Bayesian Fertility Trajectory Analysis (BaFTA) tutorial.

Fernando Colchero[1,2,*]

[1]Department of Primate Behavior and Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[2]Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark

[*]Correspondence: fernando_colchero@eva.mpg.de

Contents

1 Brief introduction

This vignette provides an overview of the data needed and the use of the functions in in the R package BaFTA. The package allows to make inference on age-specific fertility for aggregated and individual level data. For details of the age-specific fertility models included in the package see [1].

2 Installing the package

Before installing BaFTA, You will need package **snowfall** for parallel computing, which you can install as

```
# Install snowfall:
install.packages("snowfall")
```

You can download the BaFTA package from the GitHub repository. Store the tar.gz file in your computer and install the package as

```
# Install package:
install.packages("My directory/BaFTA_1.0.0.tar.gz", type = "source",
                 repos = FALSE)
```

Alternatively, you can install it directly from GitHub by typing the following code:

```
# Install devtools:
install.packages("devtools")

# Load devtools:
library(devtools)

# Install BaFTA from GitHub:
install_git("https://github.com/fercol/BaFTA", subdir = "pkg/")
```

To initialize BaFTA, simply type to the console:

```
# Load the library:
library(BaFTA)
```

which will also load **snowfall**.

3 Data formatting

3.1 Aggregated data

The data frame needed requires only the following four columns: **Age** with discrete ages; **nParents** with the number of available adults per age; **nOffspring** with the number of offspring produced in that age interval; **Fertility** with the age-specific fertility, calculated as **Fertility = nOffspring / nParents**. Here is an example of the first six rows of a simulated dataset:

```
  Age nParents nOffspring  Fertility
1   4      169          2 0.01183432
2   5      128         13 0.10156250
3   6      120         17 0.14166667
4   7      119         25 0.21008403
5   8      110         16 0.14545455
6   9      104         23 0.22115385
```

3.2 Individual data with discrete ages

When indivdiual level data with discrete age intervals is available, then the data frame for analysis in BaFTA should include the following three columns: **indID** with character string of individual IDs of the parents; **Age** with the discrete ages of the parents at the time of birth; **nOffspring** with the number of offspring produced during that reproductive event by the given parent. Here is an example on simulated data:

```
  indID Age nOffspring
1     1   4          0
2     8   4          0
3    10   4          0
4    34   4          0
5    53   4          0
6    56   4          0
```

3.3 Individual data with continuous ages and variable IBI

For individual data with continuous ages and variable IBIs, then the data needed consist of a data frame with individual rows per reproductive event, and with the following columns: **indID** with character string of individual IDs of the parents; **Age** with the discrete ages of the parents at the time of birth; **nOffspring** with the number of offspring produced during that reproductive event by the given parent, including failed events (i.e., **nOffspring** = 0); **IBI** with the time since last reproduction in years. Here is an example on simulated data:

```
  indID       Age nOffspring      IBI First
1     1  4.015909          0 0.000000     1
2     1  5.961883          2 1.945974     0
3     1  7.109653          1 1.147770     0
4     1  8.310275          0 1.200622     0
5     1 10.031800          0 1.721525     0
6     1 11.435269          1 1.403469     0
```

4 Analysis

To run an analysis on aggregated data (**aggrRepr**), using the quadratic model, it is sufficient to use function **bafta()** as follows:

```
# Quadratic:
outAG0 <- bafta(object = aggrRepr, model = "quadratic",
                dataType = "aggregated", nsim = 4, ncpus = 4,
                niter = 11000, burnin = 1001, thinning = 20)
```

where argument **model** specifies the functional form for the age-specific fertility (i.e., expected number of offspring produced by individuals of a given age), and, as its name indicates, argument **dataType** specifies the type of data, with options "**aggregated**", "**indivSimple**" for individual level data with discrete age information, and "**indivExtended**" for individual level data with continuous ages and variable IBIs.

