Dissecting SAM 2: Observations

Anders Nielsen, Ethan Lawler, & Sean Anderson

an@aqua.dtu.dk

The parts of SAM

Processes:

- The three main processes are: Recruitment $(N_{1,y})$, survival $(N_{>1,y})$, fishing $(F_{a,y})$.
- These are treated as unobserved random effects in the model
- The processes describe the development of the system we are monitoring
- Observations related to the system are used to predict these processes

Observations:

- Anything we can observe, which can help to inform about the processes
- Common options are catch-at-age $C_{a,y}$, survey index-at-age $I_{a,y}$, total catches, biomass index, tagging, lengths ...
- From the process (and a few estimated model parameters) we should be able to predict the observations

Parameters:

- Fixed effects model parameters to be estimates
- E.g. catchabilities, variance parameters, and stock-recruitment parameters.

Here we will look at the observation part

Catch-at-age

- Once the processes (recruitment, survival, and fishing) are set up we can predict observations
- The catch-at-age is predicted by the catch equation:

$$\widehat{C}_{ay} = \frac{F_{ay}}{F_{ay} + M_{ay}} N_{ay}(1 - e^{-F_{ay} - M_{ay}})$$

- N is stocksize, F is fishing mortality, M is natural mortality. Often we refer to total mortality rate as: $Z_{ay} = F_{ay} + M_{ay}$
- Taking the logarithm:

$$\log \widehat{C}_{a,y} = \log F_{ay} - \log Z_{ay} + \log N_{ay} + \log(1 - e^{-Z_{ay}})$$

- The simplest model we can construct for this is that the log of the catch follows an independent normal distribution
- we can write it as:

$$\log C_{ay} \sim \mathcal{N}(\log \widehat{C}_{a,y}, \sigma^2)$$

- or like

$$\log C_{a,y} = \log F_{ay} - \log Z_{ay} + \log N_{ay} + \log(1 - e^{-Z_{ay}}) + \varepsilon_{ay} \ , \quad \text{where } \varepsilon_{ay} \sim \mathcal{N}(0, \sigma^2)$$

- or like

$$C_{ay} = \frac{F_{ay}}{Z_{ay}} N_{ay}(1 - e^{-Z_{ay}}) e^{\varepsilon_{ay}} \ , \quad \text{where } \varepsilon_{ay} \sim \mathcal{N}(0, \sigma^2)$$

Exercise: Catch observations

- The data list in `Cobs.RData` contains a vector `Cobs` of the catch observations. It is a vector not a matrix.
- It may seem like an artificial way to represent the observations, but it will be useful later.
- In addition the data list contains a matrix `aux` with three columns (`year`, `fleet`, `age`). The i'th row in this matrix contains the information for the i'th element in the `Cobs` vector.
- Further the data list contains information on `F`, `N`, `M`, `minYear`, and `minAge`.
- Use this data to implement the model:

$$\log C_{ay} \sim \mathcal{N}(\log \widehat{C}_{a,y}, \sigma^2)$$

- Estimate the only model parameter σ
- What if we wanted separate standard deviations for some age groups?

Survey fleets

- A survey fleet produces an index-at-age I_{ay}, which we can model in a similar way to catches
- The surveys are often taken in a short time interval, and we use them as proportional to stock size at that time
- The proportionality coefficient q is expected to be time invariant, exactly because the survey aims to produce an index
- The survey indices are predicted by:
$$\widehat{I}_{ay}^{(s)} = q_a^{(s)} N_{ay} e^{-\tau^{(s)} Z_{ay}}$$
- Here $\tau^{(s)}$ it the time into the year the survey is conducted
- A first model could be:
$$\log I_{ay}^{(s)} \sim \mathcal{N}(\log \widehat{I}_{ay}^{(s)}, \sigma_s^2)$$
- Since surveys are collected over a smaller time period they are sometimes affected e.g. by bad weather, or by a few large catch events of possibly similar fish
- It can be necessary to include some correlation structure and sometimes observations can be missing.

Exercise: Adding two survey fleets

- The data list in `allfleets.RData` contains a vector `obs` with all observations (catches and surveys)
- In addition the data list contains a matrix `aux` with three columns (`year`, `fleet`, `age`). The i'th row in this matrix contains the information for the i'th element in the `obs` vector.
- Further the data list contains information on `F`, `N`, `M`, `minYear`, `minAge`
- There are three fleets in this entire data set catch and two surveys.
- Data also contain a vector `fleetTypes` with three elements. '0' indicates a catch fleet and '2' indicates a survey fleet.
- Finally the data contains a vector `sampleTimes` with three elements, which is the $\tau^{(s)}$ values (only used for the survey fleets).
- The exercise is to extend the previous exercise to also add the survey model, but we will introduce a few tricks along the way.

