Chapter 3
State-Space Models

In the previous chapter, a sequence of matrices was used to model the sequence of subprocesses, birth, survival, movement, etc., which characterize population dynamics. We find this building block perspective attractive for at least two reasons: (1) it allows one to mentally "divide and conquer" sometimes complicated population dynamics processes; (2) the resulting product of matrices is a generalization of Leslie and Lefkovitch matrices, something familiar to many biologists.

While the building block matrix model is an aid to model formulation and understanding, it may not be as useful for fitting models to data and making population projections. Matrix representations of population dynamics describe at best the expected changes in a population, for example, conditional expected numbers at time t given numbers at time $t-1$, and often, as noted in Sect. 2.5, are just approximations of these expectations, as in the case of density-dependent birth processes. Consequently, such matrix models fail to describe the variation and uncertainty around these expected outcomes. There are various ways of extending matrix models to incorporate both variability (Quinn and Deriso 1999: Sect. 7.3; Caswell 2001: Chaps. 14 and 15) and nonlinearity (Quinn and Deriso 1999: Sect. 7.4; Caswell 2001: Chap. 16). For example, variability around the expected outcome can be described by randomly selecting elements of a matrix (Caswell 2001: Sect. 14.5.5) or randomly selecting a matrix from a set of matrices (Caswell 2001: Sect. 14.5.3).

However, limitations of stochastic and nonlinear matrix models become apparent when one considers modelling population dynamics that involve a sequence of random and nonlinear sub-processes. Example 1 from the previous chapter, the BAS model, included binomial distributions for survival of two different age classes and a third binomial distribution for births. A Leslie matrix was formulated that accurately characterised the expected states at time t (conditional on $\mathbf{n}_{t-1}$). For projecting the population forward in time, however, one cannot readily, if at all, formulate a stochastic version of the matrix that accurately captures the variation of these binomials. For example, simply adding a vector of random variables to the

K.B. Newman et al., *Modelling Population Dynamics: Model Formulation, Fitting and Assessment using State-Space Methods*, Methods in Statistical Ecology,
DOI 10.1007/978-1-4939-0977-3_3, © Springer Science+Business Media New York 2014

matrix model, i.e. $\mathbf{n}_t = BAS\mathbf{n}_{t-1} + \epsilon_t$, is problematic at best. The distribution for ϵ_t will be quite complex; e.g. its components cannot be so large that a corresponding component of $\mathbf{n}_t$ is less than 0.

State-space models (SSMs) are a more flexible approach for realistically modelling population dynamics than matrix models. SSMs are the backbone of the methods discussed in this book and this chapter is an introduction to their basic structure. Similar to some of the matrix models discussed in the previous chapter, SSMs have a state model and an observation model, but each is now a *stochastic time series*. The state model is a first-order Markov process, i.e. the distribution for state $\mathbf{n}_t$ is defined conditional on the previous state $\mathbf{n}_{t-1}$, and the distribution for the observation $\mathbf{y}_t$ is defined conditional on the current state $\mathbf{n}_t$. The classic SSM is a Normal Dynamic Linear Model (NDLM, West and Harrison 1997), which consists of two normally distributed linear models conditioning on $\mathbf{n}_{t-1}$ and $\mathbf{n}_t$, respectively. The two equations below are a simplistic example of an NDLM in the context of animal abundance dynamics, where N_t is the true, but unknown, abundance of an animal population at time t and y_t is an index of N_t, i.e. $y_t = \gamma N_t$ where γ is a constant of proportionality.

$$\text{State process model} \quad N_t|N_{t-1} \sim \text{normal}\left(\lambda N_{t-1}, \sigma_N^2\right). \tag{3.1}$$

$$\text{Observation model} \quad y_t|N_t \sim \text{normal}\left(\gamma N_t, \sigma_y^2\right). \tag{3.2}$$

Here λ is the population growth rate; in a deterministic setting, $\lambda > 1$ leads to exponential growth, and $\lambda < 1$ is exponential decline. We use the terms state process model and state model interchangeably. Variation around the expected value, in this case λN_{t-1}, is sometimes called process noise or variation. In this example, the magnitude of the process variation depends on the size of σ_N^2. An example of simulated projections of states with $\lambda = 1.02$ (2 % growth rate) and $\sigma_N^2 = 4$, and unbiased observations ($\gamma = 1$) with $\sigma_y^2 = 16$, is shown in Fig. 3.1.

Thus SSMs simultaneously account for two distinct sources of variation, natural or process variation (e.g. environmental or demographic stochasticity) and observation error (e.g. sampling or measurement errors) within a single framework. SSMs are much more general and flexible than matrix models, readily accommodating multiple random nonlinear sub-processes. Given the conditionally-defined state process model, forward stochastic population projection is relatively simple so long as random samples can be generated from the distribution. When the state process is a sequence of stochastic sub-processes, simulation is often easier than evaluation of the pdf; i.e. it is easier to simulate $\mathbf{n}_t$ given $\mathbf{n}_{t-1}$ than it is to calculate $\Pr(\mathbf{n}_t|\mathbf{n}_{t-1})$. The inclusion of a stochastic observation process model for the observations provides a framework for estimating parameters and accounting for uncertainty in the data, in a way which is consistent with the underlying state process model. This is in contrast to standard usage of matrix models where vital rate parameters, or their estimates, are somehow supplied external to the model, and error in such estimates is often not accounted for.

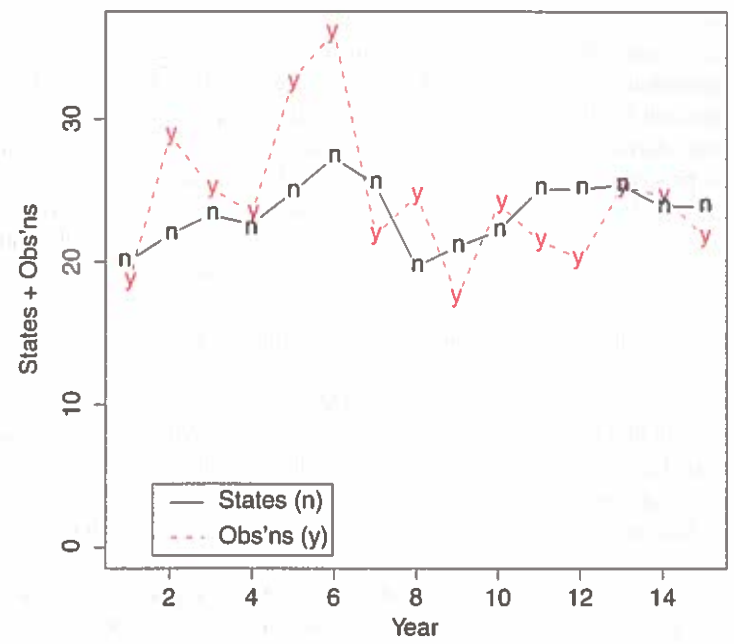


Fig. 3.1 Simulation of NDLM for animal abundance and estimates, where the population dynamics of the states are $N_t \sim \text{normal}(1.02N_{t-1}, 4)$ and estimates are unbiased, $y_t \sim \text{normal}(N_t, 16)$

In addition to the integrative nature of SSMs, there is also opportunity for a convenient division of labour between the modelling of the population dynamics and the modelling of the sampling and measurement of the population. Subject matter specialists such as ecologists can focus their attention on the underlying science through the state process model. The building block matrix models can serve as useful first approximations to the formulation of state process models, e.g. characterizing the deterministic portion of the state process model. Alternative and competing hypotheses about the underlying dynamics can be formalized by alternative state process models. Specialists in sampling, or mark-recapture, or transect sampling, or however the population is monitored, can focus on the formulation of one or more observation models. Different formulations of the observation model might result by working with summarized or derived calculations of sample data, e.g. mark-recapture-based point estimates of abundance, or the raw sample data, e.g. recaptured marks. This potential division of labour, of course, does not preclude a single individual, knowledgeable about the subject matter science and quantification of the observation process, from doing both.

Applications of state-space models to ecological data sets have steadily increased since the late 1980s. Early applications were largely restricted to the special case of NDLMs, because such models could be readily fitted using the Kalman filter (West and Harrison 1997). Several of the first applications were to fisheries data including

Mendelssohn (1988) who fitted an NDLM to Pacific mackerel *Scomber japonicus* catch data categorized by age class to estimate annual abundances and recruitment to the population, and Sullivan (1992) who fitted an NDLM to fisheries catch data categorized by length classes and then estimated parameters such as growth and survival. Advancements in computing power combined with simulation-based estimation procedures, such as Markov chain Monte Carlo (MCMC), extended the class of SSMs that could be fitted to nonlinear and non-normal distributions. Meyer and Millar (1999) gave one of the earliest demonstrations of such models for ecological data, South Atlantic albacore *Thunnus alalunga* biomass, using MCMC to fit an SSM where the state model was a univariate nonlinear, non-normal Schaefer surplus production model and the observation model used catch-per-unit-effort indices.

While many of the first applications of SSMs were to fisheries data, the diversity of species modelled by SSMs has expanded considerably. Animal species modelled by SSMs include red deer *Cervus elaphus* (Trenkel et al. 2000), grey herons *Ardea cinerea* and northern lapwings *Vanellus vanellus* (Besbeas et al. 2002), grey seals *Halichoerus grypus* (Thomas et al. 2005; Harrison et al. 2006), Chinook salmon *Oncorhynchus tshawytscha* (Newman and Lindley 2006), leatherback turtles *Dermochelys coriacea* (Jonsen et al. 2006), black bears *Ursus americanus* (Conn et al. 2008), red grouse *Lagopus lagopus scoticus* (New et al. 2009), hen harriers *Circus cyaneus* (New et al. 2011), desert bighorn sheep *Ovis canadensis mexicana* (Colchero et al. 2009), skate *Leucoraja ocellata* (Swain et al. 2009), Weddell seals *Leptonychotes weddelli* (Rotella et al. 2009), California sea lions *Zalophus californianus* (Ward et al. 2010), octopus *Octopus vulgaris* (Robert et al. 2010) and the Glanville fritillary butterfly *Melitaea cinxia* (Harrison et al. 2011).

State variables other than abundances have been considered as well. Royle and Kéry (2007) let the true state be whether or not a particular site was occupied by animals (and the observations were imperfect estimates of presence or absence). Gimenez et al. (2007) and Royle (2008), considering marked and recaptured or recovered animals, let the true state of marked animals (whether still alive or not) be the state variable. Anderson-Sprecher and Ledolter (1991) modelled the true location of radio-collared mule deer *Odoccoileus hemionus*, while Jonsen et al. (2006) did the same for individual tagged leatherback turtles. King (2014) describes how various types of ecological data can be modelled using a state-space formulation.

The biological processes explicitly modelled by the state model include mortality, birth or recruitment, and movement. Models for mortality have distinguished natural mortality for different life history stages (e.g. age 2, 3 and 4 Chinook salmon, Newman and Lindley 2006) and harvest-related mortality (e.g. black bears, Conn et al. 2008). Models for movement have included the movement of single animals (e.g. turtles, Jonsen et al. 2006), movement between areas of members of a single population (e.g. coho salmon, Newman 1998), and movement between metapopulations (e.g. four sets of pupping colonies used by grey seals, Thomas et al. 2005).

The remainder of this chapter includes a general statistical formulation of SSMs and simple examples. General approaches to inference, i.e. the fitting of SSMs to data, are discussed in Chap. 4, while Chap. 5 is a discussion of issues in SSM formulation and diagnostics. The remaining chapters show how to put specific problems, such as survival estimation (Chap. 7), into an SSM framework and include more detailed examples.

3.1 State-Space Models

Here we give a more formal and mathematically general definition of state-space models. State-space models are models for two discrete time processes running in parallel, one called the state process and the other the observation process. The state process is modelled by a conditional probability density function (pdf) or probability mass function (pmf) that describes the change of the state vector from time $t-1$ to time t, and is denoted by g_t. As will become clear in some of the later examples, g_t can be quite complicated and analytically intractable, reflecting a sequence of stochastic sub-processes. With some abuse of terminology, we will refer to pdf's and pmf's simply as pdf's. Note that we allow only discrete time indices, and to simplify notation, we assume that these are evenly spaced.

The value of the state vector at a given point in time is seldom known with certainty. Additionally, the parameters, such as survival rates and birth rates, are almost never known with certainty. If inferences about the state model are to be possible and defensible, field data must be collected. For example, various components of the state vector might be estimated from mark-recapture or line transect surveys. This leads to the observation process, which is modelled by another conditional pdf, denoted f_t, that describes the relationship between the state vector, $\mathbf{n}_t$, and a vector of observations, $\mathbf{y}_t$. As noted previously, point estimates of state vector components are typically quantities derived from sample data. We might instead use the raw sample data as observations. In contrast with the matrix model formulation, the general formulation presented here, which does not require a linear relationship between states and observations, can make modelling the raw observations more feasible.

The combination of the state model and the observation model is a state-space model and can be mathematically described as follows:

$$\text{Initial state pdf}: g_0(\mathbf{n}_0|\boldsymbol{\theta}) \tag{3.3}$$

$$\text{State } t \text{ pdf}: g_t(\mathbf{n}_t|\mathbf{n}_{t-1},\boldsymbol{\theta}) \tag{3.4}$$

$$\text{Observation } t \text{ pdf}: f_t(\mathbf{y}_t|\mathbf{n}_t,\boldsymbol{\psi}), \tag{3.5}$$

where $\boldsymbol{\theta}$ is a vector of parameters corresponding to the state model, $\boldsymbol{\psi}$ is a vector of parameters corresponding to the observation model, and t=1,..., T. A sequence of state vectors, $\mathbf{n}_a, \mathbf{n}_{a+1}, \ldots, \mathbf{n}_b$, will be denoted $\mathbf{n}_{a:b}$; $\mathbf{y}_{a:b}$ has a similar meaning.

The state process model, Eq. (3.4), is first-order Markov, i.e. the present state only depends on the previous state. Higher order Markov models can be re-written as first order Markov models by extending the dimension of the state vector to include previous states (see, for example, Schnute 1994). However, simply writing a univariate pdf for $\mathbf{n}_t$ as a function of previous states, e.g. $g_t(\mathbf{n}_t|\mathbf{n}_{t-1}, \mathbf{n}_{t-2}, \theta)$, does not pose any inference difficulties with modern approaches such as MCMC and sequential Monte Carlo, e.g., sequential importance sampling (see Chap. 4). The general class of such higher order Markov models was termed hidden process models by Newman et al. (2006).

Environmental, or temporal, variation in the state process can be made explicit by adding another level to the SSM for variability in survival, birth and other parameters.

$$\text{Parameter pdf} : h(\theta_t|\Gamma) \tag{3.6}$$

$$\text{Initial state pdf} : g_0(\mathbf{n}_0|\theta_t) \tag{3.7}$$

$$\text{State } t \text{ pdf} : g_t(\mathbf{n}_t|\mathbf{n}_{t-1}, \theta_t) \tag{3.8}$$

$$\text{Observation } t \text{ pdf} : f_t(\mathbf{y}_t|\mathbf{n}_t, \psi). \tag{3.9}$$

Such a model is an example of a random effects or hierarchical state-space model (see Sect. 2.2.2). The pdf $h(\theta_t|\Gamma)$ describes environmental stochasticity in θ_t, where Γ is a hyperparameter. Time-varying covariates could be used to model Γ.

