`hmde`: Hierarchical Methods for Differential Equations

Tess O'Brien, Fonti Kar, David Warton, & Daniel Falster

January 21, 2025

Abstract

R package implementing a hierarchical Bayesian longitudinal model for repeat observation data. The package provides a selection of differential equation models that are then fit using the hierarchical model.

1 Introduction

`hmde` (https://github.com/traitecoevo/hmde) is an R package that fits a hierarchical model to estimate parameters of a differential equation (DE) in the presence of measurement error, where the parameters may vary randomly across subjects. We estimate DE parameters from repeated observations of the process over time. The motivating application for this package comes from ecology, where we take repeated observations (with error) of organism size and want to estimate growth trajectories, which may vary across individuals. In other language, `hmde` implements a set of hierarchical Bayesian longitudinal models to fit DEs to repeat observation data and is an example of a Bayesian inverse method. The package name stands for **h**ierarchical **m**ethods for **d**ifferential **e**quations. The underlying statistical method was first used in O'Brien et al. (2024) to model tree growth, and we continue to focus on ecological applications in this paper with additional models and taxa. The package was built in order to implement the method from O'Brien et al. (2024) as no software existed to do so. Existing methods such as non-linear mixed effects models have been used for tree growth with pair-wise difference data (de Miguel et al., 2013), and hierarchical models have fit species-level trajectories (Herault et al., 2011), but the addition of the individual longitudinal structure required a new approach. Building the software allowed for precise control over the hierarchical structure, differential equation solution, and estimation method, all of which hare proven relevant to the effort.

Three ordinary differential equation (ODE) models are provided: constant, an affine (ie. linear with translation) first-order ODE known as the von Bertalanffy model (Von Bertalanffy, 1938), and a non-linear first-order ODE we call the Canham growth function (Canham et al., 2004). The underlying method is more generally applicable to repeated observation data governed by other dynamics, and the package is intended to serve as a demonstration for how such models can be implemented in Stan (Stan Development Team, 2024).

From a structural perspective, `hmde` is a wrapper for Rstan (Stan Development Team, 2019) that provides a set of pre-built Stan models and some additional functionality to make analysis easier for the end user. Figure 1 demonstrates the workflow starting from longitudinal data where individuals

differ in their behaviour over time. By fitting a suitable ODE we can extract individual parameters that allow us to fit functions to each organism, and hence estimate a sequence of sizes over time that accounts for individual variation. O'Brien et al. (2024) showed that fitting individual ODEs in this fashion to data generated from the same process with measurement error was able to smooth out those errors and provide better estimates of the true sizes over time.

This paper details the theoretical structure of the underlying mathematics and statistics. We also go through the required data structures and include example data to show how the repeat observation structure is represented in the computer. Finally, we walk through demonstrations of each implemented model using provided example data, and outline some known statistical issues that may be encountered in user-developed models that rely on numerical integration.

2 The Maths

In this section we will describe the underlying mathematical model that `hmde` implements, introduce the provided functions, and give some guidance for how a user can determine if, or which, of the provided functions are suitable for their data.

2.1 Longitudinal model

For a sequence of values over time $Y(t_j)$ governed by the differential equation

$$\frac{dY}{dt} = f(Y(t); \boldsymbol{\theta})$$

we have

$$Y(t_{j+1}) = Y(t_j) + \int_{t_j}^{t_{j+1}} f(Y(t); \boldsymbol{\theta})\, dt \tag{1}$$

for some vector of parameters $\boldsymbol{\theta}$. We assume that f is known (or at least chosen), but the parameter vector needs to be estimated. In this way, we are dealing with an inverse method: an attempt to parameterise a differential equation based on observations of the process it governs.

2.2 Pre-built functions

We have implemented three models that are available for direct use with `hmde`: constant, von Bertalanffy, and Canham. These are not the only applicable models but were chosen to demonstrate a range of implementations across biology with varying data requirements and constraints. The models have one, two, and three parameters respectively, and demonstrate linear and non-linear dynamics. All of the DEs are time independent, in the sense that the time variable t does not appear on its own in the DE itself. The von Bertalanffy and Canham models depend on $Y(t)$ explicitly, which is referred to as size-dependent growth in the ecology literature.