Handling missing observations

- Some of the observations in the `obs` vector are missing `NA`
- We could add code to simply avoid adding these to the likelihood, but that becomes problematic later when we need to work with multivatiate distributions.
- Instead we can substute them in as random effects
- We can add the random effects to the parameter list as:

```
par$missing <- numeric(sum(is.na(allfleets$obs)))  ## count them
```

- Then in the function we can use them where observations are missing

```
logobs <- log(obs)
logobs[is.na(logobs)] <- missing
```

- Finally we remember to declare them as random:

```
obj <- MakeADFun(f, par, random="missing")
```

- The rest of the program is unchanged.
- Then the model can work where observations were missing and even produce predictions of the missing (if we should need it).

Configuring parameters

- This trick could possibly be replaced by `map` in RTMB, but SAM uses this approach, and I find it very flexible
- In this data set we have 3 fleets and 9 ages, not all fleets have all ages
- If we define a data matrix like this:

```
allfleets$keyQ <- rbind(c(NA,NA,NA,NA,NA,NA,NA,NA,NA),
                        c(NA, 1, 2, 3, 4, 5, 6, 7,NA),
                        c( 8, 9,10,11,12,13,NA,NA,NA))
```

- and a parameter vector like this:

```
par$logQ <- numeric(max(allfleets$keyQ, na.rm=TRUE))
```

- Then in the function we can use this table to look up which model parameter we should use for a given fleet f and for a age a. This can be done like:

```
logPred[i] <- logQ[keyQ[f,a]]+log(N[y,a])-Z*sampleTimes[f]
```

- Notice that this can also be used to use the same parameter for multiple ages.

- Extend the previous exercise with the model for the two survey fleets:

$$\log I_{ay}^{(s)} \sim \mathcal{N}(\log \widehat{I}_{ay}^{(s)}, \sigma_s^2)$$

- Estimate the catchabilities and standard deviation parameters for the surveys
- Make a plot to convince yourself that it worked correctly.

Blocking observations

- The file `allfleetsblock.R` we carry out exactly the same calculations
- but blocked by each fleet within each year
- Hereby we prepare the code, so we can use different correlation structures
- If we need to use multivariate distributions (e.g. multivariate normal or multinomial), then we need to be able to pick out these blocks.
- Let's study the code for a blocked version of the program

Exercise: Adding covariance structure

- Add an AR(1) covariance structure across age to the survey fleets
- Why is it important to get the covariance structure right?

Irregular grid AR

- In the regular AR structure the covariance is defined as:

$$\Sigma_{ij} = \rho^{|i-j|}\sqrt{\Sigma_{ii}\Sigma_{jj}}$$

- So correlation only depende on distance between i and j, not which i and j.
- First realize that we can get the same covariance structure by:

$$\Sigma_{ij} = 0.5^{\alpha|i-j|}\sqrt{\Sigma_{ii}\Sigma_{jj}} \quad , \quad \text{where} \quad \alpha > 0$$

- Notice that this implies a regular grid.
- We can extend this structure by defining

$$\Sigma_{ij} = 0.5^{|\theta_i-\theta_j|}\sqrt{\Sigma_{ii}\Sigma_{jj}} \quad , \quad \text{where} \quad \theta_1 = 0 \leq \theta_2 \leq \cdots \leq \theta_A$$

- This corresponds to having the points on an irregular grid.
- How would we parametrize this?
- If all deltas are the same, then it is a regular AR structure
- Let's study the code in `igar.*`

Unstructured covariance

- The fully unstructured covariance can be constructed in the following way.

$$\Sigma_{ij} = (D^{-\frac{1}{2}} L L^t D^{-\frac{1}{2}})_{ij} \sqrt{\Sigma_{ii}\Sigma_{jj}}$$

- Here L is a lower triangle matrix (Cholesky of the correlation) and D is the diagonal matrix of (LL^t)

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \theta_1 & 1 & 0 \\ \theta_2 & \theta_3 & 1 \end{pmatrix}$$

- The model parameters are the elements in L and the log-standard deviations
- This is very flexible, but also requires a lot of parameters to be estimated
- It is relative simple to implement, because much of the work is done by TMB
- Let's study the code in `us.*`
- Now we have a lot of options (ID, AR, IGAR, US) for these three fleets
- Try to configure a few of the options.

- How can we go about choosing an optimal configuration?

That's SAM

- Now we have covered all parts of a standard SAM assessment.
- All processes: recruitment, survival, and fishing
- The standard data sources: catches and surveys
- Importantly the covariance structures
- It should be a small task to stitch it together