For Bayesian inference, another level is added to the state-space model formulation, namely, the prior pdf for the fixed and unknown parameters. For example, referring to the above hierarchical state-space model, Eqs. (3.6)–(3.9),

$$\text{Prior pdf} : \pi(\Gamma, \psi) \tag{3.10}$$

$$\text{Parameter pdf} : h(\theta_t|\Gamma) \tag{3.11}$$

$$\text{Initial state pdf} : g_0(\mathbf{n}_0|\theta_t) \tag{3.12}$$

$$\text{State } t \text{ pdf} : g_t(\mathbf{n}_t|\mathbf{n}_{t-1}, \theta_t) \tag{3.13}$$

$$\text{Observation } t \text{ pdf} : f_t(\mathbf{y}_t|\mathbf{n}_t, \psi). \tag{3.14}$$

As will be described in more detail in Chap. 4, the end result of Bayesian inference for a state-space model is the joint posterior distribution for the parameters and the unknown states, i.e. $\pi(\mathbf{n}_{0:T}, \Gamma, \psi|\mathbf{y}_{1:T})$.

Finally, yet another layer of uncertainty is model uncertainty, also known as structural uncertainty (Williams et al. 2001). Alternative formulations for any of the above pdf's are often postulated. Competing theories about the science underlying the population dynamics translate into different state pdf's, g_t, or parameter pdf's, h. Denoting a particular model by $\mathscr{M}$, the Bayesian hierarchical model of Eqs. (3.10)–(3.14) can be extended as follows.

$$\text{Model prior pdf} : r(\mathscr{M}) \tag{3.15}$$

$$\text{Parameter prior pdf} : \pi(\Gamma, \psi|\mathscr{M}) \tag{3.16}$$

$$\text{Parameter pdf} : h(\theta_t|\Gamma, \mathscr{M}) \tag{3.17}$$

$$\text{Initial state pdf} : g_0(\mathbf{n}_0|\theta_t, \mathscr{M}) \tag{3.18}$$

$$\text{State } t \text{ pdf} : g_t(\mathbf{n}_t|\mathbf{n}_{t-1}, \theta_t, \mathscr{M}) \tag{3.19}$$

$$\text{Observation } t \text{ pdf} : f_t(\mathbf{y}_t|\mathbf{n}_t, \psi, \mathscr{M}). \tag{3.20}$$

The pdf's and associated parameters in Eqs. (3.16)–(3.20) are conditional on the model $\mathscr{M}$, where parameters and the pdf's can vary between models. Chapter 5 will discuss model selection and model averaging.

3.2 Examples of State-Space Models

Here we give four examples of SSMs with the state pdf g_t and observation pdf f_t fully specified. The first is a nonlinear and non-normal coho salmon SSM with a scalar state and a scalar observation variable. The other examples are stochastic extensions of examples from Chap. 2.

3.2.1 Simplified Salmon Example

In this simplified SSM,[1] the state variable is a scalar N_t, the number of juvenile salmon alive in year t (at some point in time in that year) in a particular river. The dynamics of N_t are modelled as a Poisson distribution version of a Ricker stock-recruitment model (Quinn and Deriso 1999).

$$N_t|N_{t-1} \sim \text{Poisson}\left(\alpha N_{t-1} e^{-\beta N_{t-1}}\right), \quad \alpha > 0, \beta > 0. \tag{3.21}$$

Implicitly, the parameter α includes survival between time $t-1$ and t, the proportion that are female, fecundity (number of eggs produced), and survival between egg deposition and juvenile life stage. For the deterministic version of the model at least, to avoid chaotic behaviour, the value of α needs to be less than 2.69, and to avoid cycling, α must be less than 2. The parameter β is a measure of density dependence: as β increases, density dependence increases. The equilibrium value is $\ln(\alpha)/\beta$.

[1] The state process in this salmon SSM is a considerable oversimplification of the population dynamics for most, maybe all, species of salmon. Typically juvenile salmon production in a given year is the result of spawning from two or more age classes (different cohorts) and those age classes were juveniles two or more years previously, so that $N_{t-2}, N_{t-3}, \ldots$ contribute to N_t.

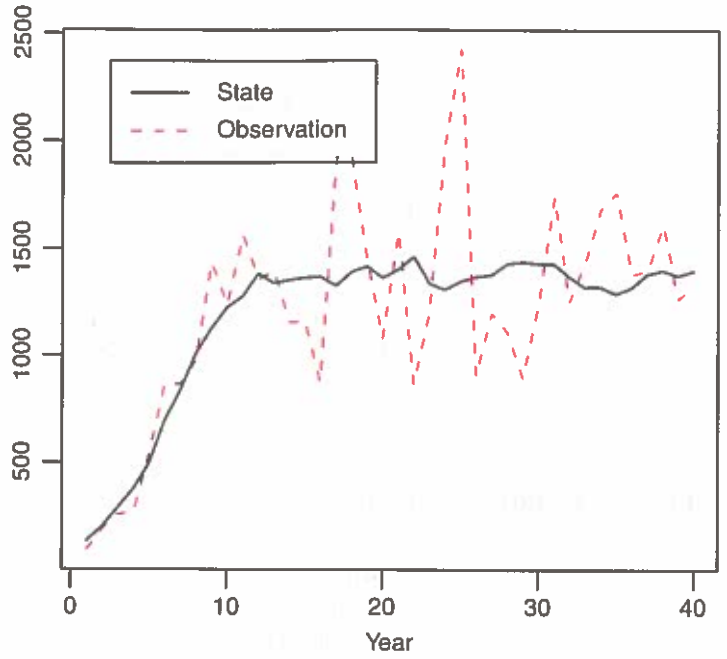


Fig. 3.2 Elementary salmon SSM with Ricker population dynamics with Poisson variation and lognormal observations. Ricker parameters are $\alpha = 1.5$ and $\beta = 0.0003$. Lognormal observations are bias-corrected with a coefficient of variation of 30 %

The observations are estimates of juvenile abundance, e.g. based upon samples take from in-river traps. A convenient probability distribution for non-negative valued observations is the lognormal distribution, although for counts of animals, continuous random variables are clearly an approximation:

$$y_t|N_t \sim \text{lognormal}\left(\log(N_t) - \sigma_y^2/2, \sigma_y^2\right). \tag{3.22}$$

This model is a bias-corrected lognormal, i.e. $E[y_t|N_t] = N_t$.

Forty years of simulated population dynamics and estimates are shown in Fig. 3.2. The estimates have a coefficient of variation of 30 %. (The code that generates this plot is given on the book website, see Sect. 1.2.)

3.2.2 *BRS Model*

For a more complex example, we return to the BRS formulation, an example of two states (e.g. immature and mature animals), summarized in Sect. 2.3. The sequence of sub-processes was survival (S), growth (R) and birth (B). Now we define stochastic processes for each of these sub-processes. The pdf g_t is difficult to

evaluate analytically because it is a function of three different pdf's, but it is easy to describe symbolically and easy to simulate from. Symbolically, we write g_t as the following composite function (with the parameter vector θ omitted to reduce notation):

$$g_t(\mathbf{n}_t|\mathbf{n}_{t-1}) = g_{3,t}\left(g_{2,t}\left(g_{1,t}(\mathbf{n}_{t-1})\right)\right). \tag{3.23}$$

The three pdf's, $g_{1,t}$, $g_{2,t}$, and $g_{3,t}$, represent the processes of survival, growth and birth. Statistically, this is a more useful formulation of the BRS model than is Eq. (2.11), because the latter tells us only the expected values of the states in $\mathbf{n}_t$, conditional on $\mathbf{n}_{t-1}$, whereas Eq. (3.23) represents the full joint distribution of the states in $\mathbf{n}_t$, conditional on $\mathbf{n}_{t-1}$.

The pdf $g_{1,t}(\mathbf{u}_{1,t}|\mathbf{n}_{t-1})$ corresponding to survival is the result of two binomial processes [Eq. (2.2)]:

$$\begin{pmatrix} u_{1(s),1,t} \sim \text{binomial}\left(n_{1,t-1}, \phi_1\right) \\ u_{1(s),2,t} \sim \text{binomial}\left(n_{2,t-1}, \phi_2\right) \end{pmatrix}. \tag{3.24}$$

Growth from immature to mature is another binomial process, so that $g_{2,t}(\mathbf{u}_{2,t}|\mathbf{u}_{1,t})$ is determined from Eq. (2.5):

$$\begin{pmatrix} u_{2(r),1,t} \sim \text{binomial}\left(u_{1(s),1,t}, 1-\pi\right) \\ u_{2(r),2,t} = u_{1(s),2,t} + \left(u_{1(s),1,t} - u_{2(r),1,t}\right) \end{pmatrix}. \tag{3.25}$$

If each adult can have at most one young, birth can be modelled as a third binomial process, so that $g_{3,t}(\mathbf{n}_t|\mathbf{u}_{2,t})$ is determined from a slight reformulation of Eq. (2.6):

$$\begin{pmatrix} n_{1,t} \equiv u_{3(b),1,t} = u_{2(r),1,t} + b_t \\ \quad \text{where} \quad b_t \sim \text{binomial}(u_{2(r),2,t}, \rho) \\ n_{2,t} \equiv u_{3(b),2,t} = u_{2(r),2,t} \end{pmatrix}. \tag{3.26}$$

In contrast to evaluation of the state pdf $\mathbf{g}_t$, simulation from the distribution is easy: one simply simulates from g_1, that output is then input to simulate from g_2, and that output is then input to simulate from g_3. Suppose $\phi_1 = 0.50$, $\phi_2 = 0.71$, $\pi = 0.60$, and $\rho = 0.80$ and let the initial numbers, $\mathbf{n}_0$, be (50,70). A simulation of the population dynamics for the two size classes for 30 years is shown in Fig. 3.3. The eventual population growth in this example is exponential and density dependence is needed in survival or birth processes to stabilize the population. The observations $y_{I,t}$ (estimated number of immature animals) and $y_{M,t}$ (estimated number of mature animals) were taken to be lognormally distributed, unbiased, with a coefficient of variation of 30 %, i.e.

$$y_{I,t} \sim \text{lognormal}\left(\ln(n_{1,t}) - \sigma_y^2/2, \sigma_y^2\right)$$

$$y_{M,t} \sim \text{lognormal}\left(\ln(n_{2,t}) - \sigma_y^2/2, \sigma_y^2\right)$$

where $\sigma_y^2 = \ln(0.3^2 + 1)$.

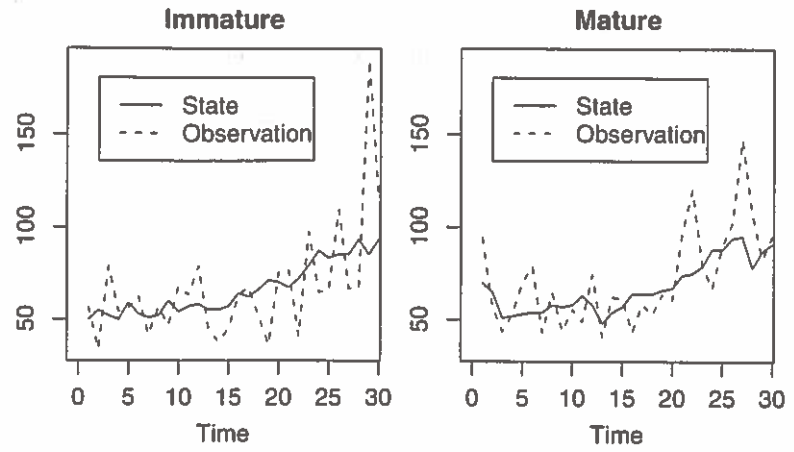


Fig. 3.3 Simulation of state process with two states, immature (I) and mature (M) animals, for $T = 30$ years, with three sub-processes, survival, growth and birth. Lognormally distributed estimates (the observations), with a coefficient of variation of 30 %, are also plotted

3.2.3 *Coho Salmon*

The processes survival, movement and harvest, that characterize the spatially partitioned coho salmon recoveries model described in Sect. 2.7.1, could be modelled using binomial (for survival and harvest) and multinomial (for movement) distributions. Here we describe a normal dynamic linear model (NDLM) approximation to those processes (Newman 1998):

$$\mathbf{n}_t|\mathbf{n}_{t-1} \sim \text{multivariate normal}\left(\mathbf{M}_t\mathbf{S}_{t-1}\mathbf{n}_{t-1}, \boldsymbol{\Sigma}_{n_t}\right), \quad t = 1, \ldots, 16 \quad (3.27)$$

$$\mathbf{y}_t|\mathbf{n}_t \sim \text{multivariate normal}\left(\mathbf{H}_t\mathbf{n}_t, \boldsymbol{\Sigma}_{y_t}\right), \quad t = 1, \ldots, 16. \quad (3.28)$$

The expected values are identical to the deterministic matrix models of Eqs. (2.18) and (2.22). The covariance matrix for the observations, $\boldsymbol{\Sigma}_{y,t}$, is a diagonal matrix with components $\sigma^2_{y,a,t}$, where

$$\sigma^2_{y,a,t} = n_{a,t}h_{a,t}(1 - h_{a,t}), \quad (3.29)$$

which is the variance for a binomial($n_{a,t}$, $h_{a,t}$) random variable. The components of the covariance matrix for the states, $\boldsymbol{\Sigma}_{n_t}$, can be constructed similarly using the variances and covariances of binomial random variables (for survival) and multinomial random variables (for movement); Newman (1998) gives a detailed example of the construction. Sullivan (1992) used a similar approach to constructing the covariance matrix in an NDLM approximation to binomial state processes.

3.2.4 *Deer Metapopulation*

The deer metapopulation model (Sect. 2.7.2) had five sub-processes characterizing the dynamics of the two deer population abundances. For convenience the expected state vector model is shown again below.

$$E[\mathbf{n}_t|\mathbf{n}_{t-1}] = \mathbf{CBAM}_t\mathbf{S}_{t-1}\mathbf{n}_{t-1}$$

Letting $\mathbf{n}'$ denote the transpose of the column vector $\mathbf{n}$,

$$\mathbf{n}'_t = [n_{A,f,1,t}, n_{A,m,1,t}, n_{A,f,2,t}, n_{A,f,3+,t}, n_{A,m,2+,t}, n_{B,f,1,t}, n_{B,m,1,t}, n_{B,f,2,t}, n_{B,f,t}, n_{B,m,t}],$$

abundances distinguished by population (A or B), sex (f or m), and age class (1, 2 or 3+ for females, and 1 or 2+ for males). The matrices correspond to the processes of survival ($\mathbf{S}_{t-1}$), movement between populations A and B ($\mathbf{M}_t$), age incrementation (**A**), births (**B**), and assignment of sex (**C**). Age incrementation is treated as a deterministic process, but the remaining processes can be modelled stochastically. Similar to the *BRS* model above, sub-process pdfs are specified, say g_S, g_M, g_A (a deterministic function), g_B and g_C. Consequently, the pdf $g_t(\mathbf{n}_t|\mathbf{n}_{t-1})$ is a composite function, analytically intractable, but potentially easy to simulate from using the individual component pdf's. Relatedly, and of relevance to issues of model fitting addressed in Chap. 4, analytic evaluation of probabilities of states is only tractable by separately evaluating the pdf's for intermediate or "latent" states, the "u's".