Constant

The constant model is given by

$$f(Y(t); \beta) = \beta, \tag{2}$$

with β as the growth rate parameter. The constant model is mathematically equivalent to a linear model for sizes over time.

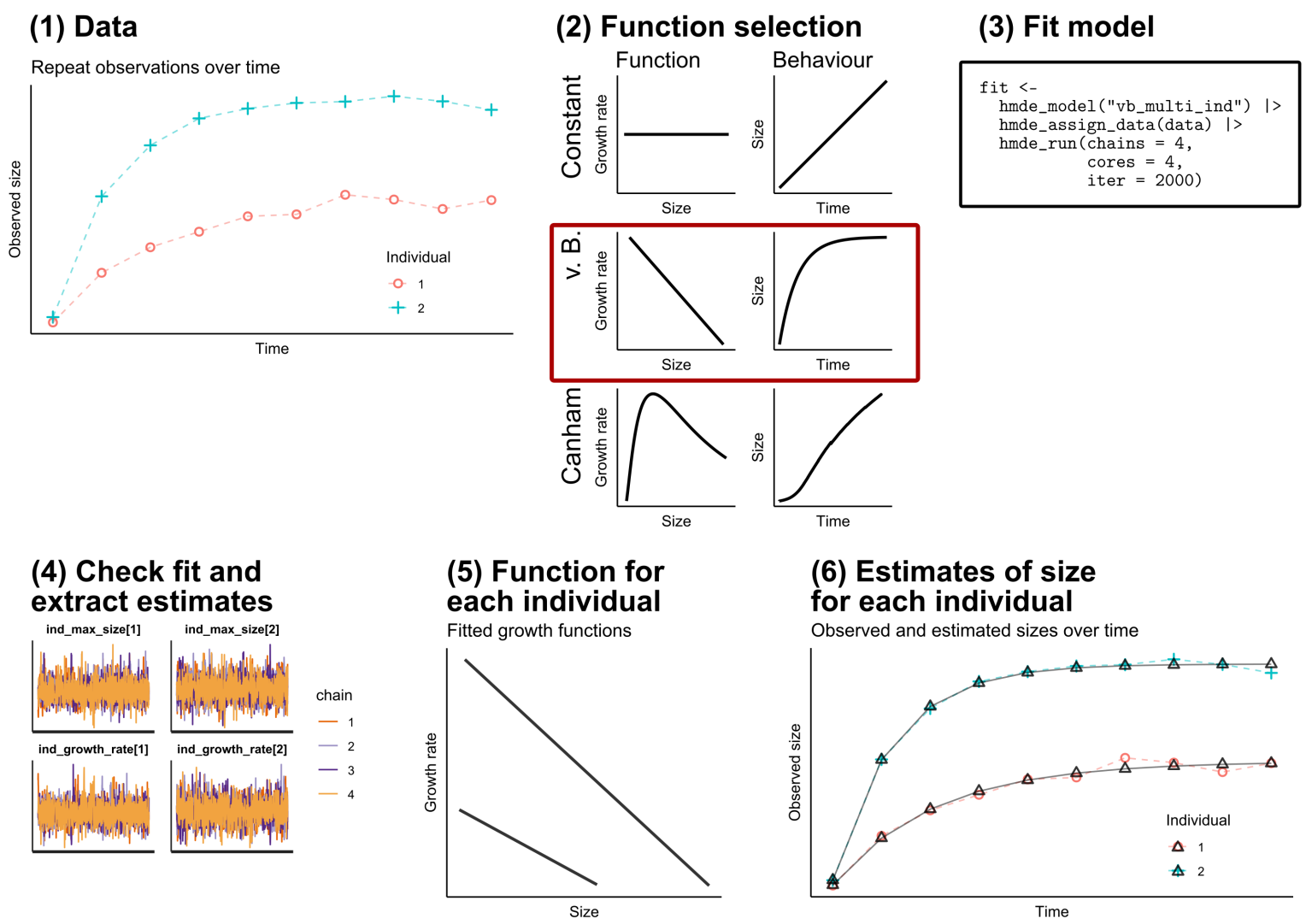

Figure 1: `hmde` is used for longitudinal data where one of three available functions would describe the growth appropriately.

von Bertalanffy

The von Bertalanffy model is given by

$$f(Y(t); S_{max}, \beta) = \beta(S_{max} - Y(t)), \tag{3}$$

where S_{max} is the asymptotic maximum size and β controls the growth rate. The implemented version in the Stan file is translated by a sample mean $\bar{y}$ in order to provide better estimation, which does not affect the function behaviour as there is a back-transformation for the output estimates.

