_{1:T} = (Y_1, \ldots, Y_T)$. Correspondingly, the vector of all state values is $X_{1:T} = (X_1, \ldots, X_T)$. In the Lapwing model, each X_t is itself a vector containing $N_1(t)$ and $N_a(t)$. The state–space model likelihood can now be written:

$$L(\theta|Y_{1:T}) = \int P(Y_{1:T}|X_{1:T}, \theta)P(X_{1:T}|\theta)dX_{1:T}. \tag{12}$$

The integral in this equation is a summation over all population trajectories, $X_{1:T}$, of the probability of the trajectory and the probability of the data given the trajectory. According to the rules of probability, this gives the marginal (i.e., total) probability of the data, which is the standard definition of the likelihood. Maximum-likelihood estimation requires finding the parameters, $\hat{\theta}$, that maximize (Eq. 12). The usual asymptotic properties of maximum-likelihood estimates, such as consistency and asymptotic normality, have been extended to state–space likelihoods (Jensen and Petersen 1999; Fuh 2006).

Note that I use "given" notation, "|", even if what follows is not a random variable. For example, in $P(X_{1:T}|\theta)$ in (Eq. 12), θ is a frequentist parameter. This allows the same notation for comparable Bayesian and frequentist equations.

Before getting into the methods for maximum-likelihood estimation, it may be helpful to examine why Bayesian computational methods are so practical for state–space models. In a Bayesian analysis, one treats θ as also following the rules of probability. One must choose a prior distribution, $P(\theta)$, that represents any initial ignorance or knowledge about θ. Then the posterior is given by:

$$P(\theta|Y_{1:T}) = \frac{P(\theta)L(\theta|Y_{1:T})}{P(Y_{1:T})}. \tag{13}$$

This is Bayes' theorem.

Modern Bayesian analysis proceeds generally as follows. The denominator (which is just a number, since the data are given) is very hard to calculate because it requires integration over all possible parameter values, θ. The important point for any inference is the relative support for different parameters, and the relative support can be characterized by the numerator. One often sees this reflected in the expression:

$$P(\theta|Y_{1:T}) \propto P(\theta)L(\theta|Y_{1:T}). \tag{14}$$

In this equation, $\propto$ means proportional to, and allows us to drop the normalizing constant $1/P(Y_{1:T})$.

This is still impractical to calculate because of the likelihood. However, if we write the likelihood integral explicitly and also express the left-hand-side as an integral, we have

$$\int P(\theta, X_{1:T}|Y_{1:T})dX_{1:T} \propto$$
$$\int P(\theta)P(Y_{1:T}|X_{1:T}, \theta)P(X_{1:T}|\theta)dX_{1:T}. \tag{15}$$

The left-hand side can be written this way because we can always view the posterior in θ as the sum over $X_{1:T}$ of the posterior in both θ and $X_{1:T}$. For example, if one wants to know the frequency of students who score 90% on an exam, and one has a table of frequencies of exam scores ("θ") along with other variables ("$X_{1:T}$", e.g., hours of lecture attended, performance in other classes), one can sum the frequencies of all combinations of categories over the other variables to obtain just the exam score frequencies. On the right-hand side, the $P(\theta)$ can be brought inside the integral because it does not involve the summation variables, $X_{1:T}$.

Equation (15) is useful because if we have a sample from $P(\theta, X_{1:T}|Y_{1:T})$, then the θ dimensions of the sample are from $P(\theta|Y_{1:T})$. Visually, a sample can be viewed as a table with one row for each entry and one column for each dimension of θ or $X_{1:T}$. A sample from $P(\theta, X_{1:T}|Y_{1:T})$ can be generated by a MCMC algorithm because, using Bayes' theorem, the integrands of Eq. (15) are proportional:

$$P(\theta, X_{1:T}|Y_{1:T}) \propto P(\theta)P(Y_{1:T}|X_{1:T}, \theta)P(X_{1:T}|\theta). \tag{16}$$

An MCMC algorithm is an omnibus approach to simulating a sample in situations like this, where the target distribution can only be calculated up to an unknown constant, i.e., can only be calculated as a ratio for any two values of $(\theta, X_{1:T})$.

Already initiated readers might find my order of explanation peculiar, and uninitiated readers might benefit from knowing why. A typical Bayesian introduction would be to view both θ and $X_{1:T}$ as random variables, write Bayes' theorem directly as Eq. (16), and be done. I want to emphasize that the Bayesian Monte Carlo approach automatically integrates over $X_{1:T}$, so that the posterior for θ is based on the likelihood (Eq. 12). This is important because the likelihood provides the connection to frequentist maximum-likelihood methods. In theory, as the amount of data increases, frequentist and Bayesian methods will give similar answers because the likelihood will overwhelm the prior.

The handy computational methods for Bayesian analysis, particularly MCMC, have contributed to the vast expansion of its use. Arguments may be made on philosophical grounds, but a great many scientists motivated to use hierarchical models have adopted the Bayesian approach largely because it is practical (de Valpine 2009). Some explanations of hierarchical modeling go so far as to present it as specifically Bayesian, but that not the case. Even when a Bayesian analysis is chosen for philosophical reasons, there can be valuable reasons to complement it

with frequentist results. The stance taken here, then, is one of "expanding the toolkit" rather than arguing that one tool should always be used.

Why frequentist analysis?

For this paper, since Bayesian methods have been more practical for state–space models, it is useful to consider why frequentist methods may nevertheless be useful. Efron (2005) predicted that "statistics is in for a burst of new theory and methodology, and that this burst will feature a combination of Bayesian and frequentist reasoning." It should be noted that "empirical Bayes" methods are frequentist in their treatment of parameters, so the maximum-likelihood methods summarized here can just as well be viewed as empirical Bayes methods. Here are several reasons one might consider a frequentist analysis either on its own or as a complement to Bayesian analysis.

1. Model selection: AIC and related methods [e.g., AICc (AIC corrected for small sample size)] are popular frequentist approaches to model selection. Improvements for AIC for state–space models have been suggested by Cavanaugh and Shumway (1997) and Bengtsson and Cavanaugh (2006).
2. Hypothesis testing: For example, in the Lapwing model, one may want to ask if the effects of year or days below freezing on demographic or reporting rates are statistically significant. Many of the arguments against hypothesis testing have included—in addition to its actual limitations—overuse, trivial use, misinterpretation, or overzealousness, but it is nevertheless fundamentally valuable (Mayo and Spanos 2006).
3. Profile likelihoods: Profile likelihoods are a way to look at parameter uncertainty using only the objective information, i.e., the likelihood. They are superior to simpler methods, such as Wald intervals (although these also may be useful), because they do not assume a perfect Gaussian shape of the likelihood (Severini 2001).
4. No need to specify a Bayesian prior: In many situations, even where Bayesian methods have been accepted and a practical knowledge of the behavior of priors has been accumulated, the priors are at best a scientific nuisance. Much has been written elsewhere about this, and here it is worth noting only that uninformative priors depend on how the model is parameterized (with the exception of Jeffreys priors), and this is arbitrary, so that there is no universal solution for choosing uninformative priors. Frequentist analysis allows the benefits of the hierarchical model structure without these difficulties.

5. As a basis for bootstrapping, cross-validation, or randomization tests: These methods require model estimation for many simulated, resampled, or partial data sets. For example, cross-validation provides a direct estimate of the prediction error distribution of a modeling procedure by fitting the model with different combinations of one or more data values omitted, which can provide a direct basis for model selection and for prediction with uncertainty (Cheng and Tong 1992; Ellner and Fieberg 2003). Karban and de Valpine (2010) used a randomization test for a nonlinear state–space model. Efron (1996, 2005) has highlighted the connection between empirical Bayes and bootstrapping.