To visualize the output of the model, functions **print()**

```
# Simple print:
outAG0

# Summary:
summary(outAG0)
```

which prints the following output to the console:

```
>
> Call:
> Model            : quadratic
> Data type        : aggregated
> Num. iterations  : 11000
> Burnin           : 1001
> Thinning         : 20
> Number of sims.  : 4
> Computing time   : 0.063 mins
>
> Coefficients:
>           Mean      SD    Lower    Upper  Rhat
> b0      0.25465 0.01965 0.218931  0.29533 1.000
> b1      0.01206 0.00177 0.008716  0.01577 1.003
> b2     10.36316 0.49623 9.401141 11.34439 1.000
> alpha   8.62592 2.19885 5.952706 14.12191 1.000
```

```
>
> Convergence:
> All parameter chains converged.
>
> Model fit:
>
> DIC = 151.09
>
> Predictive loss:
>  Good. Fit Penalty Deviance
>        387     343      730
```

```
# Gamma distributional model:
outAG1 <- bafta(object = aggrRepr, model = "gamma",
                dataType = "aggregated", nsim = 4, ncpus = 4,
                niter = 11000, burnin = 1001, thinning = 20)
```

As shown in Fig. 1, to visually inspect the traces for proper convergence, simply type

```
# Plot traces:
plot(outAG1)
```

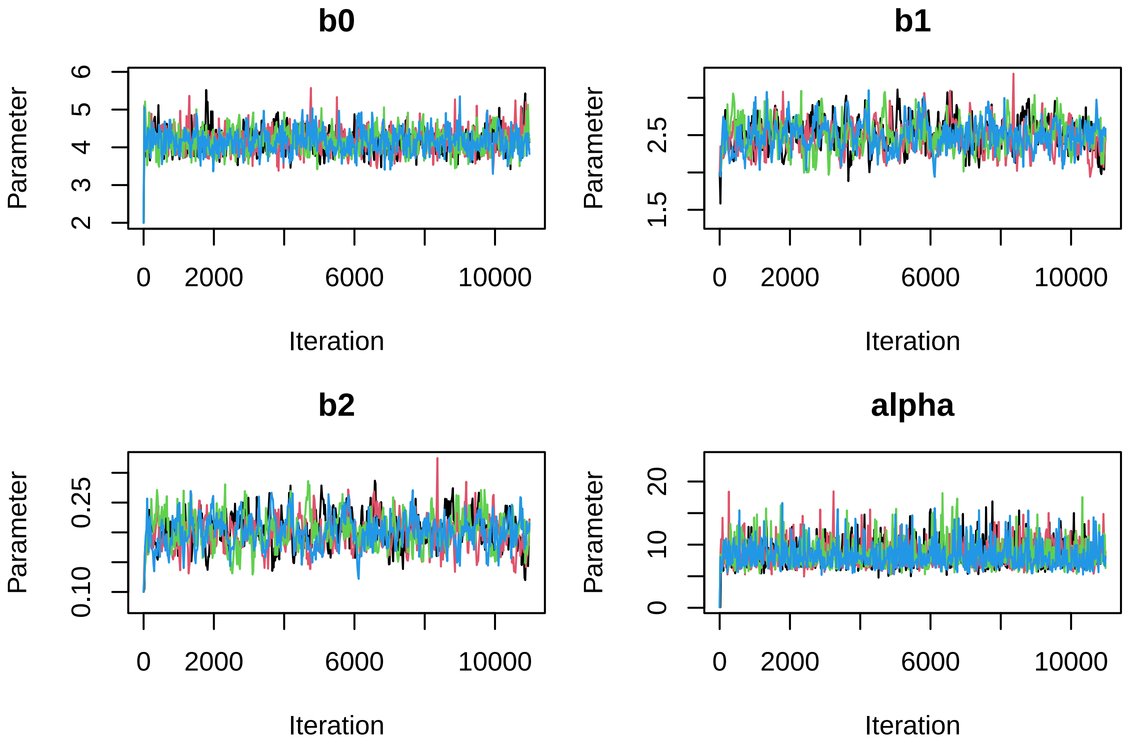

Figure 1: Plot of fertility parameter traces for BaFTA output.