Canham

The Canham function, developed in Canham et al. (2004), specifies a unimodal hump-shaped function for the derivative, given by

$$f(Y(t); f_{max}, Y_{max}, k) = f_{max} \exp\left(\frac{-1}{2}\left(\frac{\ln(Y(t)/Y_{max})}{k}\right)^2\right), \tag{4}$$

with f_{max} being the maximum growth rate, Y_{max} the Y-value at which the peak occurs, and k a spread parameter that controls how narrow or spread out the peak is.

2.3 Choosing a function

There are mathematical considerations for whether any of the existing models will work for a given data set. The constant and Canham models are constrained to have non-negative growth and Canham is not defined for $Y(t) \leq 0$. From a biological perspective non-negative growth may not always be true, for example an organism can reduce in size or mass particularly in the short term. In the constant function case this can be taken as an averaging constraint, while for Canham, the long time between observations of the trees in our demonstration data make non-negative growth a more reasonable assumption as we are smoothing out the behaviour in the five year period. The von Bertalanffy model can fit negative growth and negative $Y(t)$ even if the latter is biologically impossible. The typical use case for the von Bertalanffy model is for growth that declines as an organism approaches a maximum size, but if the data shows a value that is shrinking asymptotically to a value of $Y(t)$ (negative growth), the model can still be fit.

The second requirement is longitudinal data. The point of the model is to fit a function to the dynamics for each sampling unit (individual animal or plant in our case). We need to have repeated observations y_{ij} from that unit to do so, and know the time of each observation as well. Datasets of pairwise difference data are not suitable, the raw values should be used instead. Consistent identification of individuals is also important, but that is a matter of data cleaning and quality as we have not not implemented a structure that allows for misclassified individual.

The choice of DE depends on both data quantity and desired behaviour. From the data quantity perspective, fitting individual-level parameters requires a certain minimum number of observations for each individual, which varies across DEs. The constant function is the most generally applicable, requiring a minimum of two observations to estimate the single function parameter. The von Bertalanffy model has two parameters and a minimum of three observations. Canham, with three parameters, can theoretically be fit to four observations but we strongly recommend at least five. If the chosen

function is a good representation of the underlying dynamics, all the models will perform better with more observations per individual.

In terms of desired behaviour, we should choose a DE with a functional shape that captures the behaviours of interest. The constant model fits an average rate of change to each individual, smoothing out all dynamics. Such a model is unlikely to be realistic across a lifetime, but may be good enough if what is required is a distribution of average growth rates. The von Bertalanffy model assumes high growth at the smallest sizes, and an asymptotic maximum value which may not be observed in the data. The Canham model has accelerating growth at small sizes, and declining growth after the peak at (Y_{max}, f_{max}). As with the von Bertalanffy model, the decline to asymptotically 0 growth may not be observed in data, as seen in O'Brien et al. (2024).

2.4 Numerical methods and analytic solutions

Equation (1) invokes a differential equation that governs the dynamics of $Y(t)$ which we will need to solve to find $Y(t)$. We have implemented different methods to solve for $Y(t)$, depending on the specific DE involved. For the constant growth model we use the analytic solution which is given by

$$Y(t) = Y(0) + \beta t.$$

For the von Bertalanffy model, we have implemented the analytic solution

$$Y(t) = S_{max} + (Y(0) - S_{max}) \exp(-t\beta). \tag{5}$$

The Canham function in Equation (4) is extremely non-linear and does not have an analytic solution, so we must employ numerical methods. As Stan provides some ODE solvers, we use the inbuilt Runge-Kutta 4-5 solver for Canham which has an adaptive step size.

In the process of building `hmde` we encountered a very bimodal posterior distribution arising from numerical error in integration with the von Bertalanffy model. We have undertaken an extended investigation that can be found at `https://github.com/Tess-LaCoil/hmde-be-dragons`, while the Here Be Dragons vignette gives a short demonstration. In `hmde` itself the choice of integration methods avoids the bimodality for constant and von Bertalanffy models by implementing their analytic solutions directly, and for Canham by using the RK45 method which we found to give unbiased and unimodal results in simulation.