Methods for maximum-likelihood estimation

Methods for finding the maximum-likelihood estimates, $\hat{\theta}$, for hierarchical models have received substantial attention in the general statistics literature as well as in some ecological papers (McCulloch 1997; de Valpine 2004; Newman et al. 2009). Since a goal of this review is to encourage the exploration and adaptation of these approaches for ecological models, I will summarize a fair variety of methods. However, since all of these methods can be found in more technical detail elsewhere, I will try to explain only the key concepts and procedures for each method. Before explaining the methods, two aspects of state–space models need further elaboration: filtering and MCMC sampling.

Filtering: sequential factorization of the likelihood

Filtering methods use a factorization of the likelihood into sequential calculations for which analytical or numerical methods can be applied. The sequential factorization is:

$$P(Y_{1:T}|\theta) = P(Y_1|\theta)P(Y_2|Y_1, \theta)P(Y_3|Y_{1:2}, \theta) \\ \times \cdots \times P(Y_T|P_{1:T-1}, \theta). \quad (17)$$

This factorization states that the probability of the entire data sequence is the probability of the first observation multiplied by the probability of the second given the first multiplied by the probability of the third given the first and second, and so on up to the probability of the final observation given the previous $T - 1$ observations.

The practical value of this factorization is that it leads to ways to calculate the likelihood recursively. Each factor can itself be viewed as an integral:

$$P(Y_t|Y_{1:t-1}, \theta) = \int P(Y_t|X_t, \theta)P(X_t|Y_{1:t-1}, \theta)\mathrm{d}X_t. \quad (18)$$

This equation states that the probability of the tth observation given the previous $t - 1$ observations is the sum—over

all possible values of the tth true state—of the probability of that state given the previous $t - 1$ observations multiplied by the probability of the tth observation given that state.

The steps of filtering methods are as follows. In these steps, phrases such as "calculate the distribution" should be taken conceptually; often the distributions are not mathematically simple, and in specific methods the calculations are performed with simulated samples or other approximations.

1. Start with an assumed distribution for the true state of the system at the first time. This can be done in several ways (Besbeas et al. 2009), but the results are often not too sensitive to the details.
2. Define the first time as the "current time," or "t".
3. Calculate the probability of the observation at the current time using the distribution of the state at the current time. This is $P(Y_t|Y_{1:t-1})$ (one number), or just $P(Y_1)$ for the first time.
4. If the current time is the last time, stop.
5. Update the distribution of the state to reflect the information from the observation at the current time. In mathematical terms, obtain the conditional distribution of the state given the observation: $P(X_t|Y_{1:t})$.
6. Use the process model (with stochasticity) to calculate the distribution of the state at the next time from the updated distribution at the current time: $P(X_{t+1}|Y_{1:t})$.
7. Increment the "current time," (t) so that the distribution of the state calculated in the previous step is now the state at the "current time." Mathematically, re-label $P(X_{t+1}|Y_{1:t})$ as $P(X_t|Y_{1:t-1})$ for use in step 3 as the "distribution of the state at the current time" (given all previous data).
8. Go to step 3.

The total likelihood is the product of all of the data probabilities from each iteration of step (3).

The term "filtering" may be opaque. In early applications of state–space models, the model parameters were known, and the goal was to estimate the state of the system by balancing new observations with predictions from previous observations in light of both process noise and observation error, i.e., to update $P(X_t|Y_{1:t})$ to $P(X_{t+1}|Y_{1:t+1})$. In a sense, this represents "filtering" the noises to obtain optimal knowledge about the states and their uncertainty.

MCMC sampling for state–space models

MCMC algorithms are a general approach to generating a sample from a conditional distribution when we can only calculate the full joint distribution. For example, in the Bayesian case, a sample from $P(X_{1:T}, \theta|Y_{1:T})$ can be generated using the ability to calculate $P(Y_{1:T}|X_{1:T}, \theta)$ $P(X_{1:T}|\theta)P(\theta)$.

In the context of state–space models, there are two ways that MCMC algorithms are used. First, some maximum-likelihood algorithms, such as Monte Carlo expectation maximization, need an MCMC sampler of the states given fixed parameters and data, $P(X_{1:T}|Y_{1:T}, \theta)$. This type of algorithm uses an MCMC sample based on one value of θ to find a better value of θ, then runs the MCMC again with that θ and iterates these steps until θ converges to the maximum-likelihood estimate (which can be difficult to ascertain). Although this approach uses an MCMC sampler repeatedly, it can usually be a very efficient sampler because parameters are fixed.

Other maximum-likelihood algorithms, such as data cloning and MCKL, need a full Bayesian sampler from states and parameters given data, $P(X_{1:T}, \theta|Y_{1:T})$. MCMC samplers for parameters and states are typically much less efficient than those for states only. For these algorithms, the Bayesian view of parameters required to include them in posterior sampling can be viewed as a mathematical trick rather than a philosophical shift in the definition of probability.

Methods based on filtering

Kalman Filter and approximations

The Kalman filter calculates (Eq. 17) when all relationships in the models are linear and both process noise and observation error are Gaussian (Harvey 1991). With these assumptions, each step in a filtering algorithm involves only (possibly multivariate) normal distributions. Normality of every distribution is maintained because multiplying or adding a constant, or adding another normal variable, results in another normal distribution. For example, if $P(X_t|Y_{1:t})$ is normal and X_{t+1} is a linear function of X_t plus Gaussian noise, then $P(X_{t+1}|Y_{1:t})$ is also normal. When the model is mildly nonlinear, one can use approximations based on Taylor series expansions to approximate the filtering calculations using means and variances as if the model was linear. This is known as the extended Kalman filter (Harvey 1991).

The Kalman filter has a long history of application to population models. Much of the development of its application was inspired by fisheries time-series (Mendelssohn 1988; Sullivan 1992; Schnute 1994), as is also true for later state–space modeling developments. More recently, it has been used for models of community dynamics (Ives et al. 2003). To a substantial extent, state–space modeling efforts have moved on to numerical methods for nonlinear and/or non-Gaussian models, with a large emphasis on Bayesian applications, but the basic or extended Kalman filter is still used when it is viewed as a reasonable approximation (e.g., Ennola et al. 1998; Dennis et al. 2006; Woody et al. 2007).

Of particular interest here, Besbeas et al. (2002) showed how combining an approximately linear, Gaussian state–space model with demographic data in an integrated population model can improve results from both.

The great advantage of the Kalman filter is that it is fast and easy to calculate. This means that one does not need to spend significant amounts of time trying more difficult methods that may not yield much more biological insight. Additionally, it means that resampling methods, such as bootstrapping, could be readily applied. For example, one could estimate confidence intervals of estimated parameters using a moving block (nonparametric) bootstrap or a parametric (simulated from the model) bootstrap (Efron and Tibshirani 1993), which might somewhat address concerns if the model is obviously approximate. Another reasonable view is that population dynamics data are typically so noisy, and our explanations of them so full of uncertainty, that it may be difficult to justify a more complicated model. The disadvantages of the Kalman filter are that in many settings the required assumptions are unrealistic, and the consequent impact on estimation and inference can be unclear. Even though other methods can be harder to implement, they may be justified by the goal of learning as much as one possibly can from hard-won ecological data.

Grid-based methods (quadrature)

Grid-based methods work by splitting the range of possible state values into many small cells and tracking the probabilities that the state falls in any cell. For example, the distribution $P(X_t|Y_{1:t})$ can be represented as a set of small cells for the possible X_t values and a vector of corresponding values of $P(X_t|Y_{1:t})$ for the center of each cell. In essence, the piecewise-linear plot of the vector of $P(X_t|Y_{1:t})$ values versus the cell centers is an approximation of the full distribution $P(X_t|Y_{1:t})$. One can use this approximation to calculate the vector that represents $P(X_{t+1}|Y_{1:t})$. Similarly, one can sum over the vector that represents $P(X_{t+1}|Y_{1:t})$ to find the value of $P(Y_{t+1}|Y_{1:t})$, and so on. This method was developed by Kitagawa (1987) and used by de Valpine and Hastings (2002) for population models. Quadrature is the label for numerical integration methods that sum the area under the curve of some continuous function by splitting the range of the x-axis into many small cells and summing the area under the resulting rectangles or trapezoids. This is essentially what is being done to calculate $P(Y_{t+1}|Y_{1:t})$.

The advantages of this method are that it can be computationally efficient for models with low-dimensional X_t and it does not require Monte Carlo methods. The primary disadvantage is that it is difficult to extend to many dimensions of state or observation variables, such as in an age- or stage-structured model or a multi-species model. This difficulty arises from the curse of dimensionality. If the grid has 1,000 cells in each dimension, that will be 10^6 cells in two dimensions, 10^9 in three, and so on, so that storing and summing values on the grid is impractical. Although this disadvantage has limited its adoption for general use, it is nevertheless useful to keep grid-based integration in mind as a potential tool for some dimensions of some problems that can be combined with other tools.

Sequential Monte Carlo methods (particle filtering)

Sequential Monte Carlo methods, also known as "particle filter" methods, represent the distributions of the filtering steps using simulated samples. For example, the distribution $P(X_t|Y_{1:t})$ can be represented by a large sample of X_t values drawn from this distribution, but the samples must be updated sequentially following the steps of filtering. If we have a valid sample from $P(X_t|Y_{1:t})$, then a sample from $P(X_{t+1}|Y_{1:t})$ can be generated by simulating one value of X_{t+1} from each value in the sample of X_t. The probability of the data at $t + 1$ given previous values, i.e. $P(Y_{t+1}|Y_{1:t})$, is simply the average over the X_{t+1} values of the probability of Y_{t+1} given X_{t+1}.

Updating the points based on the information in Y_{t+1}, i.e., filtering step (5), is conceptually simple but leads to the major hitch in this method. The conceptually simple part is to weight each sample point of X_{t+1} proportionally to $P(Y_{t+1}|X_{t+1})$. If one then resamples proportionally to those weights, the result represents a sample from $P(X_{t+1}|Y_{1:t+1})$. Points that are closer to Y_{t+1} will be weighted higher and resampled more often, and vice versa for points far from Y_{t+1}. The problem is that after many time steps the quality of the approximation can degenerate. For example, if Y_{t+1} is an unlikely observation, perhaps only a few X_{t+1} points will be weighted very heavily and represent most of the resampled points. One run of a particle filter approximates one likelihood calculation, so many runs must be used in an optimization search for maximum-likelihood parameters.

Particle filters date back to Gordon et al. (1993), although early related concepts go back further (Cappé et al. 2007). In ecology, particle filters have been tried by Trenkel et al. (2000), de Valpine (2004), Fujiwara et al. (2005, for an individual growth model), Dowd (2006), and Newman et al. (2009). Thorough treatments can be found in Doucet et al. (2001), Liu (2001), and Cappé et al (2007).

Several methods have been proposed to improve the accuracy of particle filters for each likelihood calculation. One is to improve the projection of state values by disproportionately sampling X_t values that lead to X_{t+1} values that are close to Y_{t+1}, while tracking weights appropriately (Pitt and Shephard 1999). This can be done by simulating each X_t value ahead once, weighting them by how close

they come to Y_{t+1}, and then sampling the X_t according to those weights to project ahead again. Other ways to "look ahead" can also be used. This is called auxiliary particle filtering. It allows multiple values of X_{t+1} to come from the same X_t values that are likely to predict Y_{t+1}. The X_{t+1} values are still resampled proportionally to $P(Y_{t+1}|X_{t+1})$.

Another way to improve particle filtering is to use MCMC steps to replenish the particles. Gilks and Berzuini (2001) proposed that MCMC can be used to mix each set of previous trajectories, $X_{1:t-1}$, with simulated current states, X_t. In other words, once a sample for $X_{1:t-1}$ is obtained, each trajectory in it is kept together as a unit for mixing with the next state. A third way to improve particle filtering is to use kernel density smoothing of the state density at each step (Trenkel et al. 2000; Hürzeler and Künsch 2001; Thomas et al. 2005). A new set of particles can be simulated from the smoothed density. Both of these methods can go a long way to replenishing particles, but they are not a panacea in cases where predicted densities are very far from observations. Consequently, they will have difficulty providing accurate likelihoods for bad models.

A serious difficulty with particle filtering is that its likelihood approximations have a stochastic ingredient that changes for every run of the filter, making it hard to find maximum-likelihood parameters (see Fig. 4 of de Valpine 2004). When used in an optimization search, large improvements toward the maximum likelihood can be made easily because they will not be obscured by the randomness in the approximation. However, finer scale convergence to the maximum likelihood will be obscured because finding the precise peak of a stochastic mountain is not easy.

Several methods have been proposed to overcome this difficulty so that particle filtering can be used for maximum-likelihood estimation. Johansen et al. (2008) use a particle filter for its ability to sample from $P(X_{1:T}|Y_{1:T})$ and subsequently use those samples in a data cloning algorithm (see below). Ionides et al. (2006) developed a scheme for iterating particle filters with parameters allowed to vary through time, at first greatly and then progressively less until convergence is forced.

Importance sampling

In importance sampling, first a sample is drawn from a known distribution that approximates $P(X_{1:T}|Y_{1:T}, \theta)$, then a weighted average of probabilities gives the likelihood. The difficulty with this method is that a general approach to finding approximating distributions may not be easy to establish. A distribution that is too small (has tails that are too light) will give a calculation that may not even be valid, while a distribution that is too big (has tails that are too heavy) may be inefficient, i.e., has high simulation variance (Robert and Casella 1999).

Examples of importance sampling for Bayesian inference in ecology include McAllister et al. (1994) and Givens and Raftery (1996). To a large extent Bayesian methods have moved on to MCMC. For state–space maximum likelihood, Durbin and Koopman (1997) showed efficient importance sampling in the limited case of non-Gaussian observations with linear dynamics. Advances in importance sampling techniques could make it more appealing for more complex models (Neddermeyer 2009).

Methods based on MCMC sampling of states only

Monte Carlo expectation maximization

The expectation maximization or EM algorithm is one of the all-time most important algorithms for finding maximum-likelihood estimates. It is designed for models with "missing data", such as unobserved states. However, for the standard EM algorithm to work, one must be able to work with the various distributions analytically. The Monte Carlo version of the EM algorithm (MCEM; Chan and Ledolter 1995) can be used even without analytical tractability. The algorithm works as follows:

1. Start with some parameters θ.
2. Use an MCMC to sample from $P(X_{1:T}|Y_{1:T}, \theta)$.
3. Find a new value of θ that maximizes the average of $\log(P(Y_{1:T}|X_{1:T}, \theta)P(X_{1:T}|\theta))$, averaged over the sample of $X_{1:T}$ values from the previous step.
4. Repeat steps 2 and 3 until converged.