Alternatively, it is possible to plot the posterior densities of the fertility parameters (Fig. 2) by typing

```
plot(outAG1, type = "density")
```

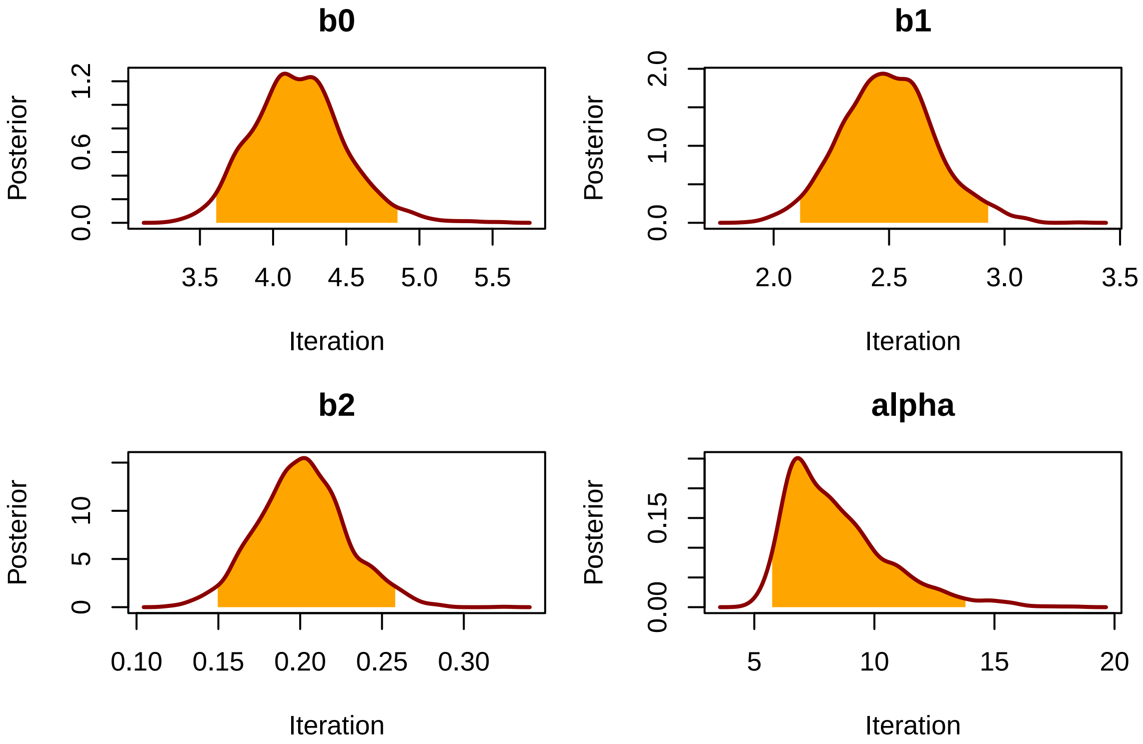

Figure 2: Plot of fertility parameter posterior densities for BaFTA output. The orange filled polygons depict the 95% credible intervals.

To plot the estimated fertility with its 95% credible intervals compared to the actual fertility just change the `type` argument as

```
plot(outAG1, type = "fertility")
```

