State–space fish stock assessment model as alternative to (semi-) deterministic VPA approaches and full parametric stochactic models.

Anders Nielsen

November 16, 2011

Abstract

This note is a simple description of a state-space assessment model.

Motivation

Fish stock assessment models are fairly complex systems, so in order to motivate the state-space approach consider the following example: Observations Y are generated from $\lambda_0 = 0$, $\lambda_i = \lambda_{i-1} + \eta_i$, $Y_i = \lambda_i + \varepsilon_i$, where $i = 1 \ldots 50$, $\eta_i \sim N(0, \sigma_p^2)$, and $\varepsilon_i \sim N(0, \sigma_\circ^2)$ all independent. The underlying unobserved quantities λ are to be estimated.

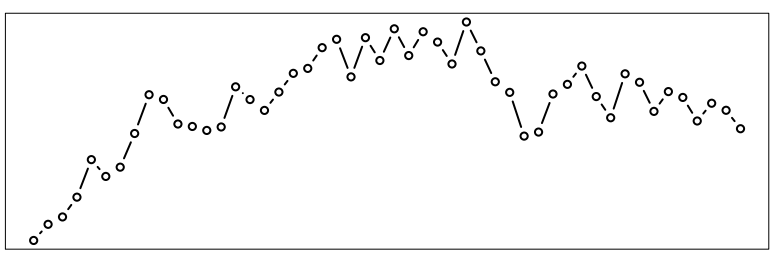

Figure 1: Simulation of a random walk with observation noise added

If we approached this system by a deterministic method (pretending zero observation noise) the logical estimator for the underlying λ_i is the corresponding observed value Y_i. This would naturally lead to a more fluctuating estimated time series than the true underlying λ if the observation noise is in fact not zero. This would not use the information fully, as it does not take advantage of the correlation between neighboring lambdas. Finally, this approach makes it impossible to quantify uncertainties in the estimated values within the model.

If we approached this system by a fully parametric statistical model we would have to add some model assumptions to make the model identifiable. One option would be to assume that $\lambda_1 = \lambda_2, \lambda_3 = \lambda_4, \ldots, \lambda_{49} = \lambda_{50}$. This pairwise coupling is naturally arbitrary, and other assumptions could have been chosen, but it illustrates trade-off we are facing. If we choose small λ-groups (here pairs) we get highly uncertain estimates, as the ratio between number of parameters and number of observations is high. If we use large λ-groups we get highly biased estimates, because the lambdas we are assuming identical are in fact very different.

The third approach presented in this section is a state-space model. In a state-space model the underlying process (here λ) is considered a random variable that is not observed. The only thing observed is a derived variable subject to measurement noise. The model parameters (here σ_p^2 and $\sigma_\circ^2$, and possibly an over all mean level) are estimated in the marginal distribution of the observations Y, and then the unobserved random variables λ are predicted via their conditional distribution given Y.

Models based on unobserved random variables are widely used in other quantitative sciences for instance agricultural studies, economics, and medical studies. Part of the reason unobserved random variables are not widely used in fisheries science is that fish stock assessment models are fairly complex. Using unobserved random variables in this setting is computer intensive, and the software tools and algorithms to make this feasible have been lagging. State-space models were introduced in fisheries by Gudmundsson (1987,1994) and Fryer (2001). Both used the extended Kalman filter to compute the likelihood. The model presented here uses new software (the random effects module for AD Model Builder), which uses a combination of automatic differentiation and the Laplace approximation (MacKay 2003) to solve high dimensional non-linear models with unobserved random effects efficiently.

Model

The model is a state–space model. The states α are the log-transformed stock sizes $\log N_1, \ldots, \log N_A$ and fishing mortalities $\log F_{i_1}, \ldots, \log F_{i_n}$ corresponding to different age classes and total international catches. In any given year y the state is the combined vector $\alpha_y = (\log N_1, \ldots, \log N_A, \log F_{i_1}, \ldots, \log F_{i_n})'$. The transition equation describes the distribution of the next years state from a given state in the current year. The following is assumed:

$$\alpha_y = T(\alpha_{y-1}) + \eta_y$$

The transition function T is where the stock equation and assumptions about stock–recruitment enters the model. The equations are:

$$\begin{aligned}
\log N_{1,y} &= \log(R(w_{1,y-1}p_{1,y-1}N_{1,y-1} + \cdots + w_{A,y-1}p_{A,y-1}N_{A,y-1})) & \\
\log N_{a,y} &= \log N_{a-1,y-1} - F^{(\cdot)}_{a-1,y-1} - M_{a-1} \ , & 2 \le a \le A \\
\log F_{a,y} &= \log F_{a,y-1} \ , & 1 \le a \le A
\end{aligned}$$

Here M_a is the age specific natural mortality parameter, which is most often assumed known from outside sources. $F^{(\cdot)}_{a-1,y-1}$ is the total fishing mortality, which include both fishing mortality from fleets with and without effort information. The function R describes the relationship between stock and recruitment. The parameters of the chosen stock–recruitment function are estimated within the model. Often it is assumed that certain F_a parameters are identical (e.g. $F_{A-1} = F_A$).