MCEM is appealing in its simplicity and relatively ease of implementation. However, it too has some serious drawbacks. One is that EM algorithms, stochastic or not, are known to suffer from slow convergence in some cases. A greater difficulty is that for each updated value of θ, a new sample of $X_{1:T}$ is generated, so that the convergence involves a stochastic surface, similarly to maximization of particle filter likelihoods. Various improvements to both problems are given by Levine and Casella (2001), Caffo et al. (2005), and Jank (2006).

Methods based on Bayesian MCMC sampling of states and parameters

While the methods for maximum-likelihood estimation have been progressing steadily, they have been slowed by the challenges mentioned above. Meanwhile, the facility of MCMC has allowed Bayesian analysis of state–space models to flourish more fully (e.g., Carlin et al. 1992; Rivot et al. 2004; Rotella et al. 2009). Two relatively recent methods take advantage of Bayesian MCMC sampling to

estimate maximum-likelihood parameters. A potential advantage of these methods over the previous ones is that they can straightforwardly be made as accurate as desired by increasing computational effort. The same is in principle true for the above methods, but in practice runs into the maximum-likelihood convergence problems mentioned (as opposed to MCMC mixing issues, which need to be considered for all of these methods).

Data cloning

If one pretends to have multiple copies of the same data, it is clear that the maximum-likelihood estimate would be the same as for the single copy of the data but that the parameter uncertainty would be spuriously reduced. The spurious reduction in parameter uncertainty would be reflected by a more peaked likelihood surface. In a Bayesian analysis, the more peaked likelihood would create a more peaked posterior. For very many copies of the same data, the posterior will be very sharply peaked at the maximum-likelihood parameters. Several authors have independently hit upon this or very similar ideas as a tool for finding maximum-likelihood parameters (Doucet and Tadic 2003; Jacquier et al. 2007; Lele et al. 2007). Lele et al. (2007) dubbed it "data cloning" and showed how Fisher information can easily be obtained from the results. The implementation of data cloning requires a separate set of latent population states, $X_{1:T}$, for each of many copies of the data. MCMC sampling is then done for the model parameters and for every set of $X_{1:T}$ values. This is conceptually straightforward and can be directly implemented in MCMC software, such as WinBUGS, but determining how many clones (copies of the data) are needed currently requires trial and error, and computation time increases with every copy. Ponciano et al. (2009) demonstrated an application of this approach.

Monte Carlo kernel likelihood

Another way to take advantage of Bayesian MCMC sampling is to estimate the likelihood from the posterior density sample. To do this, the effect of the prior must be undone, which can be accomplished by a weighted kernel density estimator. Define $K(\theta - \theta^{(i)})$ to be a kernel function such as a multivariate Gaussian density with zero mean and covariance Σ_h, also called the "bandwidth". Here, θ is any value of parameters, and $\theta^{(i)}$ is a value from a posterior sample. Then, the likelihood can be approximated up to an unknown normalizing constant by:

$$L(\theta) \propto \frac{1}{m} \sum_{i=1}^{m} \frac{K(\theta - \theta^{(i)})}{P(\theta^{(i)})} \tag{19}$$

where $P(\theta^{(i)})$ is the prior for $\theta^{(i)}$ and m is the size of the posterior sample. This is a straightforward idea but had not

been developed for the state–space model maximum-likelihood estimation until de Valpine (2004).

The main difficulty with this method is that kernel density estimation is a questionable enterprise in more than a few dimensions, i.e., for more than a few parameters. To allow a narrow bandwidth, i.e. Σ_h with small variances, one needs very many samples. Conventional wisdom from the problem of estimating full density surfaces is that the curse of dimensionality makes multivariate estimation impractical beyond three to five dimensions (Scott 1992). However, several points make the outlook more optimistic for estimating the mode location, i.e., the maximum-likelihood parameters. First, this is the region of most accurate estimation. Second, the asymptotics of mode estimation are different from those of full surface estimation (Romano 1988). Third, very large sample sizes are indeed practical because they are simulated by MCMC. And fourth, de Valpine (2004) presented two methods for improving accuracy and showed by simulation that good results can be obtained in up to 20 dimensions.

In practice, there are two sources of error in the mode estimation: bias due to smoothing and random error due to Monte Carlo estimation. Since the likelihood surface will often be approximately Gaussian, which is symmetric, bias due to smoothing can be small even for large bandwidth. One accuracy improvement is to zoom in on the maximum-likelihood estimate by using a preliminary estimate as a prior distribution for a second (or more) MCMC sample. The other is to apply an approximation from distribution theory to reduce the smoothing bias. The test problem used by de Valpine (2004) to assess accuracy was chosen to have a highly asymmetric likelihood surface in all dimensions. Areas for further development include automated bandwidth selection and better corrections to smoothing bias. In summary, this approach appears to be promising and flexible but is not fully automated and will have limits in the number of parameters that can be handled.

Methods for calculating likelihoods

Likelihoods are an example of a "normalizing constant" in Bayes' theorem for conditional distributions. To see this, write the conditional distribution for states given data, with parameters θ fixed:

$$P(X_{1:T}|Y_{1:T}, \theta) = \frac{P(Y_{1:T}|X_{1:T}, \theta)P(X_{1:T}|\theta)}{P(Y_{1:T}|\theta)} \tag{20}$$

The denominator is the likelihood. It is a normalizing constant for $P(Y_{1:T}|X_{1:T}, \theta)P(X_{1:T}|\theta)$ because $P(X_{1:T}|Y_{1:T}, \theta)$ must integrate to 1. There are several major approaches

to calculate normalizing constants numerically (Han and Carlin 2001). It is worth noting that in Bayesian analysis, marginal likelihoods are normalizing constants, but the ratios of marginal likelihoods (Bayes factors) can also be accessed by including the model set in MCMC sampling via reversible jump methods (King et al. 2008a, 2009).

Filtering methods

All of the filtering methods described above can be used to calculate the likelihood even if the maximum-likelihood parameters are found by some other method. This represents a utility of filtering methods that does not involve optimization difficulties. However, it is still relevant that particle filtering can suffer from inefficiency due to particle degeneracy.

Importance sampling

Importance sampling also may be difficult to use for estimation, but it can be more feasible for calculating likelihoods of estimated parameters. In determining an approximating distribution, one can use an MCMC sample from $P(X_{1:T}|Y_{1:T}, \theta)$ itself. However, general distributional forms for multivariate samples are technically challenging (Neddermeyer 2009).

Monte Carlo likelihood ratio approximation

Statisticians have recognized that a conditional sample of latent states (given the data) for one set of parameters or one model can be treated as an importance sampling distribution for another set of parameters or another model. The result is an approximation of a likelihood ratio between two models (Thompson and Guo 1991). Ponciano et al. (2009) used this approach to obtain likelihood profiles and make AIC comparisons among models. This is useful and easy to implement when it works well, but it will not work well if the models have substantially different conditional state distributions (see, for example, Robert and Casella (1999) on importance sampling). [Geyer and Thompson (1992) and Geyer (1996) proposed using this method iteratively to obtain maximum-likelihood estimates, but McCulloch (1997) and de Valpine (2004) did not find this to work efficiently for their examples.]