Here we describe one construction of $g_t(\mathbf{n}_t|\mathbf{n}_{t-1})$. Beginning with survival, survivors for each of the ten components of $\mathbf{n}_t$ are independent binomial random variables with various survival probabilities [e.g. Eq. (2.29)]. Given the cth component of $\mathbf{n}_{t-1}$, the survivors $u_{1(s),c,t}$ have the following distribution:

$$u_{1(s),c,t}|n_{c,t-1} \sim \text{binomial}\left(n_{c,t-1}, \phi_{c,t-1}\right). \quad (3.30)$$

The survivors in each component then move from their current population to the other population with probabilities specified previously; e.g. $\mu_{A\to B,t}$ specifies the probability that an animal moves from population A to population B, Eq. (2.30). The number moving is modelled as a binomial random variable; e.g. the movement of first year females from A to B,

$$u_{2(m),A\to B,f,1,t}|u_{1(s),A,f,1,t} \sim \text{binomial}\left(u_{1(s),A,f,1,t}, \mu_{A\to B,t}\right). \quad (3.31)$$

The new number in a given population (of sex s and age class a) is the sum of two independent binomial random variables, the "stayers" and the "movers", e.g. the first year females in B are those staying in B and those moving from A to B,

$$u_{2(m),B,f,1,t} = u_{2(m),A\to B,f,1,t} + u_{2(m),B\to B,f,1,t}. \quad (3.32)$$

As explained in Sect. 2.7.2, within each population, age incrementation shifts surviving first-year females to the second-year females group, surviving second-year and older females are combined (and labelled older females), and all surviving first-year and older males are combined (as older males). The fourth sub-process, birth, is modelled as a binomial process, with different success probabilities for second-year females and for older females, Eq. (2.32); e.g. the total births, males and females (total denoted $\cdot$) in population A, are the sum of births from the two sets of fecund females,

$$u_{4(b),A,\cdot,1,t} = u_{4(b),A,\cdot,1(2),t} + u_{4(b),A,\cdot,1(3+),t} \tag{3.33}$$

where

$$u_{4(b),A,\cdot,1(2),t} | u_{3(a),A,f,2,t} \sim \text{binomial}\left(u_{3(a),A,f,2,t}, \rho_1\right) \tag{3.34}$$

$$u_{4(b),A,\cdot,1(3+),t} | u_{3(a),A,f,3+,t} \sim \text{binomial}\left(u_{3(a),A,f,3+,t}, \rho_2\right). \tag{3.35}$$

The final sub-process, sex assignment, is treated as another binary (Bernoulli) process, where the number of females in the newly born animals is binomial; e.g. the number of first-year females in population A,

$$n_{A,f,1,t} | u_{4(b),A,\cdot,1,t} \sim \text{binomial}\left(u_{4(b),A,\cdot,1,t}, \alpha\right). \tag{3.36}$$

The observations were defined in Sect. 2.7.2 to be estimates of the total number of deer in each of the two populations, with no distinction between sex and age. Again lognormal distributions (with bias correction) are used; assuming independence between the population estimates,

$$y_{A,t} | \mathbf{n}_{A,\cdot,\cdot,t} \sim \text{lognormal}\left(\ln(\mathbf{n}_{A,\cdot,\cdot,t}) - \sigma_y^2/2, \sigma_y^2\right) \tag{3.37}$$

$$y_{B,t} | \mathbf{n}_{B,\cdot,\cdot,t} \sim \text{lognormal}\left(\ln(\mathbf{n}_{B,\cdot,\cdot,t}) - \sigma_y^2/2, \sigma_y^2\right) \tag{3.38}$$

where $\mathbf{n}_{A,\cdot,\cdot,t} = n_{A,f,1,t} + n_{A,m,1,t} + n_{A,f,2,t} + n_{A,f,3+,t} + n_{A,m,2+,t}$, and likewise for $\mathbf{n}_{B,\cdot,\cdot,t}$.

Chapter 4
Fitting State-Space Models

4.1 Introduction

This chapter is an overview of methods for using available data to make inferences about states and parameters of a state-space model. We call this "model fitting", or as Hilborn and Mangel (1997) say, "confronting models with data". Given a general SSM [Eqs. (3.3)–(3.5)],

$$\text{Initial state pdf}: g_0(\mathbf{n}_0|\theta)$$

$$\text{State } t \text{ pdf}: g_t(\mathbf{n}_t|\mathbf{n}_{t-1}, \theta)$$

$$\text{Observation } t \text{ pdf}: f_t(\mathbf{y}_t|\mathbf{n}_t, \psi),$$

model fitting is a matter of using the data, $\mathbf{y}_{1:T}$, to estimate the unknown parameters, (θ, ψ), or the unknown states, $\mathbf{n}_{0:T}$, or both—the dual estimation problem (Wan and Nelson 2001). At this point, we assume that the model structure has been specified; we later address model uncertainty. As noted previously, the definition of data, $\mathbf{y}_{1:T}$, is broad enough to include functions of sample data, e.g. point estimates of abundance and associated standard errors, as well as the raw data. The word "prediction", instead of estimation, is sometimes used when the unknown is a random variable instead of a constant; e.g. we predict the states $\mathbf{n}_{0:T}$ and estimate the parameters (θ, ψ).

The literature on fitting state-space models is extensive, from the foundational work of Kalman (1960) and Kalman and Bucy (1961) to the present day, where research activity remains high. Means of fitting SSMs have advanced considerably in recent years, in particular using computer-intensive procedures. Taken together with advances in collecting data, we can now fit increasingly complex and realistic SSMs for population dynamics. This chapter aims to provide enough detail on these recent developments to allow the statistical ecologist to understand the pros and cons of the different fitting procedures. Additional details on specific SSM fitting algorithms and on available software are available in Commandeur et al. (2011).

K.B. Newman et al., *Modelling Population Dynamics: Model Formulation, Fitting and Assessment using State-Space Methods*, Methods in Statistical Ecology,
DOI 10.1007/978-1-4939-0977-3_4, © Springer Science+Business Media New York 2014

As a guide to readers with different interests, the organisation of the remainder of this chapter is as follows. Section 4.2 looks at the fitting, or more generally the inference, objectives from the perspective of an ecologist or natural resource manager, where we argue that attention needs to be focused on the state process model in ways similar to those presented for matrix models by Caswell (2001) and for population viability analysis by Morris and Doak (2002). Section 4.3 looks at fitting from a more purely statistical perspective (as given above), namely estimation of states and parameters conditional on the data. We discuss the central role of mathematical integration in inference for SSMs and highlight Bayesian and classical inference approaches to fitting SSMs. Section 4.4 describes "classical" statistical procedures for fitting SSMs, in particular the Kalman algorithm as it is applied to normal dynamic linear models, and the extended Kalman filter for nonlinear and non-normal SSMs. Section 4.5 discusses more recent stochastic algorithms for making inferences about SSMs, such as Markov chain Monte Carlo and Sequential Monte Carlo. Such stochastic procedures are more commonly used in a Bayesian framework but classical statistical inference can be carried out using these methods as well (e.g. de Valpine 2004).

4.2 An Ecological Perspective on Inference for SSMs

The primary SSM inferential objectives from an ecologist's or wildlife biologist's perspective are guided by the underlying scientific and management objectives. Different objectives can lead to emphases on estimating past states, predicting future states, estimating parameters related to vital rates such as survival, growth and reproduction, or quantifying the effects of environmental conditions or management actions on vital rates. Below we discuss a possible (and desirable) progression in inferential goals.

4.2.1 *Knowing the Past and Present States*

Often for a particular species or population, emphasis begins with the states, i.e. the initial questions about a population pertain to its abundance, its "status", and how those abundances have changed over time. As a Google search on the key word phrase "Population status and trends" reveals, status and trend reports for a large number of biological populations are ubiquitous products of natural resource management agencies. Before anything can be done, or should be done, about a population of interest, one needs to have some idea of the abundance of the population. That abundance might be further partitioned into different biological or geographical categories if the abundances of a structured population are of interest; e.g. number of age 1 females in different regions. For apparently healthy populations that are harvested for commercial, subsistence, or recreational purposes,

estimates of current abundance, $\mathbf{n}_T$ say, are used to set harvest limits, quotas or regulations. For species that have been designated as threatened or endangered by some organization or agency such as the International Union for Conservation of Nature (IUCN) or the United States Fish and Wildlife Service, according to some criterion or law such as the U.S. Endangered Species Act, estimates of current and past abundances, $\mathbf{n}_{1:T}$, are used to determine if there are signs of population recovery, e.g. positive trends or positive step-changes in abundance.

4.2.2 *Explaining the Past and Present, and Predicting the Future*

Emphasis sometimes shifts, and we maintain that in many cases it should quickly do so, from estimating the states, or "status and trends", to estimating or quantitatively characterizing the underlying processes that drive the population dynamics. Even if one had perfect information about current and past states of the system, i.e. $\mathbf{n}_1, \ldots, \mathbf{n}_T$ were known without error, several questions would arise. Why did such a history occur? Why is the population trending downward toward extinction or quasi-extinction (Morris and Doak 2002)? Why did a positive step change occur? What will future states be? For natural resource managers, the most pressing question may be "What do I do now?", or "Given two alternate management actions, which one is better?". Predictions of future states and how different management actions affect those predictions can aid the decision-making process. The ideal model would take proposed actions A_j, $j = 1, \ldots, J$, and current system state, $\mathbf{n}_T$, as inputs and predict, without error, the next state $\mathbf{n}_{T+1}|A_j$. The manager could then compare $\mathbf{n}_{T+1}|A_j$, $j = 1, \ldots, J$, and select an action that was closest to the management objectives (with cost constraints also affecting the choice of action).

To understand the past and present and, similarly, to predict future system states, we need to know more than the system states: we need to understand the underlying dynamics that generated the states. Consider a simple univariate measure of the system, n_t, denoting the abundance of a single population. Given a sequence, $n_1, \ldots, n_T$, we can coarsely characterize the dynamics by $n_t = \lambda_{t-1} n_{t-1}$, where λ_{t-1} is the annual population growth rate. The sequence $\lambda_1, \ldots, \lambda_{T-1}$ can be quite informative in terms of characterizing the historic population growth rates. Estimates of such sequences are often the first step in carrying out a population viability analysis (PVA), "the use of quantitative methods to predict the likely future status of a population" (Morris and Doak 2002:1).

If we know the λ_ts, we would still like to know why they take those values. This requires knowing, or estimating, the underlying vital rates (survival probabilities and reproductive success, and possibly others) which are the basis of the λ_ts; e.g. what were the probabilities of annual survival for age 3 black bears for the last ten years? Thus estimation of underlying vital rates should be a next step; e.g. if $\lambda_t = (1 + \beta_t)\phi_t$, where β_t is the birth rate and ϕ_t is the survival probability, then estimation of β_t and ϕ_t is the next inferential objective.

Given estimates of vital rates, we might want to burrow down deeper to try to understand why the rates are what they are, to gain a more mechanistic understanding of the processes underlying the dynamics. In particular, we would like to understand the effects of both environmental covariates outside the direct control of humans, e.g. how winter temperatures affect bird survival (North and Morgan 1979), and deliberate management actions, e.g. crab fishery effects on Dungeness crab (Higgins et al. 1997). The quest for understanding the dynamics of a population is a continual one; the more information and understanding we gain, the more ideas will arise for managing or manipulating the population dynamics. For endangered populations in particular, there are always questions about what management actions yield the greatest improvement in population growth rates; e.g. increasing available habitat, improving the quality of current habitat, removing predators, increasing prey? We need not strive for the perfect model—if the mechanisms underlying vital rates can be explained to the degree that the effects of different management actions can be reliably predicted, this gives sufficient depth.

4.2.3 Remarks

- As SSMs become more widely used for modelling population dynamics, they will be used for similar purposes as are matrix population models (Chap. 2). Caswell (2001) and Morris and Doak (2002) demonstrate the use of matrix models to project population abundances into future years, to calculate probabilities of extinction (or quasi-extinction), to carry out sensitivity or elasticity type analyses for vital rates, and to assess the effects of environmental covariates or management actions on such vital rates and overall population growth rates.
- For ecologists and resource managers, the primary emphasis should be on the state process model, not on the observation model. While the type and quality of data are crucial to the complexity of an SSM that can be fitted, and while the observation model provides the essential link with the underlying state process model, data collection and the observation model simply provide imperfect snapshots of what the scientist and manager really need: the state vector. Even if the components of the state vector were known without error, the scientist still has substantial work to do to try to model the underlying dynamics. The science of population dynamics modelling is so challenging due to these dual problems: (1) collecting imperfect measurements on dynamic populations, e.g. using methods like mark-recapture or distance sampling or standard survey sampling procedures for fixed areas; and (2) trying to understand what causes the populations to vary as they do, e.g. whether there are density-dependent or environmental effects on survival probabilities or reproduction rates. It is important to keep a distinction in mind between the two problems, as we sometimes see field ecologists and biologists focused primarily on the data collection or population monitoring programmes and only secondarily on using those data to draw inferences about the underlying dynamics, and *vice versa* for theoretical ecologists.

- Simply postulating and mathematically formulating a state-space model for animal populations can be a useful exercise in its own right. That process alone can yield insights and greater clarity about the dynamics, the factors potentially affecting the dynamics, and which data are relevant. At the same time, it is relatively easy to formulate state process models in an SSM that are overly ambitious given the available data; e.g. a state process model for elk might project abundances by age and sex, but the only available data might be on total elk counts in September. In Chap. 5, we further discuss the merits of SSM formulation as well as the problem of data inadequacy, i.e. model overspecification. In this chapter, however, we assume that a "reasonable" SSM has been formulated in the sense that the parameters and states of the model are "estimable" given the available data, and we discuss different methods for making inferences about the SSM.

4.3 A Statistical Perspective on Inference for SSMs

As described at the beginning of this chapter, the general inferential objective is to characterize, or summarize, the unknown parameters of the SSM, $(\boldsymbol{\theta}, \boldsymbol{\psi})$, and the unknown states of the state process, $\mathbf{n}_{0:T}$, given the observations, $\mathbf{y}_{1:T}$. Sometimes the initial state, $\mathbf{n}_0$, is viewed as an unknown constant rather than the realisation of a stochastic process.

4.3.1 Inference as Integration

Inference for SSMs, whether for parameters, states, or both, ultimately involves integration, typically over a high-dimensional space (e.g. T integrals). To demonstrate this we begin with the *joint distribution of states and observations* of the most basic SSM in a non-Bayesian setting, namely Eqs. (3.3)–(3.5), which can be written as follows.

$$g_0(\mathbf{n}_0|\boldsymbol{\theta})g_{1:T}(\mathbf{n}_{1:T}, \mathbf{y}_{1:T}|\mathbf{n}_0, \boldsymbol{\theta}, \boldsymbol{\psi}) = g_0(\mathbf{n}_0|\boldsymbol{\theta})\prod_{t=1}^{T} g_t(\mathbf{n}_t|\mathbf{n}_{t-1}, \boldsymbol{\theta}) f_t(\mathbf{y}_t|\mathbf{n}_t, \boldsymbol{\psi}). \quad (4.1)$$

Inference about parameters via maximum likelihood methods is based on the *marginal distribution of the observations*, equivalently the likelihood, thus integrating out the states in the joint distribution, i.e.

$$L(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y}_{1:T}) \equiv f_{1:T}(\mathbf{y}_{1:T}|\boldsymbol{\theta}, \boldsymbol{\psi}) = \int g_0(\mathbf{n}_0|\boldsymbol{\theta})\prod_{t=1}^{T} g_t(\mathbf{n}_t|\mathbf{n}_{t-1}, \boldsymbol{\theta}) f_t(\mathbf{y}_t|\mathbf{n}_t, \boldsymbol{\psi})\, d\mathbf{n}_{0:T}. \quad (4.2)$$

Assuming the parameters are known constants, inferences about the states in the most general sense are based on the *conditional distribution of the states given the observations*, which also requires integration:

$$g_{0:T}(\mathbf{n}_{0:T}|\mathbf{y}_{1:T},\boldsymbol{\theta},\boldsymbol{\psi}) = \frac{g_0(\mathbf{n}_0|\boldsymbol{\theta})\prod_{t=1}^{T} g_t(\mathbf{n}_t|\mathbf{n}_{t-1},\boldsymbol{\theta})\, f_t(\mathbf{y}_t|\mathbf{n}_t,\boldsymbol{\psi})}{f_{1:T}(\mathbf{y}_{1:T}|\boldsymbol{\theta},\boldsymbol{\psi})}. \quad (4.3)$$

We address handling of unknown parameters in a non-Bayesian framework in Sect. 4.3.2.4. A single summary measure of a particular state $\mathbf{n}_t$, such as the expected value, involves integration of the above conditional distribution over the remaining states, i.e.

$$E[\mathbf{n}_t|\mathbf{y}_{1:T},\boldsymbol{\theta},\boldsymbol{\psi}] = \int \mathbf{n}_t\, g_{0:T}(\mathbf{n}_{0:T}|\mathbf{y}_{1:T},\boldsymbol{\theta},\boldsymbol{\psi})\, d\mathbf{n}_{0:T}.$$

Bayesian inference is based on a larger joint distribution, extending Eq. (4.1) by adding a prior distribution for the parameters, which also requires integration; this is further explained in Sect. 4.3.2.