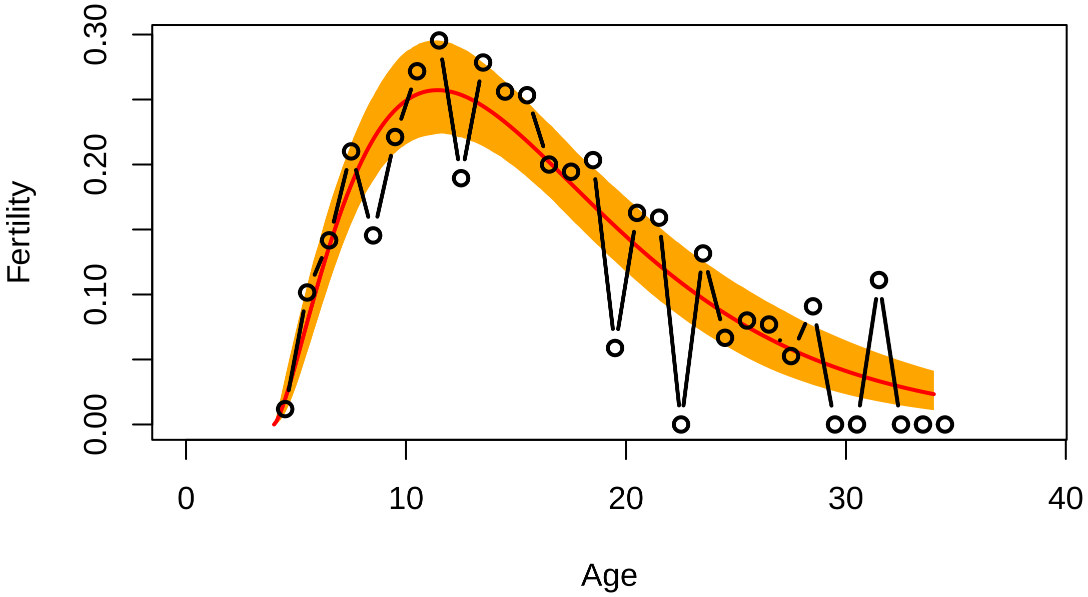

Figure 3: Plot of the estimated age-specific fertility with the 95% credible intervals (orange) against the actual fertility.

Finally, to plot the predicted number of offspring produced by parents of a given age against the actual number of offspring, modify the **`type`** argument as

```
plot(outAG1, type = "predictive")
```

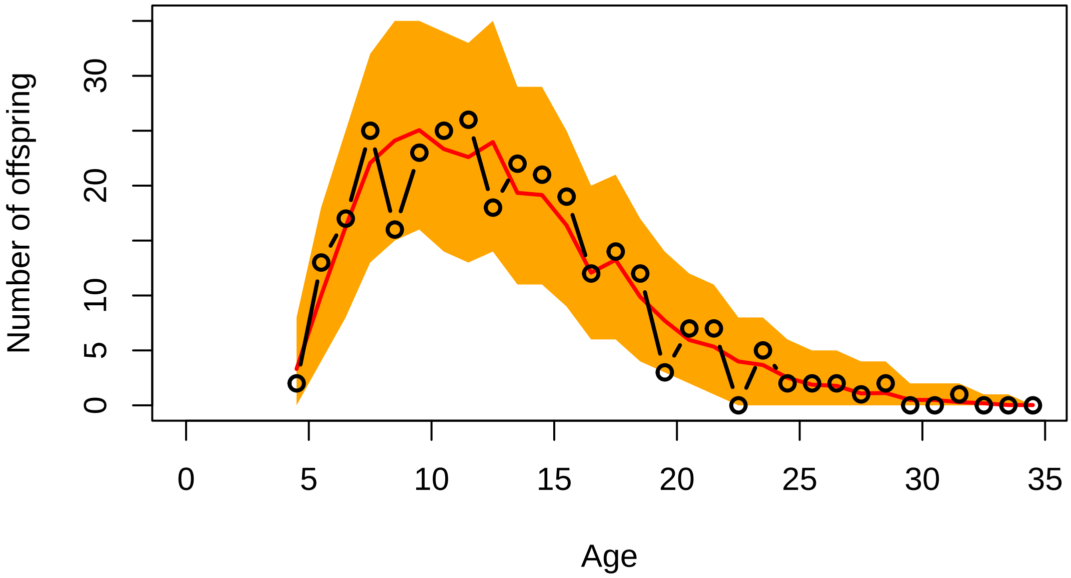

Figure 4: Plot of the predicted age-specific number of offspring with the 95% predictive intervals (orange) against the actual number of offspring.

5 Fertility models included in BaFTA

The following table shows the different models of age-specific fertility included in BaFTA. The parametrization is as in the package and therefore differs from that in Colchero [1] (i.e., $\boldsymbol{\theta}^{\top} = [b_0, b_1, \ldots, b_p]$)). Note that the column **Model** shows the model name as it needs to be specificed in argument **`model`** within function **`bafta()`**.

