Data input and stepwise model selection in R for the Punt et al. (2008) ageing error model

James T. Thorson, Ian Stewart, and André E. Punt

Technical document for the Northwest Fisheries Science Center

Running the Punt et al. (2008) model in R

Function name: RunFn()

Background:

The Punt et al. (2008) model calculates the likelihood of model parameters given an observed dataset that includes age reads (henceforth "reads") provided by multiple readers for a set of otoliths. For each reader, two sets of parameters are estimated that define the standard deviation and bias of the reads provided by that reader. Specifically, the model has parameters that approximate the expected age of each read given the true age of an otolith, and the standard deviation of a normally distributed reading error given the true age of an otolith. Each of these functional forms can be either linear or curvilinear, and each is conditioned on an unobserved "True" age for each otolith. This "True" age for each otolith can be considered a random-effect, and the software computes the resulting likelihood while summing across all possible discrete values for this "True" age for each otolith.

This summation across all possible values for a "True" age for each otolith also requires a hyperdistribution representing the "prior" probability that an otolith is any given age; this prior is parameterized using a set of hyperparameters in addition to the parameters that govern the SD and bias for each reader. Specifically, one hyperparameter is estimated for every age between (and including) a MinusAge and a PlusAge, which are defined exogenously for every model run. Ages above the PlusAge or below the MinusAge have a prior Proportion-at-Age defined as a loglinear deviation from the Proportion-at-Age for the PlusAge and MinusAge. The slope of these loglinear deviations thus constitutes an additional 1 or 2 fixed effect parameter to estimate. The "True" proportion-at-age is then calculated from these fixed effect and log-linear slope parameters by normalizing the resulting distribution so that it sums to one.

Necessary Inputs:

1. *Format data*: Data should be formatted with unique reading records as rows and readers/labs as columns (exampling in Table 1). Specifically, each column corresponds to a reader, readers, lab or labs with a unique reading error and bias; the Punt (2008) model allows for approximately 15 unique columns, so the number of "readers" must be less than this. Additionally, an additional column inserted on the left-hand side of the data matrix indicates the number of otoliths with that unique read record; this cell is generally "1", but any instances where two or more otoliths have identical

reads for all readers are combined and this cell is incremented. Any missing entries (i.e., where a reader has not read anything for a given otolith) are indicated with a "-999" in that cell. The model can be configured such that a given column (i.e. reader) has parameter values that "mirror" the parameter values for a reader to it's left. This can allow estimation of a model where readers within the same lab are estimated to have the same reading error and bias. Any instance where a particular reader (or lab) provides multiple reads for a single otolith can be dealt with by creating a 2nd column for that reader, and configuring the model so that parameters for that 2nd column mirror the parameters for the 1st column for that reader.

2. *Select inputs*: The call-function "FnRun()" in R writes data in the necessary format and then calls the Punt (2008) model. This model requires several inputs, which are listed and explained below:
    a. *Data*: This is the data set as previously formatted. If the data has multiple rows with identical reads, this will cause an error and the "XXX.rep" file will have a properly formatted data matrix which can be cut-pasted into a "XXX.dat" file for use.
    b. SigOpt: This a vector with one entry for each reader (i.e. Ncol-1 entries). Each entry specifies the functional form of reading error as a function of true age. Possible entries include:
        i. "-1", "-2", "-3", etc: This will make this reader mirror the estimated SD from another reader to it's left. "-1" causes it to mirror the estimated SD for the first reader, etc. This number has to be lower than the current entry number.
        ii. "1" : Constant CV, i.e., a 1 parameter linear relationship of SD with true age.
        iii. "2": Curvilinear SD, i.e., a 3 parameter Hollings-form relationship of SD with true age
        iv. "3": Curvilinear with CV, i.e., a 3-parameter Hollings-form relationship of CV with true age
        v. "4": No error (but potentially bias)
    c. BiasOpt: This is a vector with one entry for each reader:
        i. "-1", "-2", "-3": See SigOpt
        ii. "0": Unbiased
        iii. "1": Constant CV, i.e., a 1-parameter linear relationship of bias with true age
        iv. "2": Curvilinear, i.e., a 2-parameter Hollings-form relationship of bias with true age
    d. NDataSets: This is generally "1" and other values are not implemented in the current R-code.
    e. MinAge: The minimum possible "True" age
    f. MaxAge: The maximum possible "True" age
    g. RefAge: An arbitrarily chosen age from which "true" age-composition fixed-effects are calculated as an offset. This has no effect on the answer, but could potentially effect estimation speed.
    h. MinusAge: The minimum age for which an age-specific age-composition is estimated. Ages below this MinusAge have "true" proportion-at-age (P_a) estimated as $P_a = P_{MinusAge} \cdot e^{\beta(MinusAge-a)}$, where β is an estimated log-linear trend in the "true" proportion-at-age. If MinusAge = MinAge, β is not estimated.

i. PlusAge: Identical to MinusAge except defining the age above with age-specific age-composition is not estimated.
j. MaxSd: An upper bound on possible values for the standard deviation of reading error
k. MaxExpectedAge: Set to MaxAge
l. SaveFile: Directory where "agemat.exe" is located and where all ADMB intermediate and output files should be located.
m. EffSampleSize: Indicating whether effective sample size should be calculated. Missing values in the data matrix will cause this to be ineffective, in which case this should be set to "0"
n. Intern: "TRUE" indicates that ADMB output should be displayed in R; "FALSE" does not.