The ease with which inference can be carried out is then a function of how difficult it is to carry out the above integrations. In turn, the degree of difficulty of integration is a function of the particular formulation of the state and observation pdf's. The generality of the SSM structure invites creativity in the formulation of the pdf's, including the choice of probability distributions (e.g. normal, lognormal, Poisson), the nature of the parameterizations (e.g. linear, nonlinear), the dimensionality of the state and observation vectors, and the length of the time series, T.

Practical implementation of SSMs, however, only goes as far as the available integration techniques. Several integration techniques are currently available and have been applied to SSMs. These methods include exact analytical (e.g. Kalman filter), approximate analytical (e.g. extended Kalman filter), numerical, and Monte Carlo simulation methods. This listing is roughly a historical progression of techniques and SSM complexity approximately parallels this progression, i.e. SSMs for which inferences can be made using exact analytical methods are simpler than SSMs which can be fitted with Monte Carlo simulation methods. Statistical software developed for fitting SSMs of varying distributional structure and levels of complexity has progressed considerably over the past decade. An entire issue of the *Journal of Statistical Software* (May 2011, Volume 11) was devoted to the topic of fitting SSMs and the introductory article by Commandeur et al. (2011) is recommended reading. Here we discuss some of these methods in relatively broad terms in Sects. 4.4.1 (exact and approximate analytical methods, particularly the Kalman filter), 4.4.3 (numerical procedures), and 4.5 (Monte Carlo simulation procedures). Before doing so, we discuss the distinction between Bayesian and classical statistical approaches as applied to fitting SSMs.

4.3.2 *Bayesian Versus Classical Inference for SSMs*

This is a brief overview of the distinction between Bayesian and classical statistical procedures. We also look at how the approaches differ in the context of SSMs. Additional details can be found in many excellent references (e.g. Gelman et al. 2003; Carlin and Louis 2009; King et al. 2009).

4.3.2.1 Bayes vs Classical

Underlying both Bayesian and classical statistical inference is the likelihood, i.e., the joint probability distribution for the observed data viewed as a function of the unknown parameters. As a simple example of a likelihood, consider two passes of removal sampling of trout in a small stream (see Chap. 5 of Borchers et al. 2002). A net is swept on two consecutive occasions through the stream section, and each trout is caught with probability p; the captures are independent within and between sweeps, and trout that are captured are not returned to the stream. The observations are the numbers caught in the first and second pass, denoted y_1 and y_2. The parameters are p and N, the total number of fish present. The probability distribution for y_1 is binomial(N, p) and the distribution for y_2, conditional on the number y_1 removed previously, is binomial$(N - y_1, p)$. Thus the joint distribution is the product of the two binomial distributions:

$$f(y_1, y_2|N, p) \equiv f(\mathbf{y}|\boldsymbol{\theta}) = \binom{N}{y_1} p^{y_1}(1-p)^{N-y_1} \times \binom{N-y_1}{y_2} p^{y_2}(1-p)^{N-y_1-y_2}. \quad (4.4)$$

Classical statistical inference for the unknown parameters is usually centred on the likelihood. The expression in Eq. (4.4) is regarded as a function of the parameters given the observations: $f(\mathbf{y}|\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}|\mathbf{y})$, where $\boldsymbol{\theta} = (N, p)$. Conceptually, the parameters are considered to have fixed but unknown values. The most common point estimate of a parameter is the maximum likelihood estimate (mle); the likelihood $L(\boldsymbol{\theta}|\mathbf{y})$ has its maximum value when the mles of the parameters $\boldsymbol{\theta} = (N, p)$ are substituted in. Referring to the removal sampling example, given particular values of y_1 and y_2, the mle's for N and p are those values that maximize Eq. (4.4). These values are

$$\hat{N} = \frac{y_1^2}{y_1 - y_2}$$

$$\hat{p} = \frac{y_1 - y_2}{y_1}.$$

Note that $\hat{N}$ is only useful when $y_1 > y_2$; when $y_1 = y_2$, $\hat{N}$ is undefined, and when $y_1 < y_2$, the calculated values of $\hat{N}$ and $\hat{p}$ are negative.

Bayesian inference differs in a fundamental sense. Before collecting or analysing data, we specify probability distributions that quantify our beliefs about the values of unknown parameters. Namely, we specify a prior distribution, $\pi(\theta)$, where the parameters θ are now considered to be random variables rather than having fixed but unknown values. Then given the data, we modify our beliefs to produce a posterior distribution for the parameters, $\pi(\theta|\mathbf{y})$, where $\mathbf{y}$ represents the data. The posterior distribution is simply a conditional probability distribution, and Bayes formula is used to describe this distribution:

$$\pi(\theta|\mathbf{y}) = \frac{\pi(\theta,\mathbf{y})}{f(\mathbf{y})} = \frac{f(\mathbf{y}|\theta)\pi(\theta)}{\int f(\mathbf{y}|\theta)\pi(\theta)d\theta} \propto f(\mathbf{y}|\theta)\pi(\theta). \tag{4.5}$$

Thus Bayesian inference is based on both the likelihood, $f(\mathbf{y}|\theta) \equiv L(\theta|\mathbf{y})$, and the prior distribution for the parameters, $\pi(\theta)$.

Returning to the removal sampling example, we first specify a prior distribution for the parameters N and p. For example,

$$N \sim \text{discrete uniform}(a,b)$$
$$p \sim \text{uniform}(0,1)$$

where a and b are positive integers such that $a < b$. The discrete uniform (a,b) distribution is such that for $x = a, a+1, \ldots, b$, $\Pr(N = x) = 1/(b-a+1)$. The posterior distribution for N and p is then:

$$\pi(N,p|y_1,y_2) \propto \frac{1}{b-a+1}\binom{N}{y_1}p^{y_1}(1-p)^{N-y_1} \times \binom{N-y_1}{y_2}p^{y_2}(1-p)^{N-y_1-y_2}$$

for $N = a, a+1, \ldots, b$ and $0 \le p \le 1$.

Bayesian methods have some advantages over classical methods. One is that previous knowledge, independent of the current data, can be explicitly combined with the new data. One perspective on Bayes theorem is that, given the current data, we update or revise our prior knowledge of a process or phenomenon. Further by constraining the support (the set of allowable values) via a prior distribution, we ensure that the corresponding posterior distribution does not have support for nonsensical parameter values. As an example of the latter problem, for the removal sample example, maximum likelihood estimates are inadmissible when $y_1 \le y_2$. Adopting a Bayesian approach, we can select a prior distribution for N that constrains it to lie in the interval (a,b), thus ensuring that posterior estimates (both point and interval) remain in the permissible range.

A further advantage of Bayesian methods is that the posterior distribution is necessarily a more complete description of information about a specific parameter than classical point estimates (e.g. maximum likelihood estimates) and confidence intervals.

A potential disadvantage of Bayesian methods, particularly in the case of relatively small samples, is that the influence of the prior distribution on the posterior distribution can be sizable. As is sometimes said, "the prior is swamping the data". Individuals with vastly different prior distributions might end up with vastly different posterior distributions and arrive at conclusions that differ in important ways. Selection of non-informative or neutral priors has been a topic of considerable discussion and research (Gelman et al. 2003).

4.3.2.2 Hierarchical Models

The terminology used to describe both hierarchical models and Bayesian and classical inference for such models can be confusing. We devote time to this issue because SSMs are a special case of hierarchical models and classical and Bayesian inference procedures for SSMs are discussed next. To try to clear up potential confusion, we consider the following simple model.

$$\alpha_i \sim \text{normal}\left(\psi, \tau^2\right), \; i = 1, \ldots, p \tag{4.6}$$

$$y_{i,j}|\alpha_i \sim \text{normal}\left(\alpha_i, \sigma^2\right), \; j = 1, \ldots, q_i. \tag{4.7}$$

To make the example more concrete, suppose there are p lakes which were sampled for a species of fish one summer, and in lake i, q_i fish were caught and measured for lengths. The observation, $y_{i,j}$, is the length of the jth fish caught in the ith lake. The average lengths are allowed to differ between lakes.

Such models have been given several different names, including hierarchical, multi-level, random effects, and variance components models. The equivalence between hierarchical and multi-level is likely apparent. The model for the observations, $y_{i,j}$, can be viewed as a lower level model, while the model for the lake-specific means, α_i, is a higher level model. The random effects label refers to the random variable α_i which then becomes a parameter in the model for the observations, $y_{i,j}$. The label variance components refers to the terms τ^2 and σ^2.

Inference objectives include estimation of the fixed parameters, ψ, τ^2, and σ^2, and the random variables, α_i, i=1, …, n. A common classical inference approach to estimating the fixed parameters is maximum likelihood. The likelihood is evaluated by integrating over the unobserved α_i,

$$L(\psi,\tau^2,\sigma^2) = \prod_{i=1}^{p}\int_{\alpha_i}\left[\prod_{j=1}^{q_i}\Phi\left(\frac{y_{i,j}-\alpha_i}{\sigma}\right)\right]\Phi\left(\frac{\alpha_i-\psi}{\tau}\right)d\alpha_i$$

where $\Phi()$ denotes the standard normal density function. We note that maximum likelihood estimates of the variance components, τ^2 and σ^2, are typically biased, and REML (restricted maximum likelihood) (Harville 1977) is an alternative procedure which yields less biased or unbiased estimates.

Classical inference estimates of the random variable, α_i, can be made conditional on the estimated parameters and the observations, namely,

$$\hat{\alpha}_i = E\left[\alpha_i | \mathbf{y}, \hat{\psi}, \hat{\tau}^2, \hat{\sigma}^2\right] = \left[\frac{\hat{\sigma}^2/q_i}{\hat{\sigma}^2/q_i + \hat{\tau}^2}\right]\hat{\psi} + \left[\frac{\hat{\tau}^2}{\hat{\sigma}^2/q_i + \hat{\tau}^2}\right]\overline{y}_i. \quad (4.8)$$

Such estimates are called Empirical Bayes estimates (Casella 1985). This terminology, *Empirical Bayes*, may seem somewhat confusing in that a prior distribution was not explicitly defined in the original hierarchical model formulation. However, an alternative perspective on the formulation is that the model of Eq. (4.6) is a Bayesian prior distribution for the α_i (Casella 1985), where the values ψ and τ^2 were fixed in advance, arbitrarily. Purely Bayesian point estimates for α_i are the means of the posterior distribution, which equal Eq. (4.8) with the fixed values substituted (assuming for simplicity here that σ^2 was known).

We note that an alternative Bayesian inference procedure is to not assume known values for ψ and τ^2 and to specify a prior distribution, say $\pi(\psi, \tau^2)$ (again assume σ^2 is known, though a prior distribution could be chosen for it). Then the posterior distribution for ψ and τ^2 would be calculated,

$$\pi(\psi, \tau^2|\mathbf{y}) \propto \pi(\psi, \tau^2) \prod_{i=1}^{p} \int_{\alpha_i} \left[\prod_{j=1}^{q_i} \Phi\left(\frac{y_{i,j} - \alpha_i}{\sigma}\right)\right] \Phi\left(\frac{\alpha_i - \psi}{\tau}\right) d\alpha_i.$$

All information about the α_i would then be found in the conditional distribution for α_i given $\mathbf{y}$ and the prior distribution $\pi(\psi, \tau^2)$. Such inferences are called, somewhat confusingly again, Bayes empirical Bayes (Carlin and Louis 2009).

Finally, Clark (2005) makes a distinction between the meaning of randomness as it applies to the α_i and to ψ and τ^2. The α_i are inherently random variables, the values are varying in some temporal or spatial sense, e.g., the average lengths are varying between lakes. The probability distribution for α_i quantifies that inherent randomness. Conversely ψ and τ^2 are fixed constants but their values are unknown. A prior distribution for these fixed constants is a reflection of *uncertainty*, ignorance, say, not a reflection of temporal or spatial variability in their values. Bayesian statisticians may or may not make such a distinction and simply refer to the α_i and ψ and τ^2 as random variables.

4.3.2.3 Bayesian Inference for SSMs

In the case of state-space models and the dual estimation objective, Bayesian inference involves calculating, or generating a sample from, the posterior distribution

for the parameters and the states. Prior distributions are needed for the parameters, say $\pi(\theta, \psi)$. Unless the initial state value, $\mathbf{n}_0$, is assumed known, then a prior distribution (the initial state distribution) must also be specified. Priors for the unknown, random states, $\mathbf{n}_{1:T}$, need not be specified because the state process model itself implicitly defines their prior distribution. The Bayes formula for a state-space model can be written as follows.

$$\begin{aligned}\pi(\mathbf{n}_{0:T}, \theta, \psi|\mathbf{y}_{1:T}) &\propto \pi(\mathbf{n}_{0:T}, \theta, \psi, \mathbf{y}_{1:T}) \\ &= \pi(\theta, \psi) g_0(\mathbf{n}_0|\theta) \prod_{t=1}^{T} g_t(\mathbf{n}_t|\mathbf{n}_{t-1}, \theta) f_t(\mathbf{y}_t|\mathbf{n}_t, \psi), \quad (4.9)\end{aligned}$$

where the SSM is assumed first-order Markov.

Analytic calculation of the posterior distribution $\pi(\mathbf{n}_{0:T}, \theta, \psi|\mathbf{y}_{1:T})$ is prohibitively difficult, usually impossible, for most SSMs. The difficulty lies in the denominator of Bayes formula [Eq. (4.5)], $f(\mathbf{y})$, which is typically an intractable integral. Numerical integration methods can be used for small (say two or three) dimensional problems, but most of these methods fail in higher dimensions.

Bayesian inference has been made feasible for high-dimensional models, including SSMs, through the combination of Monte Carlo or simulation-based algorithms and high-speed computing. These computer-intensive Monte Carlo procedures can be used to produce samples from the posterior distribution. The best known and most widely used Monte Carlo sampling procedure is Markov chain Monte Carlo (MCMC). `WinBUGS` is a freely available software implementation of MCMC and is the tool we have chosen for demonstrating Bayesian inferences in this book. The open-source version of `WinBUGS` is `OpenBUGS`, and another free option is the `JAGS` software. Alternative Monte Carlo procedures designed primarily for dynamic stochastic models, including SSMs, are Sequential Monte Carlo methods with a specific approach being Sequential Importance Sampling (SIS).

4.3.2.4 Classical Inference for SSMs

In Sect. 4.4, we provide specifics on various classical statistical inference procedures for SSMs, in particular the highly influential Kalman filter algorithm. Here discussion is limited to brief general principles.