Model	Fertility function	References
`quadratic`	$g(x) = b_0 \exp\left[-b_1(x - b_2)^2\right]$, where $b_0, b_1, b_2 \geq 0$.	[2, 3, 4]
`ColcheroMuller`	$g(x) = b_0 \exp\left[b_1 x - b_2 x^2 + b_3/(x+1)\right]$, where $b_0, b_1, b_2 \geq 0$ and $b_3 \in \mathbb{R}$.	[4, 5]
`PeristeraKostaki`	$g(x) = b_0 \exp\left[-\left(\frac{x - b_3}{b(x)}\right)^2\right]$, where $b_0 \geq 0$, $b_3 \in \mathbb{R}$, and $b(x) = b_1$ if $x <= b_3$, $b(x) = b_2$ if $x > b_3$.	[6]
`Hadwiger`	$g(x) = b_0 \frac{b_1}{b_2}\left(\frac{b_2}{x}\right)^{3/2} \exp\left[-b_1^2\left(\frac{b_2}{x} + \frac{x}{b_2} - 2\right)\right]$, where $b_0, b_1, b_2 > 0$.	[7]
`gamma`	$g(x) = b_0 \left[\Gamma(b_1) b_2^{b_1}\right]^{-1} x^{b_1 - 1} e^{x/b_2}$, where $b_0, b_1, b_2 > 0$.	[8]
`beta`	$g(x) = b_0 \frac{(x - b_3)^{b_1 - 1}(b_4 - x)^{b_2 - 1}}{(b_4 - b_3)^{b_1 + b_2 - 1} B(b_1, b_2)}$, where $b_0, b_1, b_2, b_3, b_4 > 0$, and B is the beta function.	[8]
`skewNormal`	$g(x) = b_0\ 2 b_1^{-1} \phi\left(\frac{x - b_2}{b_1}\right) \Phi\left\{b_3\left(\frac{x - b_2}{b_1}\right)\right\}$, where $b_0, b_1, b_3 > 0$, $b_2 \in \mathbb{R}$, and ϕ and Φ are the standard normal PDF and CDF, respectively.	[9, 10]
`gammaMixture`	$g(x) = b_0(b_1)\left\{\left[\Gamma(b_2) b_3^{b_2}\right]^{-1} x^{b_2 - 1} e^{x/b_3}\right\}$ $+ \quad (1 - b_1)\left\{\left[\Gamma(b_4) b_5^{b_4}\right]^{-1} x^{b_4 - 1} e^{x/b_5}\right\}$, where $b_0, \ldots, b_5 > 0$.	[8]
`HadwigerMixture`	$g(x) = b_0 \left\{ b_1 \left(\frac{b_2}{b_3}\right)\left(\frac{b_3}{x}\right)^{3/2} \exp\left[-b_2^2\left(\frac{b_3}{x} + \frac{x}{b_3} - 2\right)\right]\right.$ $+ (1 - b_1)\left(\frac{b_4}{b_5}\right)\left(\frac{b_5}{x}\right)^{3/2}$ $\left. \times \exp\left[-b_4^2\left(\frac{b_5}{x} + \frac{x}{b_5} - 2\right)\right]\right\}$, where $b_0, \ldots, b_5 > 0$.	[8]

Model	**Fertility function**	**References**
`skewSymmetric`	$g(x) = b_0 \; 2b_1^{-1}\phi\left(\dfrac{x-b_2}{b_1}\right) \times \quad \Phi\left\{b_3\left(\dfrac{x-b_2}{b_1}\right) + b_4\left(\dfrac{x-b_2}{b_1}\right)^3\right\},$ where $b_0, b_1, b_3 > 0$, $b_2, b_4 \in \mathbb{R}$, and ϕ and Φ are the standard normal PDF and CDF, respectively.	[9, 10]
`skewLogistic`	$g(x) = b_0 \; 2b_1^{-1}\dfrac{e^{-(x-b_2)/b_1}}{\left(1+e^{-(x-b_2)/b_1}\right)^2} \times \quad \dfrac{1}{\left(1+e^{-b_3[(x-b_2)/b_1]-b_4[(x-b_2)/b_1]^3}\right)},$ where $b_0, b_1, b_3 > 0$, $b_2, b_4 \in \mathbb{R}$, and ϕ and Φ are the standard normal PDF and CDF, respectively.	[11]

