# Frog Growth Curves: Model and Methods
# Version 1

Shirley Pledger and Ben Bell

July 2022

# 1 Introduction

This document details the equations, assumptions and statistical methods used in Bell and Pledger, "Post-metamorphic body growth and remarkable longevity in two terrestrial New Zealand frogs (*Leiopelma archeyi* and *L. hamiltoni*)" (2022) submitted to the New Zealand Journal of Ecology for the special issue in honour of Phil Bishop. This is preliminary modelling, designed as a proof of concept for wrapping a finite mixture analysis (McLachlan and Peel, 2000) around growth curve modelling. The finite-mixture wrapping is tried in order to deal with two types of missing information, the date of birth of each frog and its sex.

The snout-vent length (SVL) data comes from long-term capture-recapture studies of frogs. Because of the capture-recapture origin the data are intermittent and sparse, with no guarantee of capture (and therefore measurement) at any chosen time. Because of adult sexual dimorphism (with females larger than males) the objective is to fit two von Bertalanffy growth curves (von Bertalanffy, 1960) of length by age, one for each sex. Each individual has two bits of missing information: (i) the age at each measurement and (ii) the sex. Age is taken to be the time since "birth", when the frog leaves its father and starts an independent life. If the age at first measurement (AFM) is known, ages at subsequent measurements are also available from the known time interval between measurements. The other missing information, SEX, is not generally available from external information, but can be estimated from adult size as there is sexual dimorphism (with females larger than males). Choosing $M$ as an age at maturity (when growing is assumed to have ceased), we set up $M + 1$ categories (classifications) where the age at the first measurement is in one of the age intervals (0,1], (1,2], (2,3], ... $(M - 1, M]$ and $(M, \infty)$ years. Individuals are cross-classified by SEX, being either Sex1 (larger at maturity) or Sex2. Each individual is assumed to belong to one of the $2(M + 1)$ groups, but its exact group membership is unknown. Because there are only finitely many groups, it is possible to use a finite mixture analysis to allocate each individual probabilistically to the $2(M + 1)$ groups. In this way a clustering by finite mixtures is wrapped around growth curve modelling.

The EM (expectation-maximisation) algorithm (Dempster, Laird and Rubin, 1977; McLachlan and Krishnan, 1997) is used to estimate missing information. It starts with an approximate guessed estimation of the von Bertalanffy curve parameters, then alternates the E step (estimating the probability for each individual to belong in each of the groups) with the M-step (maximum likelihood estimation to update parameters for the von Bertalanffy curves), until convergence to a solution is achieved.

The analysis described here (Version 1) shows that this analysis is possible, using the EM algorithm to successively approximate the missing information and to ultimately provide the maximum-likelihood fitted curves, one for each sex. However, the M step of the EM algorithm was written using the `nlme` package (Pinheiro and Bates, 2000) in **R** (R Core Team, 2020) for the curve-fitting part of the analysis, and this requires at least four SVL observations from each frog. It is possible to do the analysis including the information from other individuals with only two or three observations, so a later publication, Version 2, is planned to provide more comprehensive modelling employing data from all individuals caught at least twice. The improved modelling in Version 2 will include (i) a switch to the incremental growth curve model of Fabens (1965), and (ii) the modelling of a variance component for individual variation as well as for random residual variation (Armstrong and Brooks 2013).

## 2 Definitions, Notation and Formulae

### Von Bertalanffy (vB) curve and parameters:

We use the version of the vB curve in which length $y$ in term of age $x$ is given by

$$y = \alpha - (\alpha - \beta)\exp(-\kappa x) \tag{1}$$

This curve is seen in Figure 1. The growth has the property that at any time the slope of the curve, $\frac{dy}{dx}$ is proportional to the current shortfall from the asymptote, $\alpha - y$. Thus growth slows as the asymptote is approached.

