Introduction

A more detailed description of some of the Casal2 R package functions.

Simulating starting values

To evaluate MPD convergence when undertaking model exploration, there is a function called `generate.starting.pars()`. This function reads a casal2 input configuration file to identify `@estimate` blocks and generates starting values for the estimated parameters from the prior distributions (using the uniform distribution only if `all_uniform = T`) that are within the specified `lower_bound` and `upper_bound`. Note that wide bounds may have little effect on restricting parameter values, and may not be appropriate for starting values. A suggestion is to apply the function to a modified file that has narrower bounds that represent areas of higher density (generally more restrictive bounds than those used for estimation). Although looking for multi modes could also be of interest, with wide bounds.

For an example of the function being used, please see the R Markdown file that is embedded in the Casal2 R package.

Posterior Predictive P-values

This functionality has not been implemented yet, but these are initial thoughts on what would need to be done. This functionality would need Casal2 C++ code changes as well as a new function in the Casal2 R package to parse and format the output.

The Casal2 C++ code changes would be for the creation of a new report `posterior_predictions`. The Casal2 model configuration file syntax for this report would look something like

```
@report Label
type posterior_predictions
observation observation_label
```

This would assume a multirun input, for example a MCMC sample file. For each line of the `-i` file it would produce a replicate dataset denoted y^{rep}. Most of this functionality will be in Casal2, since generating simulated observations has been implemented. However, there are no public functions (this would be needed because the report class will be responsible for executing the simulate call) that allow an observation to simulate data, so this will have to be implemented (shouldn't be difficult) could almost keep the simulate call at the parent class..

Once this has happened the `TBD` function in the Casal2 R package will read that in and users can define different discrepancy functions $D()$, i.e., the likelihood or Pearson's residuals **discussion**: should Casal2 do the discrepency calculation or the Casal2 R package? An example of standardised residuals for a discrepancy function

$$D\left(y^{rep};\theta\right) = \sum_{i=1} \frac{y_i^{rep} - \mathbb{E}[y_i]}{Var(y_i)}$$

then a P-value can be generate as.

$$ppp\left(y\right) = P_A\left[D\left(y^{rep};\theta\right) \geq D\left(y;\theta\right)|M,y\right]$$

where, $ppp\,(y)$ is the posterior predictive p-value Hjort et al. (2006), M is the model under assessment, P_A denotes the distribution of the discrepancy posterior. An alternative formulation,

$$ppp\,(y) = \frac{1}{A}\sum_{j=1}^{A} I\left\{D\left(y_{rep};\theta\right) \geq D\left(y;\theta\right)\right\}$$

where, I is an indicator function, and A is the number of samples from the posterior.

Data Weighting

There are two data weighting functions in the Casal2 R library, `MethodTA1.8()` and `cv.for.cpue()`.

MethodTA1.8() This is a method for iterative reweighting for multinomial composition data described in Francis (2011). For completeness this section will redefine the method, and how the function works. The R function, takes two main inputs, a `casal2MPD` which is produced by using `extract.mpd()` from casal2 text output files, and a report label. This function is defined for composition data (either age or length) and assumes the likelihood is multinomial. This function calculates a weight that is then used to *update* the effective sample size of the multinomial to then be re-estimated and re-weighted (put back through this function again) until convergence ($w = 1$). If the method is applied to a single observation dataset, the function can produce a plot showing fit through the observations with the command `plot.it = T`.

Some general theory of reweighting the composition data in a model that has an observation denoted as $O_{t,b}$ (note these are proportions $\sum_{b=1} O_{t,b} = 1$) at time t for composition bin b (either age or length bin), and a model fitted value $E_{t,b}$. Data weighting aims to standardize the errors $(O_{t,b} - E_{t,b})$ so that the standardised errors have constant variance for all time steps and bins, i.e., $S_{t,b} = (O_{t,b} - E_{t,b})/X_{t,b}$, where $X_{t,b}$ is a function of the weighting parameter and $Var(S_{t,b}) = k$. Once error distribution assumptions are made for a dataset, e.g., multinomial error, the distribution of $S_{t,b}$ is defined and the aim is to find values of $X_{t,b}$ that result in $S_{t,b}$ having mean 0 and constant variance. Using the example from McAllister & Ianelli (1997) assuming the multinomial error distribution and $N_t = w\tilde{N}_t$. With the multinomial distribution the standardised error (Pearson's residuals) are sought, which involves defining the variance of the error, $Var(O_{t,b} - E_{t,b}) = E_{t,b}(1 - E_{t,b})/(w\tilde{N}_t)$. Standardised errors are then calculated as $X_{t,b} = \left[E_{t,b}(1 - E_{t,b})/\tilde{N}_t\right]^{0.5}$, which results in $k = 1/w$, where

$$w = 1/Var_{t,b}\left(O_{t,b} - E_{t,b}/\left[E_{t,b}(1 - E_{t,b})/\tilde{N}_t\right]^{0.5}\right)$$

where $Var_{t,b}$ is the finite-sample variance function for a sample. This involves an iterative process with the first stage setting initial values for the weighting variable $\tilde{N}_t$. The second stage involves calculating the standardised residuals and calculating the weighting factor w that adjusts $S_{t,b}$ towards the desired constant variance. This is achieved by updating $\tilde{N}_t = w\tilde{N}_t$ and re-running the model, and to iteratively applying this stage until a constant variance is found. A weakness of this specific example of the method is there are no explicit accounting of correlation between ages, which is often observed in age composition data due to intra-haul correlation (Pennington & Volstad 1994). There are alternative formulations for multinomial that focus on the error between mean age or length $(\bar{O}_t - \bar{E}_t)$, which can allow for correlations (method TA1.8 Francis (2011) Equation 1). This is the method commonly applied for New Zealand stock assessments (Ministry for Primary Industries 2014).

$$w = 1/Var_t\left(\bar{O}_t - \bar{E}_t/\left(v_t/\tilde{N}_t\right)^{0.5}\right) \tag{1}$$

where, $v_t = \sum_{b=1} E_{t,b} x_b^2 - \bar{E}_t^2$, where, x_b is the attribute for bin b and, i.e if $b = 3$ which corresponded to the length bin of 31cm, then $x_b = 31$ and $\bar{E}_t = \sum_{b=1} E_{t,b} x_b$

References

Francis, R. C. (2011), 'Data weighting in statistical fisheries stock assessment models', *Canadian Journal of Fisheries and Aquatic Sciences* **68**(6), 1124–1138.

Hjort, N. L., Dahl, F. A. & Steinbakk, G. H. (2006), 'Post-processing posterior predictive p values', *Journal of the American Statistical Association* **101**(475), 1157–1174.

McAllister, M. K. & Ianelli, J. N. (1997), 'Bayesian stock assessment using catch-age data and the sampling-importance resampling algorithm', *Canadian Journal of Fisheries and Aquatic Sciences* **54**(2), 284–300.

Ministry for Primary Industries (2014), Fisheries Assessment Plenary, May 2014: stock assessments and stock status, Technical Report May, Compiled by the Fisheries Science Group, Ministry for Primary Industries, Wellington, New Zealand.

Pennington, M. & Volstad, J. H. (1994), 'Assessing the effect of intra-haul correlation and variable density on estimates of population characteristics from marine surveys', *Biometrics* pp. 725–732.