Bridge sampling

The bridge sampling approach to normalizing constant calculation approaches the problem very differently. One

chooses some particular value of $X_{1:T}$, call it $X_{1:T}^*$, calculates

$P(X_{1:T}^*|Y_{1:T}, \theta)$, and then obtains
$P(Y_{1:T}|\theta) = P(Y_{1:T}|X_{1:T}^*, \theta)P(X_{1:T}^*|\theta)/P(X_{1:T}^*|Y_{1:T}, \theta).$

In other words, one only needs the normalized conditional density at one point to know the normalizing constant. The procedure involves the following steps:

1. Choose some states $X_{1:T}^*$. A sensible choice is the average from an MCMC sample from $P(X_{1:T}|Y_{1:T}, \theta)$.
2. For each t from 1 to T:

   (a) Define $f_1(X_{t:T}) = P(X_{t:T}|X_{1:t-1}^*, Y_{1:T}, \theta)$. The first $t - 1$ states are fixed.

   (b) Define

   $$f_2(X_{t:T}) = P(X_{t+1:T}|X_{1:t}^*, Y_{1:T}, \theta) \\ \times q(X_t|X_{1:t}^*, X_{t+1:T}, Y_{1:T}, \theta),$$

   where $q(\cdot)$ is a density you choose. It plays a role like a MCMC proposal density.

   (c) Run one MCMC sampler with $X_{1:t-1}$ fixed at $X_{1:t-1}^*$ and another with $X_{1:t}$ fixed at $X_{1:t}^*$. (In practice, the first sample can be re-used from the second sample of the $(t - 1)$st iteration of step 2.) The first sample is proportional to $f_1(\cdot)$.

   (d) Simulate from $q(\cdot)$ to supplement the second MCMC sample with an X_t dimension. Together these are proportional to $f_2(\cdot)$. In practice $q(\cdot)$ can be chosen adaptively once the MCMC samples are in hand.

   (e) Use the bridge sampling identity (not given) to calculate approximately the ratio of the normalizing constants of $f_1(\cdot)$ and $f_2(\cdot)$, which is:

   $$r_t = \frac{P(X_{1:t}^*|Y_{1:T}, \theta)}{P(X_{1:t-1}^*|Y_{1:T}, \theta)} \tag{21}$$

3. The product of the r_t values is $P(X_{1:T}^*|Y_{1:T}, \theta)$.

Bridge sampling methods are promising in terms of accuracy and general applicability (Han and Carlin 2001). A major advantage is that they do not require the model to be good. The basic idea for calculating ratios of normalizing constants is from Meng and Wong (1996). The idea for calculating the conditional probability at a point is from Chib (1995) for Gibbs samplers and Chib and Jeliazkov (2001) for the Metropolis–Hastings algorithm, but these authors did not relate the approach to bridge sampling. Mira and Nicholls (2004) made this connection and found immediate efficiency gains using the results of Meng and Wong (1996). De Valpine (2008) showed additional efficiency gains, mostly by adaptive estimation of $q(\cdot)$ [step 2(d) above].

Lapwing example

Next I use the MCKL and bridge sampling methods to find maximum-likelihood estimates and profile likelihood confidence intervals of the Lapwing model. The goal is to see the frequentist methods used in a somewhat intensive way—since repeated maximization and normalizing constant calculation must be done to obtain likelihood profiles—to illustrate their feasibility. A customized MCMC sampler was developed to allow complete control over sampling dimensions and proposal steps, rather than relying on the choices of WinBUGS, for example. Some salient details are as follows:

– All priors were flat.
– The prior on $\tau_y = 1/\sigma_y^2$ was flat on a log scale. Boundaries for τ_y such that $1 \le \sigma_y \le 10^5$ were used with no impact whatsoever on the results. Using a flat prior on a log scale was observed to lead to a more normal posterior (compared to a flat prior for τ_y itself), which reduces smoothing bias in the kernel density estimate of the mode.
– All proposals were normally distributed and centered on the current values. Standard deviations of proposal distributions were heuristically tuned by trial and error to achieve good mixing.
– Covariates were centered at zero. For example, the mean annual number of days below freezing was subtracted from each year's value. This allows better mixing of the slope and intercept parameters.

A sample of 300,000 parameter values, taken by recording every 30th sample from 9,000,000, was obtained for MCKL estimation of the maximum-likelihood parameters. For kernel density estimation, the posterior sample was transformed to principal components for all dimensions except for $\log(\tau)$. Since the principal components will be approximately independent, this choice allows a diagonal covariance "bandwidth" matrix for the kernel smoother to be reasonable. Omitting $\log(\tau)$ from dimensions transformed to principal components was done because it has minimum and maximum values. The kernel estimate must be re-normalized based on the area of the kernel that extends outside allowed sample values [a detail omitted from (Eq. 19)], which is simplest if a dimension with boundaries is not included in the principal components. In this case, the sample did not come close to the boundaries, and so this issue is moot. The bandwidth used in each dimension was 0.5 of the standard deviation of the marginal posterior.

Likelihoods given fixed parameters were calculated using bridge sampling. The improvements of de Valpine (2008) were not implemented for this problem, and instead

the method of Mira and Nicholls (2004) was used. Each iteration requires a sample of some states with parameters and other states held fixed. These sample sizes were 50,000, recorded as every eighth sample from 400,000.

A likelihood profile for each parameter was obtained using a grid of nine values as follows. For each of the nine values of the parameter, MCKL was used to find the maximum-likelihood estimates for the rest of the parameters. Sample sizes for this step were 200,000, thinned as every 30th value and with bandwidths chosen as above. Bridge sampling estimates of the likelihoods were then calculated at all nine sets of parameters. A spline was fit through the log likelihood as a function of the parameter being profiled, and the location of the 95% upper and lower boundaries was determined from the spline fit based on standard chi-squared quantiles. Each grid included the maximum-likelihood estimate as the central (5th) value. The extent of the grids was chosen by eye from the Bayesian marginal posterior, which was adequate as long as it was wide enough to include both upper and lower 95% boundaries.

Results are given in Table 1, and example profiles are shown in Figs. 1, 2, and 3. The main point to be taken from the figures is that the profile log-likelihoods appear to be very nearly negative quadratic, which is the expected pattern of a well-behaved likelihood surface informed by a healthy amount of data. If the methods displayed unacceptable error due to the simulation aspect of the algorithms or other problems, these would typically be revealed by poorly behaved likelihood profiles.

The estimates and confidence intervals obtained here are similar to results from Brooks et al. (2004, Table 1) and Besbeas et al. (2002, Table 1). Some apparently discrepant numbers can be explained by different scalings and centerings of the time (year) and frost data as covariates. I treated time as the integers 1–35 centered around their

Table 1 Maximum-likelihood estimates and 95% profile confidence intervals for Lapwing model

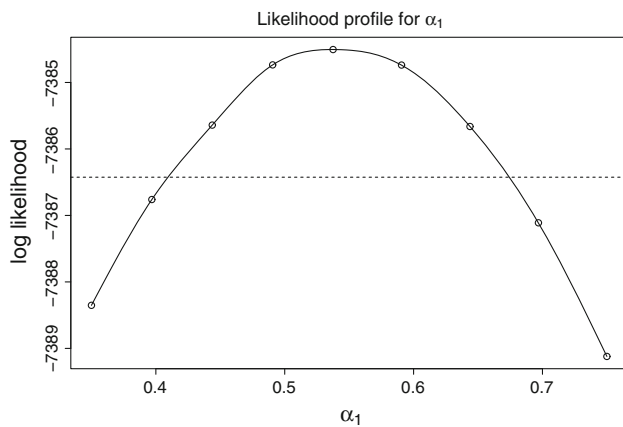| Parameter	Maximum-likelihood estimate	95% Confidence interval
α_1	0.54	(0.41, 0.67)
β_1	−0.19	(−0.31, −0.078)
α_a	1.54	(1.41, 1.68)
β_a	−0.24	(−0.32, −0.17)
α_λ	−4.57	(−4.63, −4.49)
β_λ	−0.034	(−0.042, −0.027)
α_ρ	−1.13	(−1.31, −0.97)
β_ρ	−0.027	(−0.036, −0.018)
σ_y	159	(122, 211)