3 The Stats

The underlying statistical approach in `hmde` is a hierarchical Bayesian method for inverse problems: attempting to estimate parameters for a DE based on a chosen structure for known statistical relationships governing the distribution of those parameters, and observed data of the resulting process. We assume that the dynamics are continuous, but our observations are discrete and we have only finitely many of them.[1]

[1] In the Bayesian inverse methods literature finite observations with measurement error is known as a situation of practical identifiability (Latz, 2023) and presents additional difficulties due to limited information compared to infinite, zero-error data which is used theoretically. More work needs to be done to look at the theoretical side of the model `hmde` implements.

We have implemented two sets of models, the first if a single individual is included in the data, and the second if multiple individuals are, which adds a (hierarchical) cross-individual distribution for parameters. The underlying structure is the same with one additional level for multi-individual data. We have not implemented a multi-population model due to the computational constraints associated with doing so, and because in application to cross-species variation analysis we want to avoid shrinkage towards the mean that a multi-population model would encourage.

From the 'bottom' up, the multi-individual model has the following levels:

Measurement

Values of the process over time given by Equation (1), with an additional index to track individuals. We estimate the true value $Y_i(t_j)$ from on the observation $y_{i,j}$ with the estimator $\hat{Y}_{i,j}$ based on the error model

$$y_{i,j} \sim \mathcal{N}(\hat{Y}_{i,j}, \sigma_e). \tag{6}$$

In O'Brien et al. (2024) we demonstrated that a normality assumption works well in the case of a specific non-normal (but still symmetric and centred at 0) error process, so we are comfortable with that structure. In theory more general error models can be implemented but we have not done so.

Individual

The vector of parameters for f is fitted at the individual level, giving $\boldsymbol{\theta}_i$ as the parameter vector estimate for individual i. The form of f is fixed for a given model so the individual variation is encoded by different parameter values. For example in the constant model, we get β_i for each individual.

Population

Each parameter in the vector $\boldsymbol{\theta}_i$ comes from a distribution that operates at the population level. We have parameterised our chosen growth functions to use log-Normal priors on the population-level distribution. If we take the growth parameter β from the constant model for example, the prior is

$$\beta_i \sim \log \mathcal{N}(\mu_{\ln(\beta)}, \sigma_{\ln(\beta)}),$$

with population-level hyper-parameters that govern the mean and standard deviation of the log-normal distribution. The relationships between individuals within the population are entirely encoded by the parameter distributions.

Global

The hyper-parameters have their own priors, which are treated as independent across different parameters. We use a normal distribution for the means, and a half-Cauchy distribution for the standard deviation parameters:

$$\mu_{\ln(\beta)} \sim \mathcal{N}(0, 1), \qquad 0 < \sigma_{\ln(\beta)} \sim Cauchy(0, 1).$$

The σ_e error term from Equation (6) also has a prior at the global level given by

$$0 < \sigma_e \sim Cauchy(0, 2).$$

If we are fitting a model to a single individual, we lose the global level hyper-parameters and fit the individual parameters to a log-normal distribution with specified mean and standard deviation values, for example $\beta \sim \log \mathcal{N}(0.1, 1)$ for the constant model with a single individual. The σ_e distribution is preserved as we are interested in estimating the error behaviour.

3.1 Estimation

We use a Markov Chain Monte Carlo estimation process in `hmde`, which takes samples from the posterior distribution (Gelman et al., 2021). For each estimator we primarily use the mean of the posterior samples as the estimate, but we also provide the posterior median and a central 95% credible interval of samples for individual, population, and error parameters via the `hmde_extract_estimates` function. The user is able to extract their own estimates from the samples as the `hmde_run` function returns a Stan fit object that includes the sample chains themselves. For some applications the posterior distribution represented by the samples is of interest as well.

4 The Data

This section will detail the requirements for using user data with the `hmde` package and introduce the three demonstration datasets that come with the package.