The classical approach differs from the Bayesian approach in that there are no prior distributions for the unknown parameters. Consider the following simple SSM where n_t and y_t are scalars, and where parameters from the observation model are removed to reduce notation,

$$\begin{aligned} g_t(n_t|n_{t-1}, \theta) \\ f_t(y_t|n_t). \end{aligned}$$

Inference often proceeds in two stages: first, the unknown fixed parameters ($\boldsymbol{\theta}$) are estimated, then estimates (sometimes called predictions) of the unknown states are made conditional on the fixed parameter estimates (implicitly, on the observations). Assuming that estimation of the parameters is via maximum likelihood, calculation of the marginal distribution for the observations is required, i.e. integration over the unknown states is carried out [Eq. (4.2)]. Treating these parameter estimates as constants (i.e. ignoring parameter uncertainty), the conditional distribution for the states [Eq. (4.3)] is a complete summary of information about the unknown states.

$$g_{0:T}(\mathbf{n}_{0:T}|\mathbf{y}_{1:T},\widehat{\boldsymbol{\theta}}) = \frac{g_0(\mathbf{n}_0|\widehat{\boldsymbol{\theta}})\prod_{t=1}^{T} g_t(\mathbf{n}_t|\mathbf{n}_{t-1},\widehat{\boldsymbol{\theta}})}{f_{1:T}(\mathbf{y}_{1:T}|\widehat{\boldsymbol{\theta}})}. \tag{4.10}$$

Estimated expected values for individual states can be used as point estimates and interval estimates can be based on percentiles of Eq. (4.10).

We note that this two-stage procedure is a special case of what is sometimes called parametric empirical Bayes (PEB) methods (Morris 1983). The label "Bayes" is potentially confusing as the procedure is not Bayesian. PEB methods are commonly used for random effects models. For example, consider the following simple random effects model where we retain some of the SSM notation (removing additional parameters from the observation model for simplicity):

$$\mathbf{n} \sim g\,(\mathbf{n}|\boldsymbol{\theta})$$
$$\mathbf{y} \sim f\,(\mathbf{y}|\mathbf{n})\,.$$

Both $\mathbf{n}$ and $\mathbf{y}$ are random vectors (or scalars), where $\mathbf{y}$ are observed and $\mathbf{n}$ are unknown and the random effects. Again the parameters are estimated first. Second, the conditional distribution for the random effects given the observations is calculated by substituting the mle's, say $\hat{\boldsymbol{\theta}}$, for the parameters:

$$g\left(\mathbf{n}|\mathbf{y},\hat{\boldsymbol{\theta}}\right).$$

This conditional distribution is also known as the estimated posterior distribution (Carlin and Louis 1996).

One key detail glossed over here is the degree of difficulty in carrying out the integrations, to yield the likelihood, the estimated posterior distribution, and parametric Empirical Bayes estimates. Skaug (2002), for example, uses the Laplace approximation to carry out some of the integration. A second key detail is the initial state, $\mathbf{n}_0$, whether it is viewed as a random effect (as shown here) or as a parameter (a constant), and in either case whether or not it is in fact estimable. Some discussion of how to handle $\mathbf{n}_0$ in the special case of normal dynamic linear SSMs is given in Sect. 9.5.1. Here we note that, given a long enough time series, the initial observation $\mathbf{y}_1$ can be used to estimate $\mathbf{n}_1$ in a somewhat *ad hoc* manner by inverting the observation model and solving for $\mathbf{n}_1$, and no inferences about $\mathbf{n}_0$ are made. A simple example of such an inversion would be an observation model where a scalar y_1 is proportional to n_1 (also scalar), say $y_1 \sim \text{normal}(\phi n_1, \sigma_y^2)$ for some unknown ϕ. Given an initial value for ϕ, n_1 is estimated by y_1/ϕ (Thomas et al. 2005, used this approach in a Bayesian setting to formulate a prior for the initial state value).

4.4 Classical Statistical Procedures for Fitting SSMs

The original discrete time SSM formulated by Kalman (1960) uses normal distributions of states and observations and the conditional expected values for states and for observations are linear combinations of previous states and current states, respectively, i.e. a normal dynamic linear model (NDLM). He developed a closed-form analytic algorithm, the Kalman filter, for doing the integration to calculate both the likelihood and the conditional distributions of the states, which has led to extremely widespread use for a broad range of problems. Additional explanations of the Kalman algorithms can be found in Meinhold and Singpurwalla (1983) and Harvey (1989).

4.4.1 The Kalman Filter

The Kalman filter is designed for fitting NDLMs. Even though stochastic integration techniques are more flexible because they allow fitting nonlinear, non-normal models, an overview of Kalman's algorithms is important on several counts, including their continued popular usage, relative simplicity and speed.

In a NDLM, the state and observation processes are normal random variables and the expected values are linear combinations of the conditioning variable, either $\mathbf{n}_{t-1}$ or $\mathbf{n}_t$. More concisely,

$$\mathbf{n}_0 \sim \text{normal}(\boldsymbol{\mu}_0, \mathbf{Q}_0) \tag{4.11}$$
$$\mathbf{n}_t|\mathbf{n}_{t-1} \sim \text{normal}(\mathbf{A}_{t-1}\mathbf{n}_{t-1}, \mathbf{Q}_t) \tag{4.12}$$
$$\mathbf{y}_t|\mathbf{n}_t \sim \text{normal}(\mathbf{B}_t\mathbf{n}_t, \mathbf{R}_t), \tag{4.13}$$

where $\mathbf{A}_{t-1}$ is an $m \times m$ matrix of constants, $\mathbf{B}_t$ is a $k \times m$ matrix of constants with k the number of observations in year t, and $\mathbf{Q}_t$ and $\mathbf{R}_t$ are covariance matrices for $\mathbf{n}_t$ and $\mathbf{y}_t$, respectively. Often the initial state vector, $\mathbf{n}_0$, is assumed to be a fixed value.

NDLMs may be viewed as approximations to more realistic nonlinear, non-normal SSMs for the dynamics underlying animal populations and the type of data collected from such populations. The existence of the Kalman filter (KF) makes NDLMs attractive. The KF is a recursive analytic procedure for estimating the states $\mathbf{n}_t$, given the observations up to and including time t, $\mathbf{y}_{1:t}$. In particular, the KF yields

T probability distributions for the unobserved states, n_t, $t = 1, \ldots, T$, conditional on the observations $y_{1:t}$, i.e. $f(n_t|y_{1:t})$, the so-called filtered distribution. A related algorithm, the Kalman smoother, calculates T probability distributions for the states conditional on *all* the observations, i.e. $f(n_t|y_{1:T})$. The resulting distributions in both cases are normal, thus the algorithms only need to calculate the mean vector and variance-covariance matrices. The KF can also be used to calculate the likelihood (the marginal distribution for the observations), which is necessary for calculation of maximum likelihood estimates. In comparison to computer-intensive procedures like MCMC and SIS, the Kalman filter can be very fast.

The Kalman filter is explained assuming that parameters of the pdf are known, i.e. the matrices $\mathbf{A}_t, \mathbf{B}_t, \mathbf{Q}_t, \mathbf{R}_t$ in Eqs. (4.11)–(4.13) are known. A web page devoted to Kalman algorithms which includes links to expository articles is http://www.cs.unc.edu/~welch/kalman/. The filter is a recursive algorithm in that values calculated for time $t-1$ are used to calculate values for time t. At each step of the recursion, there are two sub-steps: (a) predicting $\mathbf{n}_t$ using an estimate of $\mathbf{n}_{t-1}$; (b) "updating" $\mathbf{n}_t$ using a weighted combination of the prediction and the observation for time t, $\mathbf{y}_t$. Both the predicting and the updating steps have corresponding calculations for the variance-covariance matrices of the predictions and updated estimates.

Notation for the predicted state and its covariance matrix is $\mathbf{n}_t^{t-1}$ and $\mathbf{P}_t^{t-1}$, while that for the updated (or "filtered") state and its covariance matrix is $\mathbf{n}_t^t$ and $\mathbf{P}_t^t$.

Begin at $t = 0$ with fixed values for $\mathbf{n}_0$ and $\mathbf{P}_0^0$; often $\mathbf{P}_0^0$ is set equal to zero. Then predict the state at $t = 1$ using the state pdf, Eq. (4.12), and calculate the corresponding covariance matrix:

$$\text{Prediction of } \mathbf{n}_t: \quad \mathbf{n}_t^{t-1} = \mathbf{A}_{t-1}\mathbf{n}_{t-1}^{t-1} \tag{4.14}$$

$$\text{Covariance matrix of } \mathbf{n}_t^{t-1}: \quad \mathbf{P}_t^{t-1} = \mathbf{A}_{t-1}\mathbf{P}_{t-1}^{t-1}\mathbf{A}_{t-1}' + \mathbf{Q}_t. \tag{4.15}$$

Next update the state and calculate the corresponding covariance matrix:

$$\text{Update of } \mathbf{n}_t: \quad \mathbf{n}_t^t = \mathbf{n}_t^{t-1} + \mathbf{K}_t(\mathbf{y}_t - \mathbf{B}_t\mathbf{n}_t^{t-1}) \tag{4.16}$$

$$\text{Covariance matrix of } \mathbf{n}_t^t: \quad \mathbf{P}_t^t = (\mathbf{I} - \mathbf{K}_t\mathbf{B}_t)\mathbf{P}_t^{t-1}, \tag{4.17}$$

where $\mathbf{K}_t$ is referred to as the Kalman gain and is defined by

$$\mathbf{K}_t = \mathbf{P}_t^{t-1}\mathbf{B}_t'(\mathbf{B}_t\mathbf{P}_t^{t-1}\mathbf{B}_t' + \mathbf{R}_t)^{-1}.$$

Increment t by 1, repeat the calculations in Eqs. (4.14)–(4.17), and stop after finishing the updating for time T.

4.4.1.1 Numerical Demonstration of the Kalman Filter

The following simple NDLM is used to demonstrate the KF. The state vector has just two components. The matrix $\mathbf{A}_t$ is constant over time, as is the covariance matrix $\mathbf{Q}_t$.

Table 4.1 Observations (y_t), true unobserved states (n_t), and filtered states (n_t^t) from a simulated NDLM with $T = 10$ observations

Juveniles											
t	0	1	2	3	4	5	6	7	8	9	10
y_t	0	62	45	51	48	53	50	52	53	49	49
n_t	30	75	57	66	62	64	64	64	65	64	64
n_t^t	30	75	55	64	60	63	62	63	63	62	62
Adults											
t	0	1	2	3	4	5	6	7	8	9	10
y_t	0	42	48	46	45	44	45	46	44	45	47
n_t	50	38	44	41	43	43	43	43	42	43	43
n_t^t	50	37	43	40	41	41	41	41	42	42	42

Suppose that the $n_{t,1}$ are juveniles and $n_{t,2}$ are adults and the matrix $\mathbf{A}_t$ is a Leslie matrix. Let the state process equation be

$$\begin{bmatrix} n_{t,1} \\ n_{t,2} \end{bmatrix} = \begin{bmatrix} 0.0 & 1.5 \\ 0.3 & 0.55 \end{bmatrix} \begin{bmatrix} n_{t-1,1} \\ n_{t-1,2} \end{bmatrix} + \begin{bmatrix} q_{t,1} \\ q_{t,2} \end{bmatrix},$$

where $q_{t,1}$ and $q_{t,2}$ are assumed to be identically and independently distributed (iid) normal$(0, 0.5^2)$. We note that for this example we chose the values of the Leslie matrix carefully so that the deterministic rate of growth is exactly 1, i.e. deterministic projections of the population neither exponentially decline to extinction nor grow exponentially. This does not affect the main points we wish to demonstrate here, however.

Suppose that the observation vector has the same dimension as $\mathbf{n}_t$ and contains biased estimates of each component of $\mathbf{n}_t$, where the bias is assumed known.

$$\begin{bmatrix} y_{t,1} \\ y_{t,2} \end{bmatrix} = \begin{bmatrix} 0.8 & 0.0 \\ 0.0 & 1.1 \end{bmatrix} \begin{bmatrix} n_{t,1} \\ n_{t,2} \end{bmatrix} + \begin{bmatrix} r_{t,1} \\ r_{t,2} \end{bmatrix},$$

where $r_{t,1}$ and $r_{t,2}$ are iid normal$(0, 2^2)$.

Using R, the NDLM was simulated setting $\mathbf{n}_0 = (30, 50)$ and $\mathbf{P}_0^0 = 0$. The R code for simulation and using the Kalman filter is given on the book website (Sect. 1.2). Table 4.1 shows the simulated observations and states and the Kalman "filtered" estimates of the states. As is the case with Leslie matrix projection models (Caswell 2001), the abundances of juveniles and adults have, with the exception of relatively minor process variation, reached equilibrium values after four or five generations.

More commonly, the parameters are not known. The predicted state, $\mathbf{n}_t^{t-1}$, and corresponding covariance matrix, $\mathbf{P}_t^{t-1}$, can be used to construct the likelihood function (the marginal distribution for the observations). The likelihood can be written as follows (with conditioning on the initial state $\mathbf{n}_0$ made explicit).

$$L(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y}_{1:T}, \mathbf{n}_0) = f(\mathbf{y}_1|\mathbf{n}_0) \prod_{t=2}^{T} f(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \mathbf{n}_0). \quad (4.18)$$

It can be shown that the conditional density for each $\mathbf{y}_t$ is normal:

$$\mathbf{y}_t|\mathbf{y}_{1:t-1}, \mathbf{n}_0 \sim \text{multivariate normal}\left(\mathbf{B}_t \mathbf{n}_t^{t-1}, \mathbf{B}_t \mathbf{P}_t^{t-1} \mathbf{B}_t' + \mathbf{R}_t\right). \quad (4.19)$$

Note that the mean is the expected value of $\mathbf{y}_t$ based on the observation equation, with the predicted value of $\mathbf{n}_t$ substituted for its true, but unknown value. Example R code is given on the book website (Sect. 1.2) for calculating the likelihood for the values in the matrix A alone and for calculating the maximum likelihood estimates of the parameters. Shumway and Stoffer (1982) show how to use the EM algorithm with the KF for simultaneous inference of states and parameters.

4.4.2 *Extensions to NDLMs and the Kalman Filter*

By 1961 the Kalman filter had been extended to the case of nonlinear, but normal, SSMs (Grewal and Andrews 2010):

$$\mathbf{n}_0 \sim \text{normal}(\boldsymbol{\mu}_0, \mathbf{Q}_0) \quad (4.20)$$

$$\mathbf{n}_t|\mathbf{n}_{t-1} \sim \text{normal}\left(\mathbf{g}(\mathbf{n}_{t-1}), \mathbf{Q}_t\right) \quad (4.21)$$

$$\mathbf{y}_t|\mathbf{n}_t \sim \text{normal}\left(\mathbf{f}(\mathbf{n}_t), \mathbf{R}_t\right), \quad (4.22)$$

where g and f are vectors of differentiable functions. The Extended Kalman Filter (EKF) substitutes linear (first order) Taylor series approximations for nonlinear state and/or observation equations in prediction and update steps (Eqs. (4.14)–(4.17)). For the prediction step, prediction of the $\mathbf{n}_t$ simply involves plugging the current updated value, $\mathbf{n}_{t-1}^{t-1}$, into g, while calculation of the corresponding covariance matrix, P_t^{t-1}, includes linearization of g:

$$\mathbf{n}_t^{t-1} = \mathbf{g}\left(\mathbf{n}_{t-1}^{t-1}\right)$$
$$\mathbf{P}_t^{t-1} = \mathbf{G}_{t-1}\mathbf{P}_{t-1}^{t-1}\mathbf{G}_{t-1}' + \mathbf{Q}_t,$$

where

$$\mathbf{G}_{t-1} = \left.\frac{\partial \mathbf{g}}{\partial \mathbf{n}}\right|_{\mathbf{n}_{t-1}^{t-1}}.$$

For the update step, the updated estimate is calculated as for the regular Kalman filter, but the updated estimated covariance involves a linearization of f:

$$\mathbf{n}_t^t = \mathbf{n}_t^{t-1} + \mathbf{K}_t(\mathbf{y}_t - \mathbf{B}_t\mathbf{n}_t^{t-1})$$
$$\mathbf{P}_t^t = (\mathbf{I} - \mathbf{K}_t\mathbf{F}_t)\mathbf{P}_t^{t-1},$$

where

$$\mathbf{F}_t = \left.\frac{\partial \mathbf{f}}{\partial \mathbf{n}}\right|_{\mathbf{n}_t^{t-1}}.$$

The realizable minimum variance (RMV) filter (due to Liang and Christensen 1978, but see the concise description in Quinn and Deriso 1999:236–237) is another procedure for handling nonlinear but normal SSMs. Julier et al. (1995) proposed the Unscented Kalman Filter (UKF) as a more accurate alternative to the EKF (see the review article by Wan and van der Merwe 2001).