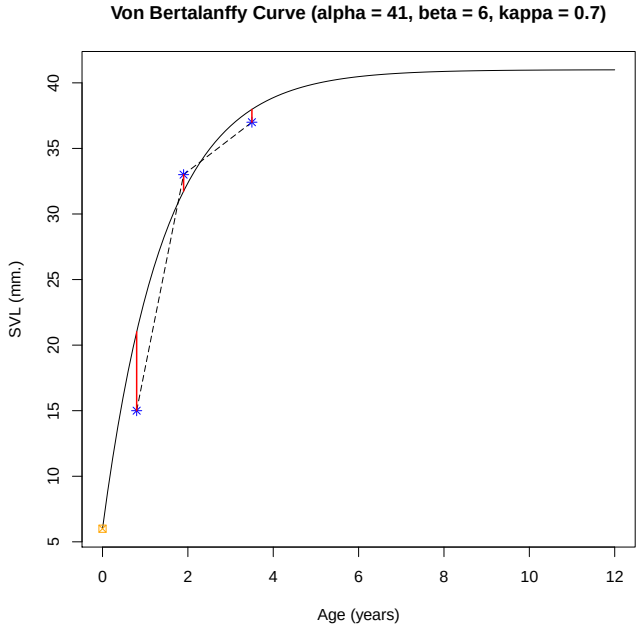


Figure 1: A von Bertalanffy curve, with parameters $\alpha$ = 41mm (asymptote), $\beta$ = 6mm (length at birth) and $\kappa$ = 0.7 (the growth parameter). If its age is known, the SVL history of one individual may be shown as the blue points. Residuals (deviations of the observed points from the expected values on the curve) are shown in red. The orange point (0,6) shows the expected length at Age 0 ("birth") . Each individual has its own chain of correlated SVL measurements; only one chain is shown.

### 2.1 Data

There are $I$ individual frogs, indexed by $i = 1, \ldots I$.
$y_{ij}$ is the SVL (mm) of individual $i$ on its measurement occasion $j$, where $j = 1, \ldots J_i$ with $J_i$ being the number of measurements of individual $i$.
The calendar date $t_{ij}$ is recorded at the time of the $j^{\text{th}}$ measurement of individual $i$. Hence each individual $i$ has a **measurement history** consisting of two vectors, the SVL measurements $y_{ij}$

and the corresponding calendar dates $t_{ij}$. The date $t$ is recorded in years and fractions of years AD so, for example, 1987.125 represents 15 Feb 1987.

The age of individual $i$ at the time of its $j^{\text{th}}$ measurement is denoted by $x_{ij}$. This is unknown, but will be estimated during the EM algorithm. The estimated age is denoted by $\widehat{x}_{ij}$. The other information missing from the dataset is the sex, where the label S1 is used for the higher-asymptote animals (females, in the case of these frogs) and S2 for the lower. The EM algorithm updates each individual's probability of being female ($p_1$) or male ($p_2 = 1 - p_1$).

## 2.2 The Groups

Individuals are grouped (or clustered) into one of twelve categories, cross-classified by the two pieces of missing information, AFM (age at first measurement) and SEX.

**AFM**

It is assumed that there is an age $M$ of maturity at which growth in length has stopped. For the frogs this was assumed to be $M = 5$ years, although it could be set higher which would give the same answers but with the penalty of more groups in the analysis. In practical terms, it is the age at which the individual is within measurement error of its asymptote; this removes theoretical complications from a model with an asymptote which can never be reached.
With $M = 5$, there are six categories: AFM is in one of the intervals A0 = 0-1 year, A1 = 1-2 years, etc. with A5 = 5+ years. Since there is an annual birth pulse in February each year, we are assigning mid-February as every frog's birthday. If it is caught while small it may possibly be classified quite accurately - e.g. if caught in September, it is only a question of whether it is 7 months old, or 1 year and 7 months, or 2 years 7 months, etc. However, if it is first caught as an adult, its AFM can only be estimated as at least 5 yr 7 months. This more accurate age estimation from small frogs was used by Bell in previous publications; our finite mixture approach here is merely building his methods into a formal probabilistic model.

**SEX**

There are two sex categories, S1 and S2. If there is enough adult sexual dimorphism, individuals with several large SVL measurements will be easily assigned to a category, while an individual only caught while still growing may not have any accurate assignment possible (having probabilities around 0.5 for each sex).

**AFM by SEX**

We work with composite groups from the cross-classification of AFM and SEX. The 12 groups are labelled: A0S1, A1S1, A2S1, A3S1, A4S1, A5S1, A0S2, A1S2, A2S2, A3S2, A4S2, A5S2.
These are indexed by $g = 1, \ldots G$ where $G$ is the number of groups. $G = 12$ in our frog data.