4.1 Data structure

The heart of the statistical model `hmde` implements is a longitudinal structure for repeated observations over time. The package requires this structure in datasets, with at least two measurements per observational unit at different times, a record of when the measurements were taken to calculate the observation interval, and if there are multiple individuals then a way of identifying which individual each measurement comes from. The basic form is a table – usually a data frame or tibble (Müller and Wickham, 2024) – with columns for observations $y_{i,j}$, the observation index j counted for each individual, time $t_{i,j}$, and individual index i. In the following example data, i is represented as `ind_id` and j as `obs_index` which are the variable names `hmde` uses internally:

```
> head(Tree_Size_Data)
# A tibble: 6 × 4
ind_id  time y_obs obs_index
<dbl> <dbl> <dbl>     <dbl>
1      1  0      2.5         1
2      1  4.77   3.3         2
3      1  9.75   4.4         3
4      1 14.8    4.6         4
5      1 19.7    5           5
6      1 25.0    5.8         6
```

The implemented models have some additional requirements that are typically calculated from the basic framework. For models fitted to single individuals the core structure within the code is

- `y_obs`: a vector of real numbers of length `n_obs` which is the observations,

- `n_obs`: an integer giving the total number of observations which is automatically calculated from the length of `y_obs`,
- `obs_index`: a vector of integers of length `n_obs` giving the index j of observations.
- `time`: a vector of real numbers of length `n_obs` giving the time since the first observation.
- `y_bar`: a real number used for the von Bertalanffy model to centralise the data, typically the mean of the observed values. Automatically calculated from `y_obs`.

For the multi-individual model there are additional values:

- `ind_id`: vector of integers of length `n_obs` that gives the individual index i.
- `n_ind`: integer giving the number of individuals in the sample which is automatically calculated from the number of unique values in `ind_id`.

The `hmde_model()` function run on the name of a model will return the ordered list of names for what data needs to be passed to that particular model. Here's the multi-individual Canham model:

```
> hmde model("canham_multi_ind")
$n_obs
NULL

$n_ind
NULL

$y_obs
NULL

$obs_index
NULL

$time
NULL

$ind_id
NULL

$model
[1] "canham_multi_ind"

attr(,"class")
[1] "hmde_object"
```

4.2 Provided datasets

For demonstration purposes we have included four in-built datasets that are prepared for immediate use with the existing models. These data represent a range of taxa, both experimental and observational data, and are taken as prepared subsets of public datasets with permission from the original authors.

Trout size data

The trout size dataset is taken from the SUSTAIN trout data (Moe et al., 2020), a set of mark-recapture data for *Salmo trutta* in a land-locked population in Norway. We have taken a stratified sample of 50 individuals, where the strata are the number of observations from the individual fish:

- 25 individuals with 2 observations,
- 15 individuals with 3 observations,
- 10 individuals with 4 observations.

Within strata, the individuals were selected by a simple random sample without replacement from a sampling frame of individual IDs. The size of the trout is measured in centimetres from end to end. As the survey structure for the trout data requires re-capture, there is no control on the time between observations, which is measured in years. Due to the limitation on observations we have chosen this as our demonstration data for the constant function, which is not size-dependent.

Lizard size data

Our example lizard size data comes from experimental data used in Kar et al. (2024) from the species *Lampropholis delicata*. Size-structured growth based on a von Bertalanffy model has been used in the literature for other lizard and reptile species (Ramírez-Bautista et al., 2016; Shine and Charnov, 1992), so we considered this data a good match for that function. Measurements are in millimetres from the lizard's snout to the top of the cloacal opening (snout-vent-length or SVL). Time is measured in days. We took a simple random sample without replacement of 50 individuals using the individual IDs as a sampling frame.

Tree size data

The tree size data comes from Barro Colorado Island (Condit et al., 2019), in this case from the species *Garcinia recondita*.

A simple random sample of 50 individuals was taken from the 400 *G. recondita* individuals used in O'Brien et al. (2024), which were already processed through data cleaning and filtered to have 6 observations each (i.e. have survived the 25 years of observation) and checked for preservation of stem and tree IDs over time. Further filtration chose individuals with more than 3 cm observed difference between first and last size to avoid model fitting problems with the smaller dataset. Size is given as diameter at breast height (DBH) in centimetres, and time is measured in years.