The EKF and RMV filters also allow calculation of the likelihood of the parameters. Wan and van der Merwe (2001) describe two dual estimation procedures using the UKF, the dual UKF and the joint extended UKF.

A variation on NDLMs, conditionally Gaussian models, which can be analysed using the standard Kalman filter, are discussed in Sect. 9.7.1.

4.4.3 *Numerical Procedures*

When the SSM is nonlinear and/or non-normal, numerical procedures are sometimes used. We begin with a method developed by Kitagawa (1987). Kitagawa noted that prior to his work, the typical approach to handling nonlinear and non-normal SSMs was "to approximate the non-normal distribution by one or several normal distributions or by some parametric function"; e.g. the extended Kalman filter, the Gaussian (i.e. normal) sum filter (Alspach and Sorenson 1972). Kitagawa, however, worked directly with non-normal distributions of the following form:

$$n_t = Fn_{t-1} + Gv_t$$
$$y_t = Hn_t + w_t,$$

where for his numerical solution n_t and y_t were *scalars*, and v_t and w_t were non-normal random variables. He formulated iterative algorithms for evaluating one-step-ahead predictive density of the state (i.e. $g(n_t|\mathbf{y}_{1:t-1})$), filtered state density (i.e. $g(n_t|\mathbf{y}_{1:t})$), and "fixed" interval smoothed state density (i.e. $g(n_t, n_{t+1}|\mathbf{y}_{1:T})$). Given those algorithms, he numerically carried out evaluations of various conditional densities and convolutions of densities using piecewise linear (first-order spline) functions. Extensions to higher dimension state and observation vectors were discussed but not demonstrated.

A computationally efficient method for calculating maximum likelihood estimates of SSM parameters and then making inferences about unobserved states in high-dimensional nonlinear and/or non-normal SSMs, or more generally hierarchical models, has been developed by Skaug and Fournier (2006; see also Skaug 2002). They have developed freely-distributed companion software, ADMB-RE

(Automatic Differentiation Model Builder—Random Effects, Skaug and Fournier 2011), for fitting SSMs. Their approach has several features.

1. The SSM is viewed as a special case of a random effects model, i.e. a *hierarchical* model, where the unobserved states are the random effects.
2. The marginal likelihood for the observations [Eq. (4.2)] is calculated approximately (typically with high accuracy) and quickly using automatic differentiation (Skaug and Fournier 2006) and the Laplace method (Tierney and Kadane 1986) to carry out the high-dimensional integration over the state vector.
3. Given the marginal likelihood, maximum likelihood estimates of the fixed SSM parameters are calculated (although Bayesian inference is an option).
4. Parametric empirical Bayes (Morris 1983) estimates of the random effects, i.e. the unobserved states, are calculated conditional on the maximum likelihood estimates, i.e. given the mle's for the parameters, ADMB-RE "automatically calculates 'maximum posterior' estimates of the random effects" (Skaug and Fournier 2011).

4.5 Monte Carlo Simulation Procedures

Computer-intensive Monte Carlo procedures such as Markov chain Monte Carlo (MCMC) and sequential importance sampling (SIS) are more commonly used within a Bayesian framework, but they can also be used for classical analyses (Geyer 1996; de Valpine 2002, 2003, 2004; de Valpine and Hilborn 2005; Ionides et al. 2006). Below we discuss MCMC and SIS as distinct approaches but note that the methods can be used in combination, as in Partical MCMC (Andrieu et al. 2010).

4.5.1 Markov Chain Monte Carlo

The general idea of Markov chain Monte Carlo (MCMC) is to generate samples from a Markov chain which has a limiting distribution equal to the desired distribution (Metropolis et al. 1953; Hastings 1970; Gilks et al. 1996; Brooks et al. 2011). The iterative procedure known as the Metropolis–Hastings algorithm is a means of constructing the appropriate Markov chain. Before describing the algorithm for the case of an SSM with unknown parameters, we begin with a simpler case of a sample of n independent identically distributed random variables, $y_1, \ldots, y_n$ with pdf $f(\mathbf{y}|\theta)$, where θ is an unknown scalar and $\mathbf{y} = \{y_1, \ldots, y_n\}$. Let $\pi(\theta)$ be the prior for an unknown parameter and let $f(\mathbf{y}|\theta)$ denote the likelihood. The Metropolis–Hastings procedure generates a sample from the posterior distribution for θ, i.e. a sample from $\pi(\theta|\mathbf{y})$, in the following manner. At each iteration i of the algorithm, let the current state of the chain be denoted by θ^i. A *candidate* value, denoted θ', is generated from some *proposal* distribution with pdf $q(\theta'|\theta^i)$. The proposed parameter is accepted with probability $\alpha(\theta^i, \theta') = \min(1, p_\theta)$, where

$$p_\theta = \frac{\pi(\theta') f(\mathbf{y}|\theta')}{\pi(\theta^i) f(\mathbf{y}|\theta^i)} \frac{q(\theta^i|\theta')}{q(\theta'|\theta^i)}. \tag{4.23}$$

If the move is accepted, θ^{i+1} is set equal to θ'; else θ^{i+1} is set equal to θ^i.

Beginning with $i = 0$, and an initial arbitrarily chosen value for θ, denoted θ^0, a sequence of θ's is generated for $i = 1, \ldots, B + N$. The first B iterations are the so-called burn-in period which is the period prior to convergence of the sample to the desired limiting distribution, and the next N iterations are the inference sample.

In the more general case, where θ is a vector, the above Metropolis–Hastings algorithm generalises immediately, where the proposal density q is a multivariate pdf. An alternative approach (and probably the most common approach in general) is the single-update Metropolis–Hastings algorithm. In this algorithm, each iteration of the Metropolis–Hastings step involves cycling through each parameter in turn and proposing a new candidate value for the given parameter being updated. This value is accepted with the probability given above (though algebraic simplifications can often be made). A single iteration is completed after each parameter has been updated. An example of the single-update algorithm is described mathematically in Sect. 4.5.1.2.

4.5.1.1 Implementation Issues

There are several decisions that must be made when carrying out MCMC simulations using the Metropolis–Hastings algorithm. Here we just list some of the issues; Gilks et al. (1996) and Brooks et al. (2011) are excellent guides to practical implementation.

1. *Burn-in length, B.* There are various ways to identify a suitable value for B. Usually, time series plots of the iterated values, also known as trace plots, are visually examined to see whether or not the simulated values are changing relatively rapidly as i increases. This is visual evidence for what is termed "good mixing", i.e. relatively rapid sampling of the posterior distribution with little "sticking" at particular values. For example, the simulated values of θ may be generally increasing from 0.1 to 0.4 for $i = 1, \ldots, 100$, and then for $i > 100$, the values are on average 0.4 but vary from 0.2 to 0.6 with no recognisable pattern. In that case, B might be set at 100.

   A less subjective approach is to start the simulation several times, i.e. to run multiple chains, each time using a different initial value, and then to examine the point at which the chains overlap consistently. A formal measure of convergence is the Brooks-Gelman-Rubin (B-G-R) statistic (Brooks and Gelman 1998), which is essentially an F-ratio statistic that compares the variation between chains with variation within chains. Values of the B-G-R statistic near 1.0 are considered indicative of convergence.

2. *Chain length, N*. The sample size can be chosen pragmatically on the basis of the Monte Carlo error of statistics of the posterior sampling distribution. For example, multiple chains are run and the 0.05 percentiles for θ are compared between chains, and if the between-chain variation, i.e. Monte Carlo variation, is considered low enough, then N is sufficient.
3. *Choice of proposal distribution, $q(\theta'|\theta_i)$*. The proposal distribution affects the mixing of the chain and time needed until convergence. There are various somewhat mechanical choices for proposals. In some cases, the full conditional distributions, e.g. $\pi(\theta_1|\mathbf{y}, \theta_2, \ldots, \theta_q)$, are known and these conditionals then serve as proposals. In this case the acceptance probabilities of Equation (4.23) are 100 %, i.e. every candidate value is kept, and the Metropolis–Hastings sampler becomes what is known as the Gibbs sampler. For other alternatives including random walk proposals and independence proposals, see Gilks et al. (1996).
4. *Blocking*. With multiple parameters, e.g. $\theta = (\beta_0, \beta_1, \sigma^2)'$ as in a simple linear regression, we can choose to implement a single-update Metropolis–Hastings algorithm or simultaneously generate candidate values for two or more parameters at one time. Simultaneously updating two or more parameters is known as blocking and is most commonly used when the parameters within a block are highly correlated as it can improve mixing (allowing a smaller value of N and reducing the number of iterations B until convergence).

4.5.1.2 MCMC for SSMs

Here we describe one way that the Metropolis–Hastings algorithm can be used to generate a sample from the joint posterior distribution of states and parameters, $\pi(\mathbf{n}_{0:T}, \boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y}_{1:T})$. For an early application of MCMC methods, in particular the Gibbs Sampler, for a non-normal, nonlinear SSM see Carlin et al. (1992).

To begin, prior distributions $\pi(\boldsymbol{\theta}, \boldsymbol{\psi})$ and $g_0(\mathbf{n}_0|\boldsymbol{\theta})$ for the parameters and the initial state vector are specified. Here we assume that a full iteration of the sampler produces a complete vector of all the state values and parameters; i.e. at the end of iteration i, all the states and parameters are "updated". Furthermore, at each iteration assume that the chain first produces individual components of the parameter vector one at a time, then the state vectors, $\mathbf{n}_0, \mathbf{n}_1, \ldots, \mathbf{n}_T$ are generated in sequence, and assume that all components of the state vector at time t are produced simultaneously.

For convenience, we denote the combined vector of state and observation parameters by $\boldsymbol{\eta} = \{\boldsymbol{\theta}, \boldsymbol{\psi}\}$ with elements η_r, $r = 1, \ldots, R$. At iteration i, we cycle through each element of $\boldsymbol{\eta}$ in turn, propose a candidate value for the given parameter, and accept/reject this candidate value with the given specified acceptance probability. Mathematically, suppose that we propose to update parameter η_r, such that (for $r > 1$) we have already updated parameters $\eta_1, \ldots, \eta_{r-1}$. We denote the set of current parameters by $\boldsymbol{\eta}^i_{[r]} = \{\eta_1^{i+1}, \ldots, \eta_{r-1}^{i+1}, \eta_r^i, \eta_{r+1}^i, \ldots, \eta_R^i\}$ and current state vectors, $\mathbf{n}^i_{0:T}$. Simulate a *candidate* value, denoted η'_r from the *proposal* distribution with pdf $q(\eta'_r|\eta_r^i)$, and set $\boldsymbol{\eta}'_{[r]} = \{\eta_1^{i+1}, \ldots, \eta_{r-1}^{i+1}, \eta'_r, \eta_{r+1}^i, \ldots, \eta_R^i\}$. The candidate

value, η'_r is accepted with probability $\alpha_{\eta_r} = \min\left(1, p_{\eta_r}\right)$, where

$$p_{\eta_r} = \frac{\pi(\boldsymbol{\eta}'_{[r]}) g_0(\mathbf{n}_0^i|\boldsymbol{\eta}'_{[r]}) \prod_{t=1}^T f_t(\mathbf{y}_t|\mathbf{n}_t^i, \boldsymbol{\eta}'_{[r]}) g_t(\mathbf{n}_t^i|\mathbf{n}_{t-1}^i, \boldsymbol{\eta}'_{[r]})}{\pi(\boldsymbol{\eta}^i_{[r]}) g_0(\mathbf{n}_0^i|\boldsymbol{\eta}^i_{[r]}) \prod_{t=1}^T f_t(\mathbf{y}_t|\mathbf{n}_t^i, \boldsymbol{\eta}^i_{[r]}) g_t(\mathbf{n}_t^i|\mathbf{n}_{t-1}^i, \boldsymbol{\eta}^i_{[r]})} \frac{q(\eta_r^i|\boldsymbol{\eta}'_{[r]})}{q(\eta'_r|\boldsymbol{\eta}^i_{[r]})}.$$

If the move is accepted, set $\eta_r^{i+1} = \eta'_r$; otherwise, set $\eta_r^{i+1} = \eta_r^i$. Following completion of each parameter update, we denote the set of updated parameters by $\boldsymbol{\eta}^{i+1}$.

Similarly for the (block) updating of the states, we cycle through each state vector in turn ($\mathbf{n}_t$ for $t = 0, \ldots, T$), simulate a candidate state vector, and accept or reject the proposed values. Consider time $t = 1, \ldots, T-1$, with current set of all state vectors, $\mathbf{n}_0^{i+1}, \ldots, \mathbf{n}_{t-1}^{i+1}, \mathbf{n}_t^i, \mathbf{n}_{t+1}^i, \ldots, \mathbf{n}_T^i$. Propose a candidate state vector for time t, denoted $\mathbf{n}'_t$, from some proposal density $q_{\mathbf{n}_t}(\mathbf{n}'_t|\mathbf{n}_t^i)$. The proposed state vector is accepted with probability $\alpha_{\mathbf{n}_t}(\mathbf{n}_t^i, \mathbf{n}'_t) = \min(1, p_{\mathbf{n}_t})$, where

$$p_{\mathbf{n}_t} = \frac{f_t(\mathbf{y}_t|\mathbf{n}'_t, \boldsymbol{\theta}^{i+1}) g_t(\mathbf{n}'_t|\mathbf{n}_{t-1}^{i+1}, \boldsymbol{\psi}^{i+1}) g_{t+1}(\mathbf{n}_{t+1}^i|\mathbf{n}'_t, \boldsymbol{\psi}^{i+1})}{f_t(\mathbf{y}_t|\mathbf{n}_t^i, \boldsymbol{\theta}^{i+1}) g_t(\mathbf{n}_t^i|\mathbf{n}_{t-1}^{i+1}, \boldsymbol{\psi}^{i+1}) g_{t+1}(\mathbf{n}_{t+1}^i|\mathbf{n}_t^i, \boldsymbol{\psi}^{i+1})} \frac{q_{\mathbf{n}_t}(\mathbf{n}_t^i|\mathbf{n}'_t)}{q_{\mathbf{n}_t}(\mathbf{n}'_t|\mathbf{n}_t^i)}$$

(explicitly using the terms $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ instead of the $\boldsymbol{\eta}$ terminology above). If the move is accepted, set $\mathbf{n}_t^{i+1} = \mathbf{n}'_t$; else set $\mathbf{n}_t^{i+1} = \mathbf{n}_t^i$. We note that for the cases $t = 0$ and $t = T$, the acceptance probability simplifies further. For $t = 0$, the term $g_t(\mathbf{n}_t|\mathbf{n}_{t-1}, \boldsymbol{\psi}) = g_0(\mathbf{n}_0|\boldsymbol{\psi})$; for $t = T$, the term $g_{t+1}(\mathbf{n}_{t+1}|\mathbf{n}_t, \boldsymbol{\psi}) \equiv 1$.