Fig. 1 Likelihood profile for α_1. This is the intercept of the logistic relationship between 1-year-old survival and number of frost days in a year (Eq. 8). *Middle circle* Maximum-likelihood estimate. Other *circles* values of α_1 that were held fixed while other parameters were maximized to obtain the profile likelihood surface, which is shown as a spline curve (*solid*). *Dashed line* Log likelihood threshold for a 95% confidence interval. Values of α_1 with log likelihoods above the *dashed line* are within the confidence interval
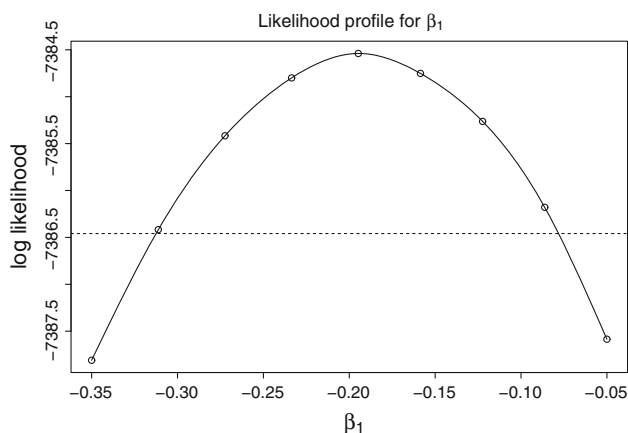
Fig. 2 Likelihood profile for β_1. This is the slope of the logistic relationship between 1-year-old survival and number of frost days in a year (Eq. 8). *Circles, solid curve*, and *dashed line* are as in Fig. 1

mean of 18, resulting in the integers from -17 to 17. Brooks et al. (2004, Annex with code) treated time as integers from 1 to 35 for the ρ regression and from 2 to 36 for the λ regression. Besbeas et al. (2002 and personal communication) scaled time as evenly spaced numbers from -1 to 1. Brooks et al. (2004) used frost data approximately centered around 0 and scaled to have standard deviation 1. I used the data directly from their paper, while Besbeas et al. (2002) used approximately centered but not rescaled frost data.

Of the eight slope and intercept parameters, the only two that appear substantially different from Brooks et al. (2004) are α_ρ and α_λ, the intercepts for the two regressions against time. If $\hat{\alpha}_\rho = -0.668$ of Brooks et al. (2004) is
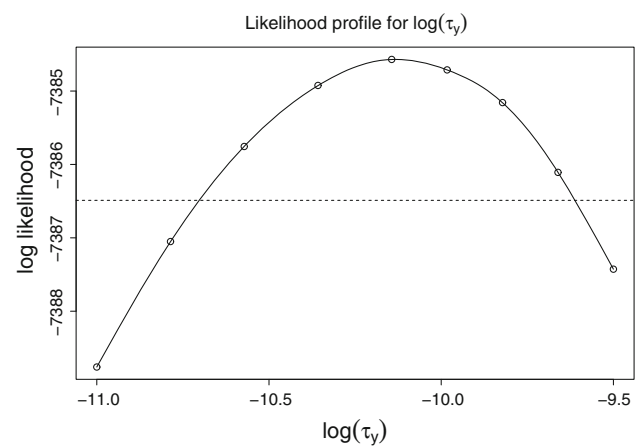

Fig. 3 Likelihood profile for $\log(\tau_y)$. This is the log of $1/\sigma^2_y$, where σ^2_y is the sampling variance for adult abundance index. *Circles, solid curve*, and *dashed line* are as in Fig. 1

shifted by $\hat{\beta}_\rho \times 18 = -0.027 \times 18$, the result is -1.15, close to -1.13 estimated here. Similar adjustment of $\hat{\alpha}_\lambda$ gives -4.56, close to -4.57 estimated here.

Turning to comparison with Besbeas et al. (2002), all four slope parameters appear to be different. It makes sense that α_ρ and α_λ are very similar, since both these authors and I centered time around 0. To compare the slopes of the time regressions, namely, β_ρ and β_λ, one can divide the slopes of Besbeas et al. (2002) by the interval representing one year on their time scaling, i.e. 2/(number of years $-$ 1), which yields values quite close to those estimated here. The ratios of slope estimates between my results and theirs for both first-year and juvenile survival are similar to the standard deviation of their frost data, again reconciling the results as different parameterizations of biologically similar models (P.T. Besbeas, personal communication). The confidence interval or region widths also appear to be generally similar among all three papers. Finally, the estimate of 159 for σ_y is identical to that of Besbeas et al. (2002) and slightly smaller than the 169 of Brooks et al. (2004).

I conclude that all three methods yield similar biological conclusions in this case. It is natural then to ask when one approach or another will be superior, but that is beyond the territory of this paper. The overarching point here is that computational maximum-likelihood methods make it possible to pursue such comparisons at all by allowing frequentist analysis of nonlinear, non-Gaussian state–space models.

Discussion

Availability of computational methods for Bayesian analysis of hierarchical models has rendered the Bayesian choice one of great pragmatism. Researchers often face a

dichotomy between frequentist analysis using a simple model for their data or Bayesian analysis using a hierarchical model that more realistically represents the multiple sources of variation and relationships in their data. A principal aim of this review has been to touch upon many areas of active statistical research to make frequentist analysis of realistic hierarchical models feasible. Some of these methods have been explored for ecological problems, and there is great potential for more.

To a reader new to these topics, the review of methods with pros and cons of each may appear to convey the message that all of them are problematic to one degree or another. This is indeed the case for these as well as Bayesian computational methods, on which there is a large body of literature of attempts to improve many recognized difficulties. However, even at their current stage of development, the methods reviewed here are practical for many ecological problems. Several of the methods leverage MCMC algorithms, which are now widely available, and all continue to see improvements developed.

What methods can be most recommended? As Newman et al. (2009) discuss, there can be tradeoffs between ease of implementation, computational efficiency, and accuracy of results. The Kalman filter is the easiest to implement but most limited in the models it can handle. For Monte Carlo methods, automated software for the algorithm steps other than MCMC is currently a limitation, but in time this may be resolved. The three methods for maximum-likelihood estimation that stand out as easiest to implement are MCEM, data cloning, and MCKL. MCEM requires iteration of MCMC and optimization. Data cloning requires almost no additional steps beyond MCMC, and so may be currently easiest to use when feasible. MCKL requires only kernel density estimation and optimization, which are both readily available in mathematical software packages (and straightforward to program). The programming required for these methods is at a level that is sometimes a barrier for applied ecologists but very feasible for many statisticians. For normalizing constant estimation, particle filtering and importance sampling are relatively straightforward to implement, while bridge sampling has the potential for greater efficiency but requires substantially more implementation effort. In summary, while these methods are not currently automatic to use, they can be made practical for many problems.