6 Inference details

6.1 Generals of the inference model

The model differs depending on the type of data. For data of type "**`indivSimple`**", we can define the nonhomogeneous Poisson process $\{N(x), x \geq 0\}$ with intensity function $\lambda(x)$, where $N(x)$ are the cumulative number of offspring produce by a given individual up to age x. At a given discrete age interval $[x, x+1)$ we can define the random variable for an average individuals as

$$Y_x = N(x+1) - N(x) \sim \text{Pois}[\Lambda(x+1) - \Lambda(x)], \tag{1}$$

where

$$\Lambda(x) = \int_0^x \lambda(t)dt. \tag{2}$$

For age intervals of one year, $\lambda(x)$ can be calculated at the midpoint of the interval $[x, x+1)$ such that

$$\int_x^{x+1} \lambda(t)dt = \Lambda(x+1) - \Lambda(x) = \lambda\left(x + \frac{1}{2}\right) + \varepsilon_x, \tag{3}$$

where, based on the error of the midpoint rule, $\varepsilon_x \leq \frac{K}{24}$ and $K \geq |\lambda''(t)|$ for all $t \in [x, x+1]$. Given that g is a concave function of age we can therefore assume that the error ε_x is negligible.

The random variable Y_x in Eq. 1 represents the number of offspring produced within the age interval $[x, x+1)$ by an average individual. However, given that individuals differ in their ability to reproduce, a hierarchical model is more appropriate [5, 4]. Here, we define the conditional random variable

$$\begin{aligned} Y_{i,x}|U_i &\sim \text{Pois}[\lambda(x, U_i)] \\ U_i &\sim Ga(\alpha, \alpha) \end{aligned} \tag{4}$$

where $\alpha > 0$, with observations $y_{i,x} \in \mathbb{N}_{\geq 0}$, $U_1, \ldots, U_{n_x}$ are i.i.d. random variables variable with $E(U_i) = 1$ for $i = 1, 2, \ldots, n_x$ where n_x is the number of parents at age x, and $g : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}_{\geq 0}$ is a $k \geq 2$ times differentiable smooth concave function. The resulting random variable is marginally distributed as $Y_{i,x} \sim \text{NB}[\alpha, \alpha/(g(x) + \alpha)]$, with likelihood function

$$\mathcal{L}(\boldsymbol{y}|\boldsymbol{\theta}, \sigma^2) = \prod_{i=1}^{n} \prod_{x \in E_i} \frac{\Gamma(y_{i,x} + \alpha)}{\Gamma(\alpha) y_{i,x}!} \left(\frac{\alpha}{g(x) + \alpha}\right)^{\alpha} \left(\frac{g(x)}{g(x) + \alpha}\right)^{y_{i,x}}, \tag{5}$$

where E_i is the set of ages at which individual i was known to be present, $\text{E}[Y_{i,x}] = g(x)$ and $\text{Var}[Y_{i,x}] = g(x)[(g(x) + \alpha)/\alpha]$

For data type **`aggregated`**, the random variable of interest is then W_x. However, since $Y_{i,x}$ for $i \in I_x$ where I_x is an indexed set of individuals alive at age x are not identically distributed (i.e., $Y_{i,x}|U_i \sim \text{Pois}[\lambda(x, U_i)]$) for all $i \in I_x$), the distribution of W_x needs to be fully specified.

As shown in [1], the resulting random variable is $W_x \sim \text{NB}(n_x\alpha, \alpha/[g(x)+\alpha])$, with probability mass function

$$f_W(w_x) = \frac{\Gamma(w + n\alpha)}{\Gamma(n\alpha)w!} \left(\frac{g(x)}{g(x) + \alpha}\right)^{w_x} \left(\frac{\alpha}{g(x) + \alpha}\right)^{n_x\alpha}, \quad \text{for } x \in \mathbb{N}_{\geq 0} \tag{6}$$

where $E[W_x] = n_x g(x)$ and $\text{Var}[W_x] = n_x g(x)[g(x) + \alpha]/\alpha$.