### 2.3 More Notation

The groups with unknown membership give rise to more notation:

$i \in g$ means individual $i$ is in group $g$,
$(Z_{ig})$ is an $I$ by $G$ indicator matrix where $z_{ig}$ = 1 if $i \in g$, otherwise 0; (missing information),
$\widehat{z}_{ig}$ is the current estimate of probability that $i \in g$,
$\widehat{x}_{ij|g}$ = estimated age of individual $i$ at occasion $j$ given $i \in g$,
$\mu_{ij|g}$ is the expected value of $y_{ij}$ given $i \in g$, assumed to be on a vB curve, and
$\widehat{\mu}_{ij|g}$ is the fitted (estimated expected) value of $y_{ij}$ given $i \in g$.

The **vB parameters** for SEX 1 and SEX 2 are (respectively):
$\alpha_1$, $\alpha_2$ = **a**symptotic length, sometimes called $L_\infty$,
$\beta_1$, $\beta_2$ = length at "**b**irth" (start of independent life, available for capture and measuring)
$\kappa_1$, $\kappa_2$ = instantaneous growth rate,
$\theta = \{\alpha_1, \beta_1, \kappa_1, \alpha_2, \beta_2, \kappa_2\}$, the set of all vB parameters.

The **proportion parameters** are $\pi_g$, $g = 1, \ \ldots \ G$, the proportions of the frogs falling in group (cluster, scenario) $g$. They are used as uninformative **prior probabilities** when updating $\widehat{z}_{ig}$.

$\phi = \{\theta, \pi\}$, the set of all vB and proportion parameters.

The $\widehat{z}_{ig}$ estimates are called the **posterior probabilities** that individual $i$ is in group $g$, after their updated estimates are found in the E step.

## 3 Probability model and distributions

For each individual there is a vector of observed SVL responses, and a predictor vector Age which will be estimated during the EM algorithm. Independence between different individuals is assumed, but the repeated measurements on individual $i$ are correlated. Thus each individual $i$ has its SVL vector $y_{ij}$ ($j = 1, \ \ldots \ J_i$) modelled by a non-linear multivariate regression, with its vector of expected SVL values $\mu_{ij}$ being the height of relevant vB curve (female or male) at the different ages. The probability distribution for individual $i$ is assumed to be multivariate normal, $N(\mu_{ij}, \mathbf{\Sigma})$ where $\mu_{ij}$ is the vector of expected values of $y_{ij}$ ($j = 1, \ \ldots \ J_i$) from the vB curve and $\mathbf{\Sigma}$ is the $J_i$ by $J_i$ variance-covariance (VCOV) matrix for the SVL measurements. With full information of Age and Sex available, this would be analysed as a **non-linear mixed-effects** (`nlme`) regression (Seber and Wild, 1989; Pinheiro and Bates, 2000). However, Age and Sex are not known.

The missing information of Age and Sex is handled in Bell and Pledger (2022) by wrapping an EM algorithm for finite mixtures (McLachlan and Peel, 2000) around an `nlme` analysis. Details and equations are in the next section. But there is another problem - many of the individuals have fewer than four measurements and cannot be included in the `nlme` analysis. Version 2 of this document will use a switch to the Fabens method (incremental growth) to allow individuals with only two or three measurements to contribute to the model fitting. Further, components of variance will also be modelled, to allow for individual variation in growth rate.

This paper is a proof of concept, that useful parameter estimates can be found despite the missing data. This preliminary analysis assumes that VCOV is diagonal. There is a possible

justified for this assumption if the data have relatively few observations on each of a large number of individuals; for a similar data set and advice, see Pinheiro and Bates (2000) section 6.4, the phenobarbital example. Our next version of this analysis will model the correlation structure and take account of rich versus sparse SVL measurement histories.