In practice, efficient implementation of an MCMC sampler for a SSM can be very difficult (Newman et al. 2009; Fearnhead 2011; King 2011). For example, correlation between states and parameters can be very high which means that very long simulations of the Markov chain may be necessary before the chain converges to the posterior distribution.

The resulting posterior sample of states is a sample from the *smoothed* distribution, i.e. all observations, $\mathbf{y}_{1:T}$, are used to make inferences about states at all time periods. This is in contrast to some inference procedures for states which produce samples from the *filtered* distribution, i.e. only observations up to and including time t, $\mathbf{y}_{1:t}$, are used to make inferences about $\mathbf{n}_t$.

4.5.2 BUGS Project Software

The BUGS (Bayesian inference Using Gibbs Sampling) project (www.mrc-bsu.cam.ac.uk/bugs/) has led to the freely available software WinBUGS and OpenBUGS for model fitting using MCMC (Lunn et al. 2000, 2013). WinBUGS is a stable version of BUGS (the final version of this software is 1.4.3) while OpenBUGS is an open-source version "on which all future development work will be focused". As suggested by the name, WinBUGS is restricted to Windows operating systems (or Windows emulators), while OpenBUGS is available for Windows and Linux

(using a command line interface). These packages have a graphical user interface (GUI) but can also be run from within R using the associated packages R2WinBUGS (for interacting with WinBUGS) and R2OpenBUGS or BRugs (for interacting with OpenBUGS). The BUGS language is also used in the software JAGS (Plummer 2003), freely available from http://mcmc-jags.sourceforge.net with versions for Windows, Linux and Mac OS X operating systems. JAGS has a command line based interface, but with an associated R package rjags for calling the program from R. For in-depth discussion of the BUGS language, numerous examples and further description of the different packages, including differences between them, see Lunn et al. (2013).

We will demonstrate the use of WinBUGS to fit a SSM but before doing so, we give a simpler example of Bayesian inference for a linear regression.

4.5.2.1 Set-Up

There are three sets of program code for most WinBUGS programs: (i) the model statement, (ii) the data statement, and (iii) the initial values.

The model statement consists of pieces of code that define the priors and the likelihood. For example, for a simple linear regression, where $y \sim \text{normal}(\beta_0 + \beta_1 x, \sigma^2)$:

```
 #(i) Model Statement
model {
 #Prior distribution for parameters
 beta1 ~ dnorm(0,0.01)
 beta0 ~ dnorm(0,0.01)
 sigma ~ dunif(0.01,10)
 tau   <- 1/(sigma*sigma)  #the precision

 #Likelihood defined
 for(i in 1:n) {
   mu[i] <- beta0+beta1*x[i]
   y[i] ~ dnorm(mu[i],tau)
 }
}
```

Note that the priors for the coefficients, β_0 and β_1, are normal($\mu = 0, \sigma^2 = 1/0.01 = 100$), as the WinBUGS syntax for normal random variables specifies the precision, or inverse of the variance. The symbol # denotes a comment in WinBUGS.

The data statement specifies the values corresponding to the observations, relevant covariates, and sample size that have been used in the model statement. Again referring to the simple linear regression example:


```
 #(ii) Data Statement

list(n=15,
     y=c( 29.4, 33.8, 24.8, 26.1, 31.6, 29.4, 14.6,
       12.6, 29.3, 27.5,28.9, 28.3 ,24.7, 28.5,
       14.9),
     x=c(8, 10,  7,  7,  9, 9,  3,  3,  8,  7,  8,
       8,  7,  8,  3))
```

The initial values are the θ^0s used to begin the chain. e.g.

```
 #(iii) Initial Values Statement

 list(beta0=0.5, beta1=1.0, sigma=3.0)
```

Note that, in general, it is possible to generate all (or some of) the initial values in WinBUGS. In this case the starting values are generated from the corresponding prior distribution specified for the parameters. However, not all prior distributions can be simulated from, in which case an error message is displayed stating that an initial value cannot be generated for the given distribution. To run the MCMC simulations, initial starting values do need to be specified for (at least) these parameters. For example, WinBUGS cannot simulate from a $\Gamma(0.001, 0.001)$ prior.

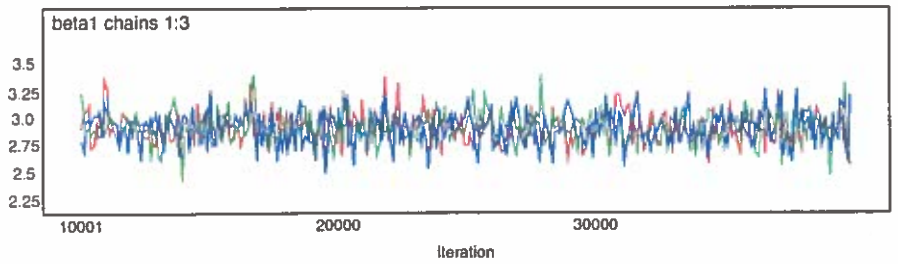
For SSMs, initial values need to be specified (or generated) for all the parameters and unknown states. However, care typically needs to be taken in specifying the initial values for the unknown states so that impossible state transitions (given the initial parameter values) are not specified. In practice, the observation vector $\mathbf{y}_{1:T}$ is often used for the observed parts of the state vector (although this can still lead to impossible state transitions). More generally, given the specified model parameters and initial value for $\mathbf{n}_0$, the remaining initial state vectors $\mathbf{n}_{1:T}$ can be obtained by simulating forward from the given model (see Sect. 4.5.3 for an example).

4.5.2.2 Program Execution

The sequence of steps to run WinBUGS and produce a sample is:

1. *Check* model statement for syntax errors.
2. *Load data.*
3. *Compile model* to check for compilation errors.
4. *Set initial values* for θ^0.
5. *Specify model outputs*, i.e. the parameters for which MCMC output is to be shown.
6. *Specify number of iterations.*

More than one chain can be generated, which then allows the B-G-R statistic to be calculated. Note that it is up to the user to specify the burn-in, B, to be used in any posterior summary estimates.



node	mean	sd	MC error	2.5%	median	97.5%
beta0	5.61	1.043	0.03399	3.50	5.56	7.738
Beta1	2.86	0.142	0.00442	2.58	2.86	3.142
Sigma	1.14	0.252	0.00856	0.78	1.10	1.778

Fig. 4.1 `WinBUGS` output from fitting a simple linear regression. The trace plot shown is for the slope coefficient β_1, based on three chains with differing starting values, $B = 10{,}000$ and $N = 30{,}000$

4.5.2.3 Program Output

The output from `WinBUGS` includes trace plots of the simulated parameter values, the B-G-R statistic if multiple chains are run, and summary statistics from the posterior distribution (including the mean, standard deviation, and various quantiles). Recall that the burn-in, B, should be initially determined (for example using the B-G-R statistic) and then discarded before calculating these summary statistics. Example output from fitting the simple linear regression model is shown in Fig. 4.1. Three chains were run with different starting values, each with a burn-in of 10,000 iterations followed by 30,000 iterations for inference. The posterior mean values for β_0, β_1 and σ shown in the figure are quite similar to least squares estimates of 5.69, 2.85 and 1.04. The MC error, Monte Carlo error, is computational error in the calculation of the mean value and can be made arbitrarily small by increasing the number of simulations.

4.5.3 Fitting the Simplified Coho Salmon SSM with `WinBUGS`

For the univariate coho salmon SSM defined previously (Sect. 3.2.1), a Ricker population dynamics model with Poisson variation and bias-corrected lognormal observations, was fitted using `WinBUGS`. Twenty years of observations were used (the first 20 shown in Fig. 3.2; see the `R` code on the book website (Sect. 1.2) for Chap. 3). These had a coefficient of variation of 30 % (`CV.obs`). The model definition code, available on the book website, estimates three parameters, α, β and n_0, assigned uniform prior distributions, while the observation CV was treated as a known constant.

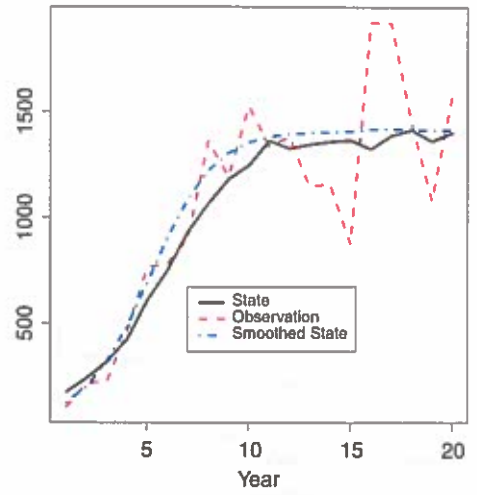


Fig. 4.2 Fitted states, true states and estimates for univariate coho salmon SSM with Ricker population dynamics (Poisson variation) and lognormal observations

Table 4.2 Posterior means for parameters of state process of the univariate coho salmon model based on MCMC sampling procedure

Parameters	True	SSM	Non-SSM
α	1.50	1.73	1.91
β	3.00e-4	3.90e-4	4.40e-4
n_0	135	76.4	75.1

Means based on the correctly specified SSM are in the column SSM while means based on fitting a non-SSM Ricker model are in the column non-SSM

Rather than inputting the model statement, data and initial values directly within `WinBUGS`, the values were passed from `R` using the function `bugs` from the `R` package `R2WinBUGS`. The relevant `R` code is also available on the book website. This is particularly convenient for generating multiple sets of initial values $\boldsymbol{\theta}_0$ (α, β and $n_{0:20}$) for running multiple MCMC chains. A case-specific function `init.value.generator` was written in this case to provide initial values for the entire state variable sequence.

The posterior means of the state variables are plotted along with the true state values and observations in Fig. 4.2. The posterior means for the parameters α, β and n_0 are shown along with the true values in Table 4.2. For comparison, a non-SSM model was fitted, where the estimated abundances were modelled according to the Ricker model with lognormal errors, i.e.

$$\hat{n}_t|\hat{n}_{t-1} \sim \text{lognormal}\left(\log\left(\alpha\hat{n}_{t-1}\exp(-\beta\hat{n}_{t-1})\right) - \sigma^2/2, \sigma^2\right).$$

Both models overestimated the productivity parameter α and the density dependence parameter β, but the SSM estimates were closer to the true values.

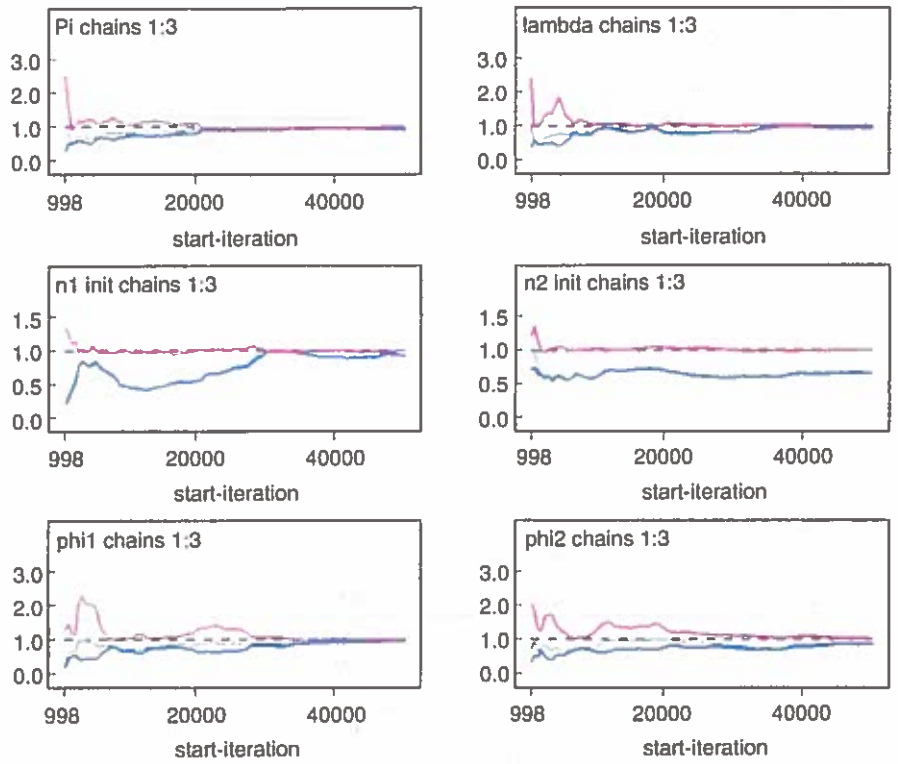


Fig. 4.3 Plot of B-G-R statistic versus iteration based on 3 independent MCMC chains for the parameters π, ρ, $n_{1,0}$, $n_{2,0}$, ϕ_1 and ϕ_2. The *green line* is the width of the central 80 % interval for all three runs pooled and based on bins of length 50, while the *blue line* gives the average width of the 80 % intervals within all three runs individually, and the *red line* is the ratio R of pooled to averaged within credible interval widths. Ideally R converges to 1 and the pooled and within lines converge

4.5.4 Fitting the BRS SSM with `WinBUGS`

For a somewhat more complex SSM, the BRS model (Sect. 3.2.2) was fitted using `WinBUGS`. The `WinBUGS` code is available on the book website (Sect. 1.2) as is the generating `R` code. As mentioned previously, with a sequence of three sub-processes, the pdf for $\mathbf{n}_t|\mathbf{n}_{t-1}$ is intractable. The `WinBUGS` code deals with this by explicitly defining the intermediate or "latent" states, the u's, and their corresponding pdf's. Three different chains were run; each had a different set of initial values for the parameters as well as for the latent states, all of which were generated within `R`.

The B-G-R statistic (Fig. 4.3) indicated that the chain was not converging for the initial population size $n_{1,0}$, which is possibly indicative of non-identifiability or weak identifiability problems. However, the posterior mean (smoothed) values for the states matched well with the true states (Fig. 4.4).

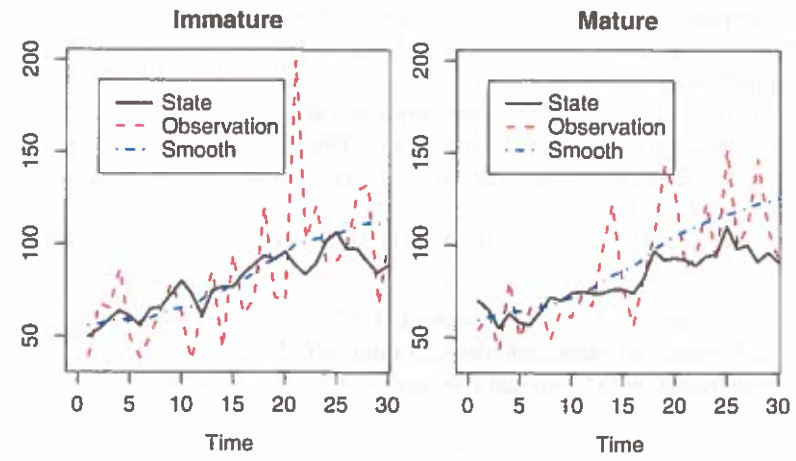


Fig. 4.4 Fitted states, true states and estimates for BRS SSM with lognormal observations

Table 4.3 Parameters and posterior means for the BRS model based on 30 years of simulated data

Parameter	True value	Posterior distribution 2.5 %	Mean	97.5 %
ϕ_1	0.60	0.13	0.50	0.92
ϕ_2	0.70	0.37	0.76	0.94
ρ	0.70	0.37	0.72	0.93
$n_{1,0}$	50	13	89	318
$n_{2,0}$	70	13	54	122

The true parameter values along with posterior distribution summaries, based on the combined output of three MCMC chains of length 50,000 following a burn-in of 50,000, are shown in Table 4.3.