If the data are **indivExtended** then $x > 0$ and we define two additional random variables, namely W_i with $w_i \geq 0$ for the time between the minimum age at maturity and the first birth, with exponential PDF, and $Z_{i,x} = \delta_{i,x} - \tau$ where $\delta_{i,x} \geq 0$ is the interbirth interval for individual i at age x and τ is the minimum gestation time (also including weanning if relevant). Here we need also a hierarchical model given by

$$\begin{aligned} Y_{i,x}|U_i, W \geq w_i \vee Z_x = z_{i,x} &\sim F_Y(y_{i,x}|u_i, w_i, z_{i,x}) \\ U_i &\sim N(0, \sigma_u^2) \\ W_i &\sim F_W(w_i) \\ Z_{i,x}|V_i &\sim F_Z(z_{i,x}, v_i) \\ V_i &\sim N(0, \sigma_v^2), \end{aligned} \tag{7}$$

and with likelihood

$$\mathcal{L}(\ldots) = \begin{cases} \gamma e^{-\gamma w_i} f_Y(y_{i,x}|\boldsymbol{\theta}, u_i) f_U(u_i) & \text{for } x = w_i \\ \eta e^{-e^{v_i}\eta z_{i,x}} f_Y(y_{i,x}|\boldsymbol{\theta}, u_i) f_U(u_i) & \text{for } x > w_i \end{cases} \tag{8}$$

The posterior for the most parameterized model (i.e., **indivExtended**) is approximated as

$$\begin{aligned} p(\boldsymbol{\theta}, \eta, \gamma, \sigma_u, \sigma_v|\boldsymbol{y}, \boldsymbol{w}, z) \quad \propto \quad & p(\boldsymbol{y}, \boldsymbol{w}, z|\boldsymbol{\theta}, \eta, \gamma, \sigma_u, \sigma_v) \\ & \times p(\boldsymbol{\theta}|\boldsymbol{\mu}_\theta, \boldsymbol{\sigma}_\theta) p(\eta|\mu_\eta, \sigma_\eta) p(\gamma|\mu_\gamma, \sigma_\gamma) \\ & \times p(\sigma_u|s_1, s_2) p(\sigma_v|r_1, r_2), \end{aligned} \tag{9}$$

where the first term on the right hand side of Eq. 9 is the likelihood function while the following terms are prior densities for the parameters. For the other two types of datasets, the posterior is simplified accordingly.

6.2 MCMC performance diagnostics

After the MCMC algorithms are finished, a range of diagnostics are calculated on the parameter chains. If multiple simulations are implemented and all of them run through, then potential scale reduction is calculated for each parameter to estimate convergence [12]. This diagnostic is calculated as $\hat{R} = \sqrt{\hat{v}^+/W}$, where W is a measure of the within-sequence variance and $\hat{v}^+$ is a weighted average of the between-sequence variance (B) and W. Convergence is attained when $\hat{R}$ is close to 1. As a rule of thumb, we have assigned an arbitrary upper bound of $\hat{R} < 1.05$ above which it is assumed that parameters have not reached convergence.

6.3 Model fit

If all parameters have converged, BaFTA calculates deviance information criterion (DIC; [13]), which has been described as a measure of predictive power and a criterion for model fit. DIC approximates the expected predictive deviance, and is calculated as

$$\text{DIC} = 2\hat{D}_{avg}(y) - D_{\hat{\theta}}(y)$$

where y denotes the observed data, $\hat{D}_{avg}(y)$ is the mean discrepancy between the data and the model as a function of the parameters θ, averaged over the posterior distribution, while $D_{\hat{\theta}}(y)$ is the discrepancy at the posterior mode (here represented by the point estimate $\hat{\theta}$). It is important to outline that the use of DICs is still controversial and therefore the results need to be taken with caution (see responses in [13]). In order to improve the measure provided, BaFTA's DIC is calculated as an approximation to the group-marginalized DIC presented by Millar [14].

