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 a decription of 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/ and 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 are formulated through 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 and useful overview of R formulas here http://science.nature.nps.gov/im/datamgmt/statistics/r/formulas/.

Contents

1 Background

The `a4a` stock assessment model 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 index 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 are formulated through the use of 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()`, a usefull overview of model formulas in R can be found here `http://science.nature.nps.gov/im/datamgmt/statistics/r/formulas/`. 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 helps to improve communication 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

2 The basic approach

The data required to fit the `a4a` stock assessment model are observations of catch $C_{ay}^{(\text{obs})}$ for age $a = a_0, a_0+1, \ldots$ and year $y = y_0, y_0+1, \ldots$, and observations of abundance indices $I_{ays}^{(\text{obs})}$ for age a and year y from the sth survey or CPUE series, $s = 1, 2, \ldots$

The model is an age structure model where the number of fish at age a at the start of year y is N_{ay} are assumed to die through the year at a constant rate given by $e^{-Z_{ay}}$, where Z_{ay} is always positive, and that this rate is solely due to natural causes M_{ay} and fishing F_{ay}. At the start of the following year $y+1$ the number of fish $N_{a+1,y+1}$ is the number of fish, 1 year older, that survived the perils of year y. This results in an expression that describes the simplest type of population dynamics in the `a4a` model

$$N_{a+1,y+1} = N_{ay} e^{-Z_{ay}} \tag{1}$$

And to initialise this population, it requires a vector of numbers of recruiting fish $R_y = N_{a_0 y}$, $y = y_0, y_0+1, \ldots$ and a vector of the age structure N_{ay_0}, $a = a_0+1, a_0+2, \ldots$ in year y_0. These two vectors, along with an estimate of mortality Z_{ay} are what is required to generate a model of the fish population, and are refered to as recruitment and initial-age-structure.

Unfortuntely, recruitment and the initial population structure are not directly observed, but the modelled numbers at age can be used to generate predictions that are directly observed in the form of survey indices and commercial catches. Abundance indices are (in most cases) observations of the relative abundance because they do not detect every fish but rather a fixed proportion Q_{ays} that can depend on age and year, and be survey specific. This model of how the survey index relates to the numbers of fish in the population, allows a prediction of the survey indices I_{ays} to be made, which can be compared to the observed survey indices $I_{ays}^{(\text{obs})}$.

$$I_{ays} = Q_{ays} N_{ay} \tag{2}$$

As mentioned, it is also nessisary to estimate the mortality rate in the population. Usually, only mortality due to fishing is observed, and so natural mortality M_{ay} is assumed to be known, and is set to a sensible value, guided by expert judgement. Fishing mortality F_{ay}, on the other hand is observed via the numbers removed through fishing. Because F_{ay} and M_{ay} are constant through the year, catches arise as a fraction of those fish that died $N_{ay} - N_{a+1,y+1}$, and is written here as the familiar Baranov catch equation. Note the second line arises by substituting the population equation (1) for $N_{a+1,y+1}$

$$\begin{aligned} C_{ay} &= \frac{F_{ay}}{Z_{ay}} \left(N_{ay} - N_{a+1,y+1}\right) \\ &= \frac{F_{ay}}{Z_{ay}} \left(1 - e^{-Z_{ay}}\right) N_{ay} \end{aligned} \tag{3}$$

This equation, like (2) gives predictions based on the population model and provides a value of catches which can be compared to the observed catches $C_{ay}^{(\text{obs})}$. Taken together, the comparison of the observations of catches and survey index to their predictions is the basic approach for estimating the numbers of fish in the population and the fishing mortality rate in `a4a` .

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:

$$C_{ay} = \frac{\boldsymbol{F}_{ay}}{\boldsymbol{F}_{ay} + M_{ay}} \left(1 - e^{-(\boldsymbol{F}_{ay} + M_{ay})}\right) \boldsymbol{R}_y e^{-\sum(\boldsymbol{F}_{ay} + M_{ay})} \tag{4}$$

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:

$$I_{ays} = \boldsymbol{Q}_{ays} \boldsymbol{R}_y e^{-\sum(\boldsymbol{F}_{ay} + M_{ay})} \tag{5}$$

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:

$$\log C_{ay}^{(obs)} \sim \text{Normal}\Big(\log C_{ay}, \boldsymbol{\sigma}_{ay}^2\Big) \qquad \log I_{ays}^{(obs)} \sim \text{Normal}\Big(\log I_{ays}, \boldsymbol{\tau}_{ays}^2\Big) \tag{6}$$

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)

$$\ell_C = \sum_{ay} w_{ay}^{(c)} \; \ell_N\Big(\log C_{ay}, \boldsymbol{\sigma}_{ay}^2;\; \log C_{ay}^{(obs)}\Big) \tag{7}$$

and the log-likelihood of the observed survey indices

$$\ell_I = \sum_{s} \sum_{ay} w_{ays}^{(s)} \; \ell_N\Big(\log I_{ays}, \boldsymbol{\tau}_{ays}^2;\; \log I_{ays}^{(obs)}\Big) \tag{8}$$

giving the total log-likelihood

$$\ell = \ell_C + \ell_I \tag{9}$$

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.

$$\log F_{ay} = \sum_k \beta_k x_{ayk} \tag{10}$$

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

$$\ell_{SR} = \sum_y \ell_N\Big(\log \tilde{R}_y(a,b), \phi_y^2;\ \log R_y\Big) \tag{11}$$

giving the total log-likelihood

$$\ell = \ell_C + \ell_I + \ell_{SR} \tag{12}$$

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

$$\tilde{R}_y(a_y, b_y) = a_y S_{y-1} e^{-b_y S_{y-1}} \tag{13}$$

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.

$$\log R_y \sim \text{Normal}\Big(\log \tilde{R}_y, \phi_y^2\Big) \tag{14}$$

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.

$$\log R_y = \gamma_y \tag{15}$$

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 (9) or (12) 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

$$N_{at} = \begin{cases} R_t & \text{if} \quad a = 1 \\ R_{t-a+1} e^{-\sum_{i=1}^{a-1} Z_{a-i,t-i}} & \text{if} \quad a > 1 \end{cases} \tag{16}$$

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

$$n_{at} = \begin{cases} r_t & \text{if} \quad a = 1 \\ r_t - \sum_{i=1}^{a-1} Z_{a-i,t-i} & \text{if} \quad a > 1 \end{cases} \tag{17}$$

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

$$\log F_{at} = \gamma_a + \delta_t \tag{18}$$

with suitable constraints. The observation equations are

$$c_{at} \sim N\left(\log\left(\frac{F_{at}}{Z_{at}} \left(1 - e^{-Z_{at}}\right) \right) + n_{at}, \quad \kappa \right) \tag{19}$$

and

$$s_{atk} \sim N\left(q_{ak} + n_{at}, \quad \tau_k \right) \tag{20}$$

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. Structured random effect models using GMRFs require a Bayesian approach. If the variance of these is assumed known then these random effects can be integrated out directly, but if the variance is to be estimated it must be done in the outside iteration along with the observation error and the F parameters.

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