Stepwise model selection in R

Function name: StepwiseFn()

Background:

Stepwise model selection allows many different model configurations to be explored: in this code, I have used AIC as the metric for comparison among model structures, although BIC or other criteria could be used. AIC seems appropriate to select among possible PlusAge values, because this parameter determines the number of estimated fixed effect hyperparameters that are used to define the true "Proportion-at-age" hyperdistribution. This hyperdistribution in turn is used as a "prior" when integrating across a "True Age" associated with each otolith. This "True Age" latent effect can be interpreted as a random effect (one for each observation), so the use of AIC to select among parameterizations of the fixed effects defining this hyperdistribution is customary (Pinheiro and Bates 2009). Additionally, the use of AIC to select the value of the PlusAge parameter appears (in preliminary analysis using Sablefish ageing error data) to lead to a "True" proportion-at-age that is biologically plausible.

Necessary Inputs:

1. *Format data*: Same as for a single-run
2. *Select inputs*: Most inputs are the same as for a single-run. However, the "SigOpt" "BiasOpt" and "PlusAge" are now specified using a matrix called "PossibleMat", which has 2*Nreaders+2 rows and as many columns as necessary. Row #1-#Nreaders specify the SigOpt for each reader; Next are the BiasOpt for each reader, followed by the PlusAge. The first entry in each row specifies the starting value for that parameter in the search algorithm; a value must be specified in the first column for each parameter. Any parameter for which the search algorithm should search across possible values has other possible values in the 2nd, 3rd, and subsequent cells in that row. An example is given in Table 2.

Diagnostic figures in R

Function name: PlotOutputFn()

Background:

There are many ways to visualize the results that are provided by the Punt et al. (2008) model. Some of these allow comparison with observed data. However, any comparison with observed data is a little problematic, as comparisons must generally be conditioned on a "True" age that is not observed. In place of a "True" age, the diagnostic plots that we present generally condition on an "Estimated" age, which is fixed as the mode of the conditional probability-at-age for each otolith.

Diagnostic plots include:

1. *Error and bias by reader*: A panel graph where each panel shows the expected and standard deviation in age reads for that reader. This is displayed against a scatterplot of the "Read" and "Estimated" ages for each otolith that was read by that reader.
2. *Proportion-at-age histogram*: The estimated "Proportion-at-age" can be plotted as a histogram, and is displayed against the "observed" distribution of read ages. This is useful to determine if the estimated "proportion-at-age" is generally plausible, e.g., whether it has too many ages where the estimated proportion-at-age approaches zero (which is unlikely in a composite sample with moderate effective sample sizes). This plot can also be used as a diagnostic to confirm that AIC has selected reasonable values for the MinusAge and PlusAge parameters.

Necessary Inputs:

1. The plotting function reads the "XXX.rep" and "XXX.par" files that are located in the directory that is provided. It also requires specifying the MaxAge and Data, as formatted and defined earlier.

Bibliography

Pinheiro, J. and Bates, D. 2009. *Mixed-Effects Models in S and S-PLUS*. Springer.

Table 1: A hypothetical “Data” matrix for a situation with 1 dataset and 3 readers. In this example, missing entries are indicated using the value “-999”; the left-most column indicates that there is only one unique otolith each with the 1,2,4,and 5 reading histories, but two unique otoliths with the reading history in row 3. The 2nd, 3rd, and 4th columns correspond to the read ages for the three readers.

1	1	2	1
1	1	2	2
2	2	-999	2
1	-999	1	2
1	-999	2	2

Table 2: A hypothetical “PossibleMat” for a situation with 1 dataset and 3 readers. In this example, the SigOpt for readers 1-2 is allowed to explore {Linear; Curvilinear SD; Curvilinear Bias}, while the SigOpt for reader 3 is constrained to mirror that of reader 2; and the BiasOpt for reader 1-2 is allowed to explore {Linear; Curvilinear; None}, while that for reader 3 again mirrors that of reader 2. The “MaxAge” starts at “30” and will search across all other values with a proposal kernel of {-10; -4; -1; +0; +1; +4; +10}

1	2	3	NA	NA	NA	NA
1	2	3	NA	NA	NA	NA
-2	NA	NA	NA	NA	NA	NA
1	2	0	NA	NA	NA	NA
1	2	0	NA	NA	NA	NA
-2	NA	NA	NA	NA	NA	NA
30	20	26	29	31	34	40