Dissecting SAM 2: Observations

Anders Nielsen & Olav Nikolai Breivik
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

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 sometimed 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
- On the R-side we can add the random effects as:

```
par$missing <- numeric(sum(is.na(allfleets$obs)))  ## count them
obj <- MakeADFun(allfleets, par, random="missing", DLL="allfleets")
```

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

```
int idxmis=0;
for(int i=0;i<nobs;i++){
  if(isNA(obs(i))){
    obs(i)=exp(missing(idxmis++));
  }
}
```

- 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).
- The `isNA` helper function is defined on the next page

Small helper function to test for missing values

```
template<class Type>
bool isNA(Type x){
  return R_IsNA(asDouble(x));
}
```

To be pasted in right below the line `#include <TMB.hpp>`

Configuring parameters

- This trick could possibly be replaced by `map` in TMB, 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 an integer data matrix like this:

```
allfleets$keyQ <- rbind(c(NA,NA,NA,NA,NA,NA,NA,NA,NA),
                        c(NA, 0, 1, 2, 3, 4, 5, 6,NA),
                        c( 7, 8, 9,10,11,12,NA,NA,NA))
```

- and a parameter vector like this:

```
par$logQ <- numeric(max(allfleets$keyQ, na.rm=TRUE)+1)
```

- Then in the C-code we can use the 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:

```
case 2:
  logPred(i) = logQ(keyQ(f,a))+log(N(y,a))-Z*sampleTimes(f);
break;
```

- 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 data list in `allfleetsblock.RData` contains two additional matrices `idx1` and `idx2`.
- these have a row per fleet and a column year.
- `idx1`(f,y) is the index of the first observation from fleet f in year y
- `idx2`(f,y) is the index of the last observation from fleet f in year y
- The observations are sorted acordingly (year, fleet, age), so these two define the vector of observations from fleet f in year y.
- 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 from the last exercise.

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 constructured 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?

Configuration options for observations in SAM

```
$maxAgePlusGroup
# Is last age group considered a plus group for each fleet (1 yes, or 0 no).
 1 0 0
```

```
$keyLogFpar
# Coupling of the survey catchability parameters (nomally first row is not used, as that is covered by fishing
     mortality).
  -1   -1   -1   -1   -1   -1
   0    1    2    3    4   -1
   5    6    7    8   -1   -1
```

```
$keyVarObs
# Coupling of the variance parameters for the observations.
   0    1    2    2    2    2
   3    4    4    4    4   -1
   5    6    6    6   -1   -1
```

```
$obsCorStruct
# Covariance structure for each fleet ("ID" independent, "AR" AR(1), or "US" for unstructured). | Possible
     values are: "ID" "AR" "US"
 "ID" "ID" "ID"

$keyCorObs
# Coupling of correlation parameters can only be specified if the AR(1) structure is chosen above.
# NA's indicate where correlation parameters can be specified (-1 where they cannot).
#1-2 2-3 3-4 4-5 5-6
  NA   NA   NA   NA   NA
  NA   NA   NA   NA   -1
  NA   NA   NA   -1   -1
```

```
$fracMixObs
# A vector with same length as number of fleets, where each element is the fraction of t(3) distribution used
    in the distribution of that fleet
 0 0 0
```

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