4.5.5 Sequential Importance Sampling

We discuss sequential importance sampling (SIS) in the context of Bayesian inference for a SSM. Once again the objective is to produce a sample from the posterior distribution for the parameters and states. The collection of papers edited by Doucet et al. (2001) is a good reference for details of sequential importance sampling in its various forms.

4.5.5.1 Importance Sampling

Before discussing SIS, we describe "ordinary" importance sampling. Suppose we want a sample of values of a random variable X from a distribution with pdf $f(x)$.

Further suppose that generating samples directly from $f(x)$ is difficult, but given a particular value x, $f(x)$ can be evaluated. Let $g(x)$ be the pdf for a random variable with support including the support of $f(x)$, i.e. if $f(x) > 0$, $g(x) > 0$. This simply means that if $f(x)$ can produce values between 0 and 20, say, then $g(x)$ needs to be able to produce values between 0 and 20, too. The distribution of interest, $f(x)$, is referred to as the target distribution and the generating distribution, $g(x)$, is the trial distribution (Liu 2001).

The following importance sampling algorithm yields an independent sample from $f(x)$:

1. Generate a sample of N values from $g(x)$, $x_1^*, \ldots, x_N^*$.
2. For each generated value, calculate the ratio, $w'(x_i^*)=f(x_i^*)/g(x_i^*)$.
3. Scale the ratios, $w'(x_i^*)$, so that they sum to 1:

$$w(x_i^*) = \frac{w'(x_i^*)}{\sum_{i=1}^{N} w'(x_i^*)}.$$

4. Resample the x^* by sampling values x_i^* with replacement and with probability $w(x_i^*)$.

We note that, with this particular implementation of an importance algorithm, $f(x)$ need only be evaluated up to an unknown constant of proportionality. The rescaling has the effect of cancelling out the unknown proportionality constant. Thus, as with MCMC, this is particularly relevant for Bayesian inference in that the normalizing constant (the denominator in the posterior pdf, Eq. (4.5)) need not be calculated.

4.5.5.2 Sequential Importance Sampling for SSMs

To use ordinary importance sampling to generate the posterior sample for the states and parameters of an SSM, we need to specify a relatively high-dimensional pdf $g(\mathbf{n}_0, \ldots, \mathbf{n}_T, \boldsymbol{\theta})$. This has sometimes been done using distributions such as a multivariate t-distribution (e.g. Cunningham 2002). An alternative, "divide and conquer" approach is sequential importance sampling (Liu and Chen 1998), which is, as the name suggests, a sequential implementation of importance sampling. Doucet et al. (2001) give a nice description of the method and our explanation below largely follows theirs.

We will just explain how SIS proceeds for an SSM where the parameters and initial state $\mathbf{n}_0$ are known. Dual inference, about the unknown parameters as well as the states, can be done using a method developed by Liu and West (2001).

At time $t = 1$, generate $\mathbf{n}_1^*$ from a trial distribution pdf $h_1(\mathbf{n}_1)$ where h_1 can depend upon $\mathbf{n}_0$ and $\boldsymbol{\theta}$. Evaluate the following weight:

$$w_1 = \frac{g_1(\mathbf{n}_1^*|\mathbf{n}_0, \boldsymbol{\theta}) f_1(\mathbf{y}_1|\mathbf{n}_1^*, \boldsymbol{\psi})}{h_1(\mathbf{n}_1^*)}.$$

Next, generate $\mathbf{n}_2^*$ from a pdf $h_2(\mathbf{n}_2)$ where h_2 can depend upon parameters, previous states, and previous and current observations. Evaluate a new weight

$$w_2 = w_1 \times \frac{g_2(\mathbf{n}_2^*|\mathbf{n}_1^*, \boldsymbol{\theta}) f_2(\mathbf{y}_2|\mathbf{n}_2^*, \boldsymbol{\psi})}{h_2(\mathbf{n}_2^*)}.$$

Proceed in a similar manner for the remaining time periods. After the last time period,

$$w_T = \prod_{t=1}^{T-1} w_t \times \frac{g_T(\mathbf{n}_T^*|\mathbf{n}_{T-1}^*, \boldsymbol{\theta}) f_T(\mathbf{y}_T|\mathbf{n}_T^*, \boldsymbol{\psi})}{h_T(\mathbf{n}_T^*)}$$
$$= \prod_{t=1}^{T} \frac{g_t(\mathbf{n}_t^*|\mathbf{n}_{t-1}^*, \boldsymbol{\theta}) f_t(\mathbf{y}_t|\mathbf{n}_t^*, \boldsymbol{\psi})}{h_t(\mathbf{n}_t^*)}.$$

Note that a special case of a trial distribution is where h_t is g_t, the state process pdf. In this case, the weights are proportional to the observation process pdf, $f_t(\mathbf{y}_t|\mathbf{n}_t^*)$.

The weights are subsequently scaled to cancel out the proportionality constant $1/f(\mathbf{y}_1, \ldots, \mathbf{y}_T)$, and the generated sequences can be re-sampled according to these scaled weights to yield a sample from the posterior distribution. An alternative to accumulating the weights in this manner is to re-sample at various points in time, with probabilities proportional to the current weights, according to some re-sampling schedule. One extreme schedule is to re-sample every time period and eliminate the need to store weights. Such re-sampling schedules can affect the degree of Monte Carlo variation in the posterior samples.

In the case of unknown parameters and initial state $\mathbf{n}_0$, we generate a sample of parameters and initial states from a sampler with joint pdf $h_{\boldsymbol{\theta},\boldsymbol{\psi}} \times h_0$, say. The weight at time T is then

$$w_T = \frac{\pi(\boldsymbol{\theta}, \boldsymbol{\psi}) g_0(\mathbf{n}_0|\boldsymbol{\theta})}{h_{\boldsymbol{\theta},\boldsymbol{\psi}}(\boldsymbol{\theta}, \boldsymbol{\psi}) h_0(\mathbf{n}_0|\boldsymbol{\theta})} \prod_{t=1}^{T} \frac{g_t(\mathbf{n}_t^*|\mathbf{n}_{t-1}^*, \boldsymbol{\theta}) f_t(\mathbf{y}_t|\mathbf{n}_t^*, \boldsymbol{\psi})}{h_t(\mathbf{n}_t^*)}.$$

As for MCMC, efficient implementation of SIS is not necessarily simple. One of the problems is that the distribution of weights can be quite asymmetric with relatively few sample vectors having most of the weight, a problem known as particle depletion. This can be problematic even with very large initial sample size. There are numerous techniques aimed at reducing particle depletion, such as the auxiliary particle filter (Pitt and Shephard 1999; Liu and West 2001), residual resampling (Liu and Chen 1998), and partial rejection control (Liu 2001).

The algorithm of Liu and West (2001) adds the step of kernel smoothing the selected parameter values, after an importance sampling step, to lessen the degree of particle depletion. R code implementing this algorithm for the univariate coho salmon SSM is available on the book website (Sect. 1.2). The program was run

Table 4.4 Posterior means for parameters of state process of the univariate coho salmon model based on both MCMC and SIS sampling procedures

Parameters	True	MCMC	SIS
α	1.50	1.73	1.64
β	3.00e-4	3.90e-4	3.54e-4
n_0	135	76.4	104.6

with 200,000 particles, where the initial parameter values were drawn from the same prior distributions used in the MCMC implementation. The degree of kernel smoothing was set at 5 %, i.e. 95 % of the weighting in the smooth came from the original, unsmoothed value. Particle depletion was 99.6 %; i.e. of the 200,000 original particles, only 850 of the original simulated values (ignoring effects of kernel smoothing on parameters which perturbs *all* parameters) were left by the 20th year of simulations. In this particular instance, the SIS posterior means for α, β and n_0 (Table 4.4) were closer to the true values than the MCMC values, but this should not be overstated due to Monte Carlo variation and perhaps due to the bias introduced by kernel smoothing (Newman et al. 2009). The prior and posterior densities for the parameters are shown in Fig. 4.5.

4.6 Selecting a Fitting Procedure

We have discussed several approaches for fitting state-space models, distinguishing classical and Bayesian frameworks as well as analytical and simulation-based solutions. How does one determine which fitting procedure to use? No single procedure is consistently the best choice; e.g. Monte Carlo solutions are not necessarily better than analytic solutions. To make a decision, consideration of the following criteria may be useful.

1. Code development time or ease of implementation.
2. Computational efficiency, namely the computing time to produce parameter and state estimates.
3. Accuracy of estimated parameters and states.
4. Degree of fidelity to reality.
5. Risk associated with decisions made using the SSM.

Newman et al. (2009) compared MCMC and SIS using the first three criteria and concluded that there was no clearcut choice between the two approaches as these criteria are affected by the SSM formulation and the available data. They noted that MCMC and SIS are similar in that one must choose a "sample generating" distribution (a proposal distribution for MCMC and an importance distribution

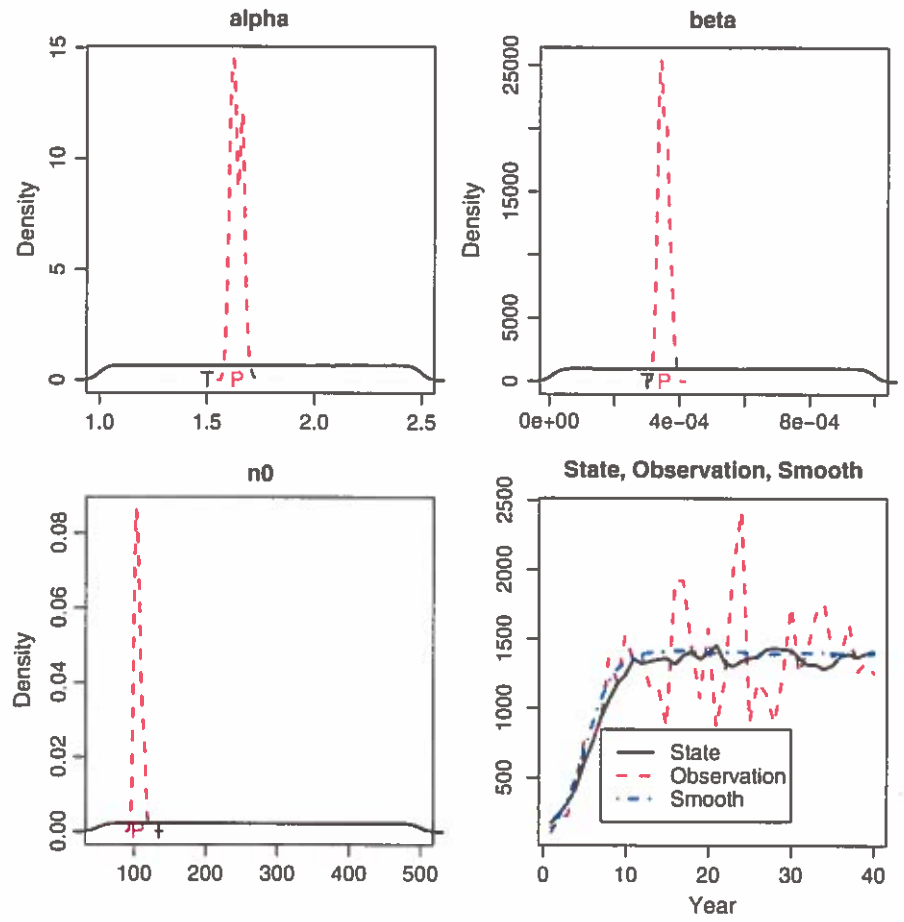


Fig. 4.5 Results from the SIS fit to the univariate coho salmon SSM with Ricker population dynamics (Poisson variation) and lognormal observations. Prior (*solid*) and posterior (*dashed*) for the parameters α, β and n_0 are shown along with marks "T" at the true values (1.50, 3.0e-4, 135) and "P" at the posterior mean values (1.64, 3.5e-4, 105). The *lower right plot* shows the fitted states, true states and estimates

for SIS), and those choices can affect accuracy of estimates as well as ease of implementation.

The fourth criterion, fidelity to reality, refers to how similar the postulated SSM is thought to be to reality. There is an interplay between selecting a model-fitting procedure and model formulation. If time to develop and execute computer code is quite limited and if one is willing to approximate a nonlinear, non-normal SSM, considered relatively realistic, with a normal dynamic linear model, considered less realistic, then maximum likelihood estimation using the Kalman filter to calculate the likelihood may be a good choice.

If there is model uncertainty and the set of models to fit is relatively large, then run execution time can be a deciding criterion and analytic fitting procedures may be preferable.

The fifth criterion, risk associated with use of the SSM, can be related to accuracy and model fidelity. If the consequences of poor decisions resulting from using a less-than-adequate SSM are high, then model fidelity and accuracy of estimates can be extremely important. Devoting considerable financial resources to code development and computational time may be preferable to the potential risk of poor management decisions. If a relatively complex realistic SSM is considered necessary, then Monte Carlo techniques may be the only choice.

The length of the time series, the length of the state vector, and the length of the observation vector, as well as the nature of the relationship between observation vector components and state vector components, can all affect accuracy of inferences about parameters and states. As time series length increases, particle depletion in some sequential Monte Carlo procedures in particular can reduce accuracy.

Issues pertinent to model formulation are discussed further in Sect. 5.1.

Chapter 5
Model Formulation and Evaluation

The previous two chapters have presented the state-space model as a general framework for modelling population dynamics and discussed alternative ways of fitting SSMs to data. In this chapter, we address model formulation and model evaluation.

How does one formulate a model, or models, for an animal population in the first place? Model construction was considered from the perspective of using matrix models as building blocks in Chap. 2. In Sect. 5.1 we take a more general look at the issue. The initial formulation of a model may be one that cannot be practically fit to data, in the sense that model parameters and state vector values cannot be estimated. Even if the model is biologically sensible, the available data might be inadequate, e.g. the state vector consists of age classes $0, 1, \ldots, 9, 10+$, but data are only available for age 0 and ages $1+$. Or there are too many intermediate sub-processes given the temporal resolution of the state vector. In Sect. 5.2, a general approach to the issue of determining model "over-parameterization", which we label parameter redundancy, is described. In Sect. 5.3, we demonstrate how parameter redundancy can be determined in state-space models.

Once one or more models have been formulated and fitted to the data, how do we determine if a model is adequate, or if one model is better than another? We consider several facets of model evaluation including model comparison and selection (Sect. 5.4), model averaging (Sect. 5.5), and individual model diagnostics (Sect. 5.6).

5.1 Model Formulation

Before formulating a model, and attempting to fit it, the purpose of modelling needs to be clearly articulated. Starfield and Bleloch (1991) generally advise that a model should be formulated to answer one or more specific questions, and to make the model no more complex than necessary. Questions thus motivate and guide

K.B. Newman et al., *Modelling Population Dynamics: Model Formulation, Fitting and Assessment using State-Space Methods*, Methods in Statistical Ecology,
DOI 10.1007/978-1-4939-0977-3_5, © Springer Science+Business Media New York 2014