The multivariate normal R package `mvtnorm` is called in to provide probability densities and likelihoods during the calculations. The probability density function is $f(x, y; \phi)$, specifying the probability density of the $x, y$ data given the parameters $\phi$ (where $\phi$ consists of the set of vB parameters $\theta$ and the $\pi$ parameters). However, the same formula interpreted the other way around is $f(\phi; x, y)$, the likelihood of the parameters given the data. These concepts are used in the EM algorithm, described in detail below. Starting with guessed values for the $\phi$ parameters, the E (expectation) step builds probability densities to estimate the expected values of the missing data, then the M (maximisation) step uses the combined known and estimated data to update estimates for the parameters.

# 4 The EM Algorithm

The expectation-maximisation (EM) algorithm (Dempster, Laird and Rubin, 1977) was designed to estimate values of missing data while still providing maximum likelihood estimates of parameters of interest. It is initialised with guessed values for the parameters. Next the E (expectation) step builds probability densities using the known data to estimate the expected values of the missing data, then the M (maximisation) step uses the combined known and estimated data to find the maximum likelihood estimates for the parameters. But now the parameter estimates have changed, so it returns to the E step to rebuild the expected missing data. The algorithm continues to alternate the two steps, the E-step (expectation of missing data values) and the M-step (doing maximum likelihood estimation of the parameters of interest), until there are no further discernable changes in the parameter estimates when another EM loop is done, for example that the estimates are constant when taken to 5 decimal places and so acceptable convergence is reached. Dempster, Laird and Rubin (1977) found this idea being used in ad hoc ways by various authors in various areas of applied statistics, and they pulled the ideas together, gave them a common notation, and proved that there is continual improvement towards the true maximum-likelihood parameter estimates.

### Step 0: Initialise

Guess at $\theta$, the set of three starting parameters for each curve: $\theta = \{\alpha_1, \beta_1, \kappa_1, \alpha_2, \beta_2, \kappa_2\}$ for sexes 1 and 2. Without loss of generality let Sex1 label the curve with the higher asymptote. Also give starting estimates of $\pi_g$, the proportions of individuals in each group. This may be initilised as uninformative, e.g. a vector of length $G$ with all components $1/G$; it will soon get updated as the EM algorithm proceeds. The set of all these parameters, $\{\theta, \pi_g\}$ is labelled $\phi$.

### Step 1: E-step, Expectation

In the E step, the most recent estimates of $\phi$ (the vB parameters and the $\pi$ proportions) are used to update the estimated posterior probabilities $\widehat{z}_{ig}$, that individual $i$ is in group $g$.

For each individual $i$ and each group $g$, find the corresponding theoretical age vector $x_{ij|g}$ using the value of AFM from group $g$, complete the data with this age vector together with $i$'s SVL vector, (ii) Use the completed data to find the likelihood of parameters $\phi$ given the data. For individual $i$ and group $g$ this is

$$L_{ig} = \pi_g \prod_{j=1}^{J_i} f(\theta_g; y_{ij}, x_{ij|g})$$

where $f()$ is the multivariate normal probability density function of dimension $J_i$ and $\theta$ is the set of current vB parameters. The value of $g$ tells us which SEX to use for the $\theta_g$ parameters and which AFM to use for adjusting the Age vector. The likelihood $L_{ig}$ provides a measure of goodness of fit between the data for individual $i$ and the model for group $g$.

For individual $i$ the relative values of the likelihood over the different groups $g$ are what matters, where a high likelihood indicates a good fit of parameters to data. The values of the likelihood vector of $L_{ig}$ values (for $g = 1, \ldots G$ ) may be turned into probabilities by rescaling them to add to one. Dividing by the sum of the vector components ensures they add to one, which turns them into the posterior probabilities that $i \in g$:

$$\widehat{z}_{ig} = \frac{L_{ig}}{\sum_{g=1}^{G} L_{ig}}.$$

The posterior probabilities are put in the $\widehat{Z}$ matrix with rows $i = 1, \ldots I$ and columns $g = 1, \ldots G$ with row sums of 1. This is the expectation step, as $\widehat{z}_{ig}$ is the estimated expected value of a binary group-membership random variable ($Z_{ig} = 1$ if $i \in g$, otherwise 0).

### Step 2: M-step, Maximisation

The posterior probabilities for AFM and SEX from the E step (matrix $\widehat{Z}$) are now used together with the original data ($Y$ = SVL sequences and their time spacings for the individuals) to update estimates of the vB parameters $\theta$ (vector $(\alpha_1, \beta_1, \kappa_1, \alpha_2, \beta_2, \kappa_2)^T$) and the proportion parameters $(\pi_1, \ldots, \pi_G)^T$).