The prediction noise η is assumed to be uncorrelated Gaussian with zero mean, and three separate variance parameters. One for recruitment σ^2_R, one for survival σ^2_S, and one for the yearly development in fishing mortality σ^2_F.

This completes the description of the unobserved state process. One unique feature of this model is that the survival process is stochastic. Stock assessment methods frequently assume deterministic survival process, which means that full knowledge of N_a, M_a, and F_a in the previous year imply full knowledge of N_{a+1} in the current year. This assumption originates from historic purely deterministic assessment methods where F_a was considered equivalent to a known catch that was simply subtracted from N_a.

In fully parametric statistical stock assessment models the assumption of deterministic survival is combined with structural assumptions on the F parameters (e.g. multiplicative), which is inconsistent, as an approximated F cannot give an exactly known number of survivors. In this model F_a is considered a mortality rate, and even full knowledge of N_a, M_a, and F_a in the previous year only gives a prediction of N_{a+1} in the current year, and the uncertainty of this prediction is estimated within the model.

The observation part of the state–space model describes the distribution of the observations for a given state α_y. Here the vector of all observations from a given year y is denoted x_y. The elements of x_y are residual log-landings $\log C^{(\circ)}_{a,y}$ (which equals total landings if no other commercial fleets are present), log-catches from commercial fleets with effort data $\log C^{(f)}_{a,y}$, and log-indices from scientific surveys $\log I^{(s)}_{a,y}$. The combined observation equation is:

$$x_y = O(\alpha_y) + \varepsilon_y$$

The observation function O consists of the familiar catch equations for fleets and surveys, and ε_y of independent measurement noise with separate variance parameters for separate fleets and surveys. An expanded view of the observation equation becomes:

$$\begin{aligned}
\log C^{(\circ)}_{a,y} &= \log\left(\frac{F_{a,y}}{Z_{a,y}}(1-e^{-Z_{a,y}})N_{a,y}\right) + \varepsilon^{(\circ)}_{a,y} \\
\log C^{(f)}_{a,y} &= \log\left(\frac{E^{(f)}_y Q^{(f)}_a}{Z_{a,y}}(1-e^{-Z_{a,y}})N_{a,y}\right) + \varepsilon^{(f)}_{a,y} \\
\log I^{(s)}_{a,y} &= \log\left(Q^{(s)}_a e^{-Z_{a,y}\frac{D^{(s)}}{365}} N_{a,y}\right) + \varepsilon^{(s)}_{a,y}
\end{aligned}$$

Here Z is the total mortality rate $Z_{a,y} = M_a + F_{a,y} + \sum_f E^{(f)}_y Q^{(f)}_a$, $D^{(s)}$ is the number of days into the year where the survey s is conducted, $Q^{(f)}_a$ and $Q^{(s)}_a$ are model parameters describing catchabilities. Finally $\varepsilon^{(\circ)}_{a,y} \sim N(0,\sigma^2_\circ)$, $\varepsilon^{(f)}_{a,y} \sim N(0,\sigma^2_f)$, and $\varepsilon^{(s)}_{a,y} \sim N(0,\sigma^2_s)$ are all assumed independent.

The likelihood function for this is set up by first defining the joint likelihood of both random effects (here collected in the α_y states), and the observations (here collected in the x_y vectors). The joint likelihood is:

$$L(\theta,\alpha,x) = \prod_{y=2}^{Y}\{\phi(\alpha_y - T(\alpha_{y-1}),\Sigma_\eta)\}\prod_{y=1}^{Y}\{\phi(x_y - O(\alpha_y),\Sigma_\varepsilon)\}$$

Here θ is a vector of model parameters. Since the random effects α are not observed inference should be obtain from the marginal likelihood:

$$L_M(\theta,x) = \int L(\theta,\alpha,x)d\alpha$$

This integral is difficult to calculate directly, so the Laplace approximation is used. The Laplace approximation is derived by first approximating the joint log likelihood $\ell(\theta,\alpha,x)$ by a second order Taylor approximation around the optimum $\hat{\alpha}$ w.r.t. α. The resulting approximated joint log likelihood can then be integrated by recognizing it as a constant term and a term where the integral is know as the normalizing constant from a multivariate Gaussian. The approximation becomes:

$$\int L(\theta,\alpha,Y)d\alpha \approx \sqrt{\frac{(2\pi)^n}{\det(-\ell''_{\alpha\alpha}(\theta,\alpha,Y)|_{\alpha=\hat{\alpha}_\theta})}}\exp(\ell(\theta,\hat{\alpha}_\theta,Y))$$

Taking the logarithm gives the Laplace approximation of the marginal log likelihood

$$\ell_M(\theta, Y) = \ell(\theta, \hat{u}_\theta, Y) - \frac{1}{2}\log(\det(-\ell''_{uu}(\theta, u, Y)|_{u=\hat{u}_\theta})) + \frac{n}{2}\log(2\pi)$$

References

Fryer, R. (2001). TSA: is it the way? ICES Working document for the Working Group on Methods of Fish Stock Assessment.

Gudmundsson, G. (1987). Time series models of fishing mortality rates. ICES C.M. **d:6**.

Gudmundsson, G. (1994). Time series analysis of catch-at-age observations. Appl. Statist. **43**:117-126.

MacKay, D. J. C. (2003). Information Theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge.