Coverpage

```
## Document build date: Tue Dec 20 14:55:52 2016
## Working directory :
##       D:/projects/flr/drafting-doc
## Current contents of .GlobalEnv:
##      .First .Last thm
##
## Session information:
## R version 3.3.2 (2016-10-31)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 10586)
##
## locale:
## [1] LC_COLLATE=English_United Kingdom.1252
## [2] LC_CTYPE=English_United Kingdom.1252
## [3] LC_MONETARY=English_United Kingdom.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United Kingdom.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] xtable_1.8-2       lattice_0.20-34    RColorBrewer_1.1-2
## [4] formatR_1.4        knitr_1.15.1
##
## loaded via a namespace (and not attached):
## [1] magrittr_1.5  tools_3.3.2   stringi_1.1.2 grid_3.3.2    digest_0.6.10
## [6] stringr_1.1.0 evaluate_0.10
```

Description of Catch at Age model

Colin P. Millar, Ernesto Jardim

Abstract

This document presents the statistical catch-at-age stock assessment model developed in the JRC Assessment For All (`a4a`) initiative. The stock assessment model framework is a non-linear catch-at-age model implemented in R `http://www.r-project.org/` / FLR `http://www.flr-project.org/` / ADMB `http://www.admb-project.org/` that can be applied rapidly to a wide range of situations with low parametrization requirements. The model structure is defined by submodels, which are the different parts that require structural assumptions. There are 5 submodels in operation: a model for F-at-age, a model for the initial age structure, a model for recruitment, a (list) of model(s) for abundance indices catchability-at-age, and a list of models for the observation variance of catch-at-age and abundance indices. The submodels form use linear models. This opens the possibility of using the linear modelling tools available in R: see for example the mgcv `http://cran.r-project.org/web/packages/mgcv/index.html` gam formulas, or factorial design formulas using. Detailed model formulas, several diagnostic tools and a large set of models are presented in the document. Additionaly, advanced features like external weighting of the likelihood components and MCMC fits are also described. The target audience for this document are readers with some experience in R and some background on stock assessment. The document explains the approach being developed by a4a for fish stock assessment and scientific advice. It presents a mixture of text and code, where the first explains the concepts behind the methods, while the last shows how these can be run with the software provided.

Contents

1 Background

The stock assessment model framework is a non-linear catch-at-age model implemented in R/FLR/ADMB that can be applied rapidly to a wide range of situations with low parametrization requirements.

In the `a4a` assessment model, the model structure is defined by submodels, which are the different parts of a statistical catch at age model that require structural assumptions.

There are 5 submodels in operation:

- a model for F-at-age,
- a (list) of model(s) for abundance indices catchability-at-age,
- a model for recruitment,
- a list of models for the observation variance of catch-at-age and abundance indices,
- a model for the initial age structure,

In practice, we fix the variance models and the initial age structure models, but in theory these can be changed.

The submodels form use linear models. This opens the possibility of using the linear modelling tools available in R: see for example the mgcv gam formulas, or factorial design formulas using `lm()`. In R's linear modelling language, a constant model is coded as ~ 1, while a slope over age would simply be $\sim age$. For example, we can write a traditional year/age separable F model like $\sim factor(age) + factor(year)$.

The 'language' of linear models has been developing within the statistical community for many years, and constitutes an elegant way of defining models without going through the complexity of mathematical representations. This approach makes it also easier to communicate among scientists

- 1965 J. A. Nelder, notation for randomized block design
- 1973 Wilkinson and Rodgers, symbolic description for factorial designs
- 1990 Hastie and Tibshirani, introduced notation for smoothers
- 1991 Chambers and Hastie, further developed for use in S

There are two basic types of assessments available in `a4a` : the management procedure fit and the full assessment fit. The management procedure fit does not compute estimates of covariances and is therefore quicker to execute, while the full assessment fit returns parameter estimates and their covariances at the expense of longer fitting time.

2 The data and a simple model

The data are

C_{at} catch at age a and year t

S_{atk} abundance index for age a and year t from the kth survey or CPUE series, $k = 1, 2, \dots$

The model is an age structure model where the number of fish in a given cohort N at the start of the following year is the number of fish that survived the perils of the current year. We assume that fish die through the year at a constant rate e^{-Z} (Z is positive), and that this rate is solely due to natural causes (M) and fishing (F) so that the total mortality rate is $Z = F + M$. This results in the model

$$N_{a+1,t+1} = N_{at}e^{-Z_{at}}$$

Abundance indices are observations of the relative abundance not of absolute abundance. This is because trawl surveys do not detect every fish but a fixed proportion Q. This proportion depends on age through length and means the index is proportional to abundance

$$S_{at} = Q_a N_{at}$$

If F and M are constant through the year catches arise as a fraction of those fish that died, and is written here as the familiar Baranov catch equation

$$\begin{aligned} C_{at} &= \frac{F_{at}}{Z_{at}} \left(N_{at} - N_{a+1,t+1} \right) \\ &= \frac{F_{at}}{Z_{at}} \left(1 - e^{-Z_{at}} \right) N_{at} \end{aligned}$$

These last two equations show that in there own way, catches and abundance indices are both observations of the numbers of fish in the population. Neither is sufficient to estimate the absolute abundances N but together they can be used to estimate both N and F. One way of doing this is using a statistical catch at age approach

$$c_{at} \sim N\left(\log\left(\frac{F_{at}}{Z_{at}} \left(1 - e^{-Z_{at}} \right) \right) + n_{at}, \quad \sigma_c^2 \right)$$

and

$$s_{at} \sim N\left(q_a + n_{at}, \quad \sigma_s^2 \right)$$

where in these equations we have written logs in lower case i.e. $c = \log C$.

3 a4a Model details

Modelled catches C are defined in terms of the three quantities, natural mortality M, fishing mortality F and recruitment R, using a modified form of the well known Baranov catch equation:

$$C_{ay} = \frac{\boldsymbol{F}_{ay}}{\boldsymbol{F}_{ay} + M_{ay}} \left(1 - e^{-(\boldsymbol{F}_{ay}+M_{ay})}\right) \boldsymbol{R}_y e^{-\sum(\boldsymbol{F}_{ay}+M_{ay})}$$

where a and y denote age and year. Modelled survey indices I are defined in terms of the same three quantities with the addition of survey catchability Q:

$$I_{ays} = \boldsymbol{Q}_{ays} \boldsymbol{R}_y e^{-\sum(\boldsymbol{F}_{ay}+M_{ay})}$$

where s denotes survey or abundance index and allows for multiple surveys to be considered. Observed catches $C^{(obs)}$ and the observed survey indices $I^{(obs)}$ are assumed to be log-normally distributed, or equivalently, normally distributed on the log-scale, with age, year and survey specific observation variance:

$$\log C^{(obs)}_{ay} \sim \text{Normal}\Big(\log C_{ay}, \boldsymbol{\sigma}^2_{ay}\Big) \qquad \log I^{(obs)}_{ays} \sim \text{Normal}\Big(\log I_{ays}, \boldsymbol{\tau}^2_{ays}\Big)$$

The full log-likelihood for the a4a statistical catch at age model can now be defined as the sum of the log-likelihood of the observed catches (ℓ_N is the log-likelihood of a normal distribution)

$$\ell_C = \sum_{ay} w^{(c)}_{ay}\ \ell_N\Big(\log C_{ay}, \boldsymbol{\sigma}^2_{ay};\ \log C^{(obs)}_{ay}\Big)$$

and the log-likelihood of the observed survey indices

$$\ell_I = \sum_s \sum_{ay} w^{(s)}_{ays}\ \ell_N\Big(\log I_{ays}, \boldsymbol{\tau}^2_{ays};\ \log I^{(obs)}_{ays}\Big)$$

giving the total log-likelihood

$$\ell = \ell_C + \ell_I$$

which is defined in terms of the strictly positive quantites, M_{ay}, F_{ay}, Q_{ays} and R_y, and the observation variances σ_{ay} and τ_{ays}. As such, the log-likelihood is over-parameterised as there are many more parameters than observations. In order to reduce the number of parameters, M_{ay} is assumed known (as is common), and the remaining parameters are written in terms of a linear combination of covariates x_{ayk}, e.g.

$$\log F_{ay} = \sum_k \beta_k x_{ayk}$$

where k is the number of parameters to be estimated and is sufficiently small. Using this tecnique the quantities $\log F$, $\log Q$, $\log \sigma$ and $\log \tau$ (in bold in the equations above) can be described by a reduced number of parameters. The following section has more discussion on the use of linear models in a4a .

Stock recruitment relationships

The a4a statistical catch at age model can addiionally allow for a functional relationship to be imposed that links predicted recruitment $\tilde{R}$ based on spawning stock biomass and modelled recruitment R, included as a fixed variance random effect. Options for the relationship are the hard coded models Ricker, Beverton Holt, smooth hockeystick or geometric mean. This is implemented by including a third component in the log-likelihood

$$\ell_{SR} = \sum_y \ell_N\Big(\log \tilde{R}_y(a, b), \phi^2_y;\ \log R_y\Big)$$

giving the total log-likelihood

$$\ell = \ell_C + \ell_I + \ell_{SR}$$

Using the (time varying) Ricker model as an example, predicted recruitment is

$$\tilde{R}_y(a_y, b_y) = a_y S_{y-1} e^{-b_y S_{y-1}}$$

where S is spawning stock biomass derived from the model parameters F and R, and the fixed quantites M and mean weights by year and age. It is assumed that R is log-normally distributed, or equivalently, normally distributed on the log-scale about the (log) recruitment predicted by the SR model $\tilde{R}$, with known variance ϕ^2, i.e.

$$\log R_y \sim \text{Normal}\Big(\log \tilde{R}_y, \phi_y^2\Big)$$

which leads to the definition of ℓ_{SR} given above. In all cases a and b are strictly positive, and with the quantities F, R, etc. linear models are used to parameterise $\log a$ and/or $\log b$, where relevant.

By default, recruitment R as apposed to the reruitment predicted from a stock recruitment model $\tilde{R}$, is specified as a linear model with a parameter for each year, i.e.

$$\log R_y = \gamma_y$$

This is to allow modelled recruitment R_y to be shrunk towards the stock recruitment model. However, if it is considered appropriate that recruitment can be determined exactly by a relationship with covariates, it is possible, to instead define $\log R$ in terms of a linear model in the same way as $\log F$, $\log Q$, $\log \sigma$ and $\log \tau$.

Model fitting

Model fitting is done by optimising the combined likelihood (above) in ADMB.

4 Implementation

We require two functions that return an objective

A inputs are F at age and observation error and arguments are the data and the hat matrix

B inputs are F at age and any variance parameters taking as arguments the design matrix H, the weight matrix W and the structural prior matrix (not mentioned yet but lets call it Q)

The full objective function is then

1. take input parameters (F pars, variances, recruitments (if SRR model being used))
2. convert F pars into F at age
3. calculate objective value using one of the two functions A or B above
4. add on SRR density and prior densities for variances if necessary

$$N_{at} = \begin{cases} R_t & \text{if} \quad a = 1 \\ R_{t-a+1} e^{-\sum_{i=1}^{a-1} Z_{a-i,t-i}} & \text{if} \quad a > 1 \end{cases} \tag{1}$$

The data are assumed to be gaussian observations of numbers at age in the population, which itself is generated by the common age structured model

$$n_{at} = \begin{cases} r_t & \text{if} \quad a = 1 \\ r_t - \sum_{i=1}^{a-1} Z_{a-i,t-i} & \text{if} \quad a > 1 \end{cases} \tag{2}$$

where $Z_{at} = F_{at} + M_{at}$, where M_{at} is known and F_{at} is modelled as a seperable function

$$\log F_{at} = \gamma_a + \delta_t \tag{3}$$

with suitable constraints. The observation equations are

$$c_{at} \sim N\left(\log\left(\frac{F_{at}}{Z_{at}}\left(1 - e^{-Z_{at}}\right)\right) + n_{at}, \quad \kappa \right) \tag{4}$$

and

$$s_{atk} \sim N\left(q_{ak} + n_{at}, \quad \tau_k \right) \tag{5}$$

The parameters to be estimated in this model are: log recruitment r_t, F at age and year, F_{at}, log survey catchability q_{ak}, and the precisions $\theta = (\kappa, \tau_1, \ldots)$.

5 Extending the model

First potential models for q_a. Linear forms can be included my simply changing the design matrices. By linear forms i mean spline smoothers with fixed degrees of freedom. An interesting addition would be to use penalised splines or better (i think) structured random effects

The model for F has not been considered so far. Options for this are a simple separable model, a model with several separable periods all these models can be expressed as linear models on the log link. Within the same framework is simple then to include smoothers (splines) with fixed degrees of freedom. More interesting but perhaps out with the scope of this project are structured random effect models for F. These include seasonal models (treating the number of ages as the season length) and correlated random walks.

6 Summary