We treat the two types of missing information differently. The missing AFM is not of intrinsic interest; it is merely used to decide which Age vector to use for each individual. The generalised EM algorithm (GEM) permits simply plugging in the current best choice of AFM for each individual $i$, rather than trying to retain and carry forward the current posterior probabilities from the $\widehat{Z}$ matrix. This choice of Age vector is not set in stone; the position and shape of the fitted curves may change enough further into the algorithm for a different choice of age vector to be made. Build a pseudo-dataset which has the known columns ID, SVL, with one row per SVL measurement, then add in the selected Age vector for each ID. This is not a data set which actually occurred, as the Age vector is not really known, but we have inserted our best guess so far. The Age vector column will be updated every time the M step is run. If this GEM process of assigning the Age vectors converges, it will be to the correct MLE of the parameters (Turnbull, 1976; Dempster, Laird and Rubin, 1977; McLachlan and Krishnan, 1997). Turnbull referred to the pseudo-individuals, each with exact assignment to its best possible age vector, as "ghosts".

The other missing data, the Sex of each individual, is of particular importance, as we have a model with two curves. We do not assign each exactly to the most probable sex, but instead

retain the probabilities for updating throughout the EM cycles. At each M step, $i$ has estimated probability $p_1 = \sum_{g=1}^{M+1} \widehat{z}_{ig}$ of being of Sex 1, and estimated probability $p_2 = 1 - p_1$ of being of Sex 2. These values may be stored in the pseudo-data frame as columns `prob1` and `prob2`, as appropriate for each individual $i$, and updated each time through the M step.

The two fitted curves now have their parameter estimates updated. Here we use the non-linear least squares function `nls()` from the $R$ `nlme` package to update estimates of the two curves, using the current $p_1$ and $p_2$ as weights. The updated vB parameter estimates are ready to be used in the next E step.

Using `nls()` assumes the `pdDiag` correlation structure for the repeated measures, as suggested in Pinheiro and Bates (2000, Section 6.4) for data with many short SVL histories. Later versions of this work will use more detailed models.

### Step 3: Recycle to convergence

Continue to alternate steps 1 and 2 until convergence is reached - no change of log likelihood or parameter estimates. At each cycle the parameter estimates move closer to the maximum likelihood estimates. When there is (practically) no change in the estimates from one cycle to the next, effectively the maximum likelihood estimates have been reached. They are identical to as many decimal places as we specify.

## 5 Discussion

Assignment to AFM class is very close to the age estimation done by Bell in earlier work; our model here simply formalises that work in a probabilistic framework. We have confirmed that a finite mixture and EM algorithm may be wrapped around traditional growth curve analysis, to deal with the missing information of AFM and SEX.

The current model will be developed to use the Fabens method (dealing with the curves in increments), to model more appropriately the correlation stucture of the repeated measures within individuals, to test the models and to evaluate their accuracy at prediction of sex.

# 6 References


Armstrong D.P. and Brooks R.J. (2013). Application of hierarchical biphasic growth models to long-term data for snapping turtles. *Ecological Modelling* **250**, 119–125.

Dempster A.P., Laird N.M. and Rubin D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B* **39**, 1–38.

Fabens A.J. (1965). Properties and fitting of the von Bertalanffy growth curve. *Growth* **29** 265–289.

McLachlan, G.J. and Krishnan, T. (1997). *The EM Algorithm and Extensions.* John Wiley and Sons, New York.

McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models.* John Wiley and Sons, New York.

Pinheiro J.C. and Bates D.M. (2000). *Mixed-Effects Models in S and S-PLUS.* Springer-Verlag, New York.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Seber G.A.F. and Wild C.J. (1989). *Nonlinear Regression.* Wiley, New York.

Turnbull, B.W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society Series B (Methodological)* **38**, 290–295.

von Bertalanffy, L. (1960). Principles and theory of growth. In: Nowinski WW (Ed) *Fundamental spects of Normal and Malignant Growth.* Elsevier, Amsterdam, pp. 137–259.