References

Barry SC, Brooks SP, Catchpole EA, Morgan BJT (2003) The analysis of ring-recovery data using random effects. Biometrics 59(1):54–65

Bengtsson T, Cavanaugh J (2006) An improved Akaike information criterion for state-space model selection. Comput Stat Data Anal 50(10):2635–2654

Besbeas P, Freeman SN (2006) Methods for joint inference from panel survey and demographic data. Ecol Lett 87(5):1138–1145

Besbeas P, Freeman SN, Morgan BJT, Catchpole EA (2002) Integrating mark-recapture-recovery and census data to estimate animal abundance and demographic parameters. Biometrics 58(3):540–547

Besbeas P, Lebreton JD, Morgan BJT (2003) The efficient integration of abundance and demographic data. J R Stat Soc Ser C Appl Stat 52:95–102

Besbeas P, Borysiewicz RS, Morgan B (2009) Completing the ecological jigsaw. In: Thomson DL, Cooch EG, Conroy MJ (eds) Environmental and ecological statistics series, vol 3: modeling demographic processes in marked populations. Springer, New York, pp 513–539

Borchers DL, Buckland ST, Zucchini W (2002) Estimating animal abundance: closed populations. Springer, London

Brillinger DR, Guckenheimer J, Guttorp P, Oster G (1980) Empirical modelling of population time series data: the case of age and density dependent vital rates. In: Oster GF (ed) Lectures on mathematics in the life sciences, vol 13: some mathematical questions in biology. The American Mathematical Society, Providence, pp 65–90

Brooks SP, King R, Morgan BJT (2004) A Bayesian approach to combining animal abundance and demographic data. Anim Biodiv Conserv 27(1):515–529

Buckland ST, Anderson DR, Burnham KP, Laake JL, Borchers DL, Thomas L (2001) Introduction to distance sampling: estimating abundance of biological populations. Oxford University Press, New York

Buckland ST, Newman KB, Thomas L, Koesters NB (2004) State-space models for the dynamics of wild animal populations. Ecol Model 171(1–2):157–175

Burnham K, White G (2002) Evaluation of some random effects methodology applicable to bird ringing data. J Appl Stat 29:245–264

Caffo BS, Jank W, Jones GL (2005) Ascent-based Monte Carlo expectation-maximization. J R Stat Soc Ser B Stat Methodol 67:235–251

Calder C, Lavine M, Muller P, Clark JS (2003) Incorporating multiple sources of stochasticity into dynamic population models. Ecol Lett 84(6):1395–1402

Cam E, Link W, Cooch E, Monnat J, Danchin E (2002) Individual covariation in life-history traits: seeing the trees despite the forest. Am Nat 159(1):96–105

Cappé O, Godsill SJ, Moulines E (2007) An overview of existing methods and recent advances in sequential Monte Carlo. Proc IEEE 95(5):899–924

Carlin BP, Polson NG, Stoffer DS (1992) A Monte-Carlo approach to nonnormal and nonlinear state-space modeling. J Am Stat Assoc 87(418):493–500

Catchpole EA, Morgan BJT, Freeman SN, Peach W (1999) Modelling the survival of British Lapwings, *Vanellus vanellus*, using ring-recovery data and weather covariates. Bird Study 46:S5–S13

Cavanaugh J, Shumway R (1997) A bootstrap variant of AIC for state-space model selection. Stat Sinica 7(2):473–496

Chan KS, Ledolter J (1995) Monte-Carlo EM estimation for time-series models involving counts. J Am Stat Assoc 90(429):242–252

Cheng B, Tong H (1992) On consistent nonparametric order determination and chaos. J R Stat Soc Ser B Stat Methodol 54(2):427–449

Chib S (1995) Marginal likelihood from the Gibbs output. J Am Stat Assoc 90:1313–1321

Chib S, Jeliazkov I (2001) Marginal likelihood from the Metropolis–Hastings output. J Am Stat Assoc 96(453):270–281

Clark JS (2007) Models for ecological data: an introduction. Princeton University Press, Princeton

Clark JS, Bjornstad CN (2004) Population time series: process variability, observation errors, missing values, lags, and hidden states. Ecol Lett 85(11):3140–3150

Cressie N, Calder CA, Clark JS, Ver Hoef JM, Wikle CK (2009) Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. Ecol Appl 19:553–570

Davison A, Hinkley A (1997) Bootstrap methods and their application. Cambridge University Press, New York

Dennis B, Ponciano JM, Lele SR, Taper ML, Staples DF (2006) Estimating density dependence, process noise, and observation error. Ecol Monogr 76(3):323–341

Dennis B, Ponciano JM, Taper ML (2010) Replicated sampling increases efficiency in monitoring biological populations. Ecol Lett 91(2):610–620. doi:10.1890/08-1095.1

de Valpine P (2002) Review of methods for fitting time-series models with process and observation error and likelihood calculations for nonlinear, non-Gaussian state-space models. Bull Mar Sci 70(2):455–471

de Valpine P (2004) Monte Carlo state–space likelihoods by weighted posterior kernel density estimation. J Am Stat Assoc 99(466):523–536

de Valpine P (2008) Improved estimation of normalizing constants from Markov chain Monte Carlo output. J Comput Graph Stat 17:333–351

de Valpine P (2009) Shared challenges and common ground for Bayesian and classical analysis of hierarchical models. Ecol Appl 19:584–588

de Valpine P, Hastings A (2002) Fitting population models incorporating process noise and observation error. Ecol Monogr 72(1):57–76

de Valpine P, Hilborn R (2005) State-space likelihoods for nonlinear fisheries time-series. Can J Fish Aquat Sci 62(9):1937–1952

Doucet A, Tadic VB (2003) Parameter estimation in general state-space models using particle methods. Ann Inst Stat Math 55(2):409–422

Doucet A, De Freitas N, Gordon N (2001) Sequential Monte Carlo methods in practice. Statistics for engineering and information science. Springer, New York

Dowd M (2006) A sequential Monte Carlo approach for marine ecological prediction. Environmetrics 17(5):435–455

Durbin J, Koopman SJ (1997) Monte Carlo maximum likelihood estimation for non-Gaussian state space models. Biometrika 84(3):669–684

Durbin J, Koopman SJ (2001) Time series analysis by state space methods. Oxford University Press, Oxford

Efron B (1996) Empirical Bayes methods for combining likelihoods. J Am Stat Assoc 91(434):538–550

Efron B (2005) Bayesians, frequentists, and scientists. J Am Stat Assoc 100(469):1–5

Efron B, Tibshirani R (1993) An introduction to the bootstrap. Chapman and Hall, New York

Ellner SP, Fieberg J (2003) Using PVA for management despite uncertainty: effects of habitat, hatcheries, and harvest on salmon. Ecol Lett 84(6):1359–1369

Ennola K, Sarvala J, Devai G (1998) Modelling zooplankton population dynamics with the extended Kalman filtering technique. Ecol Model 110(2):135–149

Fuh CD (2006) Efficient likelihood estimation in state space models. Ann Stat 34(4):2026–2068

Fujiwara M, Kendall BE, Nisbet RM, Bennett WA (2005) Analysis of size trajectory data using an energetic-based growth model. Ecol Lett 86(6):1441–1451

Gauthier G, Besbeas P, Lebreton JD, Morgan BJT (2007) Population growth in snow geese: A modeling approach integrating demographic and survey information. Ecol Lett 88(6):1420–1429

Gelman A, Carlin JB, Stern HS, Rubin D (2004) Bayesian data analysis. 2nd edn. Chapman & Hall/CRC, Boca Raton

Geyer CJ (1996) Estimation and optimization of functions. In: Gilks WR, Richardson S, Spiegelhalter DJ (eds) Markov Chain Monte Carlo in practice. Chapman and Hall/CRC, New York, pp 241–258

Geyer CJ, Thompson EA (1992) Constrained Monte-Carlo maximum-likelihood for dependent data. J R Stat Soc Ser B Stat Methodol 54(3):657–699

Gilks WR, Berzuini C (2001) Following a moving target—Monte Carlo inference for dynamic Bayesian models. J R Stat Soc Ser B Stat Methodol 63:127–146