5 The Workflow

All of the models provided in `hmde` leverage the same workflow. In the next sections we will walk through implementation but first we will detail the workflow in theory. As an example, let's say we choose the constant function as our model. The following code will take the provided trout size data, convert it in to the structure required to fit a constant model, fit the model, and extract the posterior estimates at measurement, individual, and population levels. The step label comments indicate the workflow process given in more detail below. We store the model fit object from step 4 in order to save it, and for traceplots or other diagnostic analysis.

```
# Constant function chosen as Step 1
trout_constant_fit <-
hmde_model("constant_multi_ind") |> #Step 2
hmde_assign_data(data = Trout_Size_Data)  |> #Step 3
hmde_run(chains = 4, cores = 4, iter = 2000) #Step 4
trout_estimates <-
hmde_extract_estimates(trout_constant_fit, #Step 5
input_measurement_data = Trout_Size_Data)
```

Here are the first few rows of each element in the estimates list:

```
> trout_constant_estimates
$measurement_data
# A tibble: 135 × 5
ind_id  time y_obs obs_index y_hat
<dbl> <dbl> <dbl>     <dbl> <dbl>
1      1  0       52         1  53.7
2      1  1.91    60         2  61.3
3      1  4.02    70         3  69.6

$individual_data
# A tibble: 50 × 5
ind_id ind_beta_mean ind_beta_median ind_beta_CI_lower
<int>      <dbl>         <dbl>          <dbl>
1    1        3.96          3.95           2.32
2    2        2.42          2.40           1.16
3    3        4.34          4.38           2.44

ind_beta_CI_upper
<dbl>
5.67
3.80
6.30

$error_data
```

```
# A tibble: 1 × 5
par_name            mean median CI_lower CI_upper
<chr>              <dbl>  <dbl>    <dbl>    <dbl>
1 global_error_sigma  3.93   3.90     3.21     4.81

$population_data
# A tibble: 2 × 5
par_name        mean median CI_lower CI_upper
<chr>          <dbl>  <dbl>    <dbl>    <dbl>
1 pop_beta_mu    0.882  0.894    0.578    1.12
2 pop_beta_sigma 0.470  0.473    0.117    0.795
```

In greater detail, the 5 steps to the fundamental workflow for `hmde` are:

1. **Choose model:** To see a list of model names run `hmde_model_names()`.

2. **Data structures:** `hmde` requires specific data structures for the Stan models. A detailed list of each can be seen by running the function `hmde_model` and giving it the name of the relevant model. The user can build each element in the list themselves, or construct a table with observations from each individual, an ordering of observations for the individual, the time at which they occurred, and information on which individual the observation came from. The table structure of `ind_id`, `time`, `y_obs`, and `obs_index` is the easiest to work with as `hmde_assign_data` is built to use those column names and the list output by `hmde_model` to give the Stan data structure.

3. **Convert data to Stan model structure:** The Stan model requires a specific list structure with agreement between the lengths of some vectors and corresponding integer values. We provide the `hmde_assign_data` function that converts provided data into the required list format and checks for the necessary size agreements. There are a handful of ways to pass datasets depending on the level of control required. A data frame or tibble structure with columns named `y_obs`, `obs_index`, `time`, and `ind_id` can be passed as the `data` argument and `hmde_assign_data` will format the list for the model. Specific list elements can be assigned directly by passing an argument of that name to `hmde_assign_data`, with the value of the argument being the data to be assigned.

4. **Run the model:** The function `hmde_run` loads the chosen model and runs the Stan MCMC sampler on the provided data, returning a Stan fit object. If no control parameters are provided the sampler will run 4 chains with 2000 iterations on a single CPU core as default for Stan. Sampler controls can be passed to `hmde_run` as they would to the sampler directly, for example `chains = 2` will run only two chains.

5. **Extract posterior estimates:** We provide `hmde_extract_estimates` to make posterior processing easier. The function takes the Stan fit object output by `hmde_run` and the data it was fit to in the tibble structure described in step 2. Which model was chosen is named in the Stan fit object. The output is a list of tibbles with posterior parameter estimates for each level in the hierarchical structure. The function gives the mean of posterior samples as the

estimate for $\hat{Y}_{i,j}$, but additionally the posterior sample median, and a central 95% credible interval for the individual- and population-level parameter estimates.

Steps can be strung together using the pipe operator `|>` (native to R, the Magrittr pipe