In addition to the DIC, BaFTA calculates predictive loss as described by Gelfand *et al.* [15]. This measure requires producing a vector of predicted number of offspring $\hat{\boldsymbol{y}}|\boldsymbol{y}$, where $\boldsymbol{y}$ are the observed number of offsprings. With these $\hat{\boldsymbol{y}}$ we can construct a prediction distribution, integrated across the posterior densities of the parameters as

$$p(\hat{\boldsymbol{y}}|\boldsymbol{y}) = \int \ldots \int p(\hat{\boldsymbol{y}}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}) d\theta_0 \ldots d\theta_p. \tag{10}$$

The integrals in Eq. 10 are evaluated numerically by means of the converged parameter traces, and the expected predictions are approximated as

$$E[\hat{\boldsymbol{y}}|\boldsymbol{y}] = \frac{1}{M} \sum_{m=1}^{M} E[\hat{\boldsymbol{y}}|\hat{\boldsymbol{\theta}}_m], \tag{11}$$

where M is the total number of MCMC iterations, and $\hat{\boldsymbol{\theta}}_m$ is the vector of parameters estimated at step m. Mean and upper and lower 95% predictive intervals are included as part of the BaFTA output. With these, BaFTA calculate predictve loss, D, by first calculating a measure of goodness of fit

$$G = \sum_{i=1}^{N} \left(E[\hat{\boldsymbol{y}}|\boldsymbol{y}] - \boldsymbol{y}\right)^2 \tag{12}$$

and a measure of model dispersion, or penalty term, given by the predictive variance

$$P = \sum Var[\hat{\boldsymbol{y}}|\boldsymbol{y}]. \tag{13}$$

Predictive loss is then calculated as $D = G + P$.

References

[1] Colchero, F. Inference on age-specific fertility in ecology and evolution. Learning from other disciplines and improving the state of the art. in prep. (2023).

[2] Sharp, S. P. & Clutton-Brock, T. H. Reproductive senescence in a cooperatively breeding mammal. *Journal of Animal Ecology* **79**, 176–183 (2010).

[3] Dugdale, H. L., Pope, L. C., Newman, C., Macdonald, D. W. & Burke, T. Age-specific breeding success in a wild mammalian population: selection, constraint, restraint and senescence. *Molecular Ecology* **20**, 3261 – 3274 (2011).

[4] Colchero, F., Eckardt, W. & Stoinski, T. Evidence of demographic buffering in an endangered great ape: Social buffering on immature survival and the role of refined sex-age classes on population growth rate. *Journal of Animal Ecology* **90**, 1701–1713 (2021).

[5] Muller, M. N. *et al.* Sexual dimorphism in chimpanzee (Pan troglodytes schweinfurthii) and human age-specific fertility. *Journal of human evolution* **144**, 102795 (2020).

[6] Peristera, P. & Kostaki, A. Modeling fertility in modern populations. *Demographic Research* **16**, 141–194 (2007).

[7] Hadwiger, H. Eine analytische Reproduktionssunktion für biologische Gesamtheiten. *Scandinavian Actuarial Journal* **1940**, 101–113 (1940).

[8] Hoem, J. M. *et al.* Experiments in modelling recent Danish fertility curves. *Demography* **18**, 231–244 (1981).

[9] Mazzuco, S. & Scarpa, B. Fitting age-specific fertility rates by a skew- symmetric probability density function. Tech. Rep., University of Padua (2011).

[10] Mazzuco, S. & Scarpa, B. Fitting age-specific fertility rates by a flexible generalized skew normal probability density function. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **178**, 187–203 (2015).

[11] Asili, S., Rezaei, S. & Najjar, L. Using Skew-Logistic Probability Density Function as a Model for Age-Specific Fertility Rate Pattern. *BioMed Research International* **2014**, 790294 (2014).

[12] Gelman, A. *et al. Bayesian Data Analysis*. Chapman and Hall/CRC (Chapman and Hall/CRC, 2013), third edn.

[13] Spiegelhalter, D., Best, N., Carlin, B. & Linde, A. V. D. Bayesian measures of model complexity and fit. *Journal Of The Royal Statistical Society Series B-Statistical Methodology* **64**, 583 – 639 (2002).

[14] Millar, R. B. Comparison of hierarchical bayesian models for overdispersed count data using dic and bayes' factors. *Biometrics* **65**, 962–969 (2009).

[15] Gelfand, A. & Ghosh, S. Model Choice: A Minimum Posterior Predictive Loss Approach. *Biometrika* **85**, 1 – 11 (1998).