Gimenez O, Choquet R (2010) Individual heterogeneity in studies on marked animals using numerical integration: capture-recapture mixed models. Ecol Lett 91(4):951–957

Gimenez O, Rossi V, Choquet R, Dehais C, Doris B, Varella H, Vila JP, Pradel R (2007) State-space modelling of data on marked individuals. Ecol Model 206(3–4):431–438

Givens G, Raftery A (1996) Local adaptive importance sampling for multivariate densities with strong nonlinear relationships. J Am Stat Assoc 91:132–141

Gordon NJ, Salmond DJ, Smith AFM (1993) Novel-approach to nonlinear non-Gaussian Bayesian state estimation. IEE Proc F Radar Signal Proc 140(2):107–113

Han C, Carlin BP (2001) Markov chain Monte Carlo methods for computing Bayes factors: a comparative review. J Am Stat Assoc 96(455):1122–1132

Harvey AC (1991) Forecasting, structural time series models and the Kalman filter. Cambridge University Press, Cambridge

Hürzeler M, Künsch HR (2001) Approximating and maximising the likelihood for a general state-space model. In: Doucet A, De Freitas N, Gordon N (eds) Sequential Monte Carlo Methods in practice. Springer, New York, pp 159–175

Ionides EL, Bretó C, King AA (2006) Inference for nonlinear dynamical systems. Proc Natl Acad Sci USA 103(49):18438–18443

Ives AR, Dennis B, Cottingham KL, Carpenter SR (2003) Estimating community stability and ecological interactions from time-series data. Ecol Monogr 73(2):301–330

Jacquier E, Johannes M, Polson N (2007) MCMC maximum likelihood for latent state models. J Economet 137:615–640

Jank W (2006) Ascent EM for fast and global solutions to finite mixtures: An application to curve-clustering of online auctions. Comput Stat Data Anal 51(2):747–761

Jensen JL, Petersen NV (1999) Asymptotic normality of the maximum likelihood estimator in state space models. Ann Stat 27(2):514–535

Johansen AM, Doucet A, Davy M (2008) Particle methods for maximum likelihood estimation in latent variable models. Stat Comput 18:47–57

Karban R, de Valpine P (2010) Population dynamics of an Arctiid caterpillar-tachinid parasitoid system using state-space models. J Anim Ecol 79(3):650–661

King R, Brooks SP, Coulson T (2008) Analyzing complex capture-recapture data in the presence of individual and temporal covariates and model uncertainty. Biometrics 64(4):1187–1195

King R, Brooks SP, Mazzetta C, Freeman SN, Morgan BJT (2008) Identifying and diagnosing population declines: a Bayesian assessment of Lapwings in the UK. J R Stat Soc Ser C Appl Stat 57:609–632

King R, Morgan B, Gimenez O, Brooks S (2009) Bayesian analysis for population ecology. CRC Press, New York

Kitagawa G (1987) Non-Gaussian state-space modeling of non-stationary time-series. J Am Stat Assoc 82(400):1032–1041

Knape J (2008) Estimability of density dependence in models of time series data. Ecol Lett 89(11):2994–3000

Lele SR, Dennis B, Lutscher F (2007) Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. Ecol Lett 10(7):551–563

Levine RA, Casella G (2001) Implementations of the Monte Carlo EM algorithm. J Comput Graph Stat 10(3):422–439

Link W, Cooch E, Cam E (2002) Model-based estimation of individual fitness. J Appl Stat 29:207–224

Liu JS (2001) Monte Carlo strategies in scientific computing. Springer series in statistics. Springer, New York

Mayo DG, Spanos A (2006) Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. Br J Philos Sci 57(2):323–357

McAllister MK, Pikitch EK, Punt AE, Hilborn R (1994) A Bayesian-approach to stock assessment and harvest decisions using the sampling importance resampling algorithm. Can J Fish Aquat Sci 51(12):2673–2687

McCulloch CE (1997) Maximum likelihood algorithms for generalized linear mixed models. J Am Stat Assoc 92(437):162–170

Mendelssohn R (1988) Some problems in estimating population sizes from catch-at-age data. Fish Bull 86(4):617–630

Meng XL, Wong WH (1996) Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. Stat Sinica 6(4):831–860

Mira A, Nicholls G (2004) Bridge estimation of the probability density at a point. Stat Sinica 14(2):603–612

Neddermeyer JC (2009) Computationally efficient nonparametric importance sampling. J Am Stat Assoc 104(486):788–802

Newman KB, Fernandez C, Thomas L, Buckland ST (2009) Monte Carlo inference for state-space models of wild animal populations. Biometrics 65(2):572–583

O'Hagan A (1994) Kendall's advanced theory of statistics, vol. 2B: Bayesian Inference. Oxford University Press, New York

Pitt MK, Shephard N (1999) Filtering via simulation: auxiliary particle filters. J Am Stat Assoc 94(446):590–599

Ponciano J, Taper M, Dennis B, Lele S (2009) Hierarchical models in ecology: confidence intervals, hypothesis testing, and model selection using data cloning. Ecol Lett 90(2):356–362

Rivot E, Prévost E, Parent E, Baglinière J (2004) A Bayesian state-space modelling framework for fitting a salmon stage-structured population dynamic model to multiple time series of field data. Ecol Model 179:463–485

Robert CP, Casella G (1999) Monte Carlo statistical methods. Springer, New York

Romano JP (1988) On weak-convergence and optimality of kernel density estimates of the mode. Ann Stat 16(2):629–647

Rotella JJ, Link WA, Nichols JD, Hadley GL, Garrott RA, Proffitt KM (2009) An evaluation of density-dependent and density-independent influences on population growth rates in Weddell seals. Ecol Lett 90(4):975–984

Royle J, Link W (2002) Random effects and shrinkage estimation in capture–recapture models. J Appl Stat 29:329–351

Royle JA, Dorazio RM (2008) Hierarchical modeling and inference in ecology. Academic Press, London

Schaub M, Gimenez O, Sierro A, Arlettaz R (2007) Use of integrated modeling to enhance estimates of population dynamics obtained from limited data. Conserv Biol 21(4):945–955

Schnute JT (1994) A general framework for developing sequential fisheries models. Can J Fish Aquat Sci 51(8):1676–1688

Scott DW (1992) Multivariate density estimation: theory, practice, and visualization. Wiley, New York

Severini TA (2001) Likelihood methods in statistics. Oxford University Press, New York

Sullivan PJ (1992) A Kalman filter approach to catch-at-length analysis. Biometrics 48(1):237–257

Tavecchia G, Besbeas P, Coulson T, Morgan BJT, Clutton-Brock TH (2009) Estimating population size and hidden demographic parameters with state-space modeling. Am Nat 173(6):722–733

Thomas L, Buckland ST, Newman KB, Harwood J (2005) A unified framework for modelling wildlife population dynamics. Aust N Z J Stat 47:19–34

Thompson EA, Guo SW (1991) Evaluation of likelihood ratios for complex genetic models. IMA J Math Appl Med Biol 8(3):149–169

Trenkel V, Elston D, Buckland S (2000) Fitting population dynamics models to count and cull data using sampling importance sampling. J Am Stat Assoc 95:363–374

Williams BK, Nichols JD, Conroy MJ (2002) Analysis and management of animal populations, 1st edn. Academic Press, New York

Woody ST, Ives AR, Nordheim EV, Andrews JH (2007) Dispersal, density dependence, and population dynamics of a fungal microbe on leaf surfaces. Ecol Lett 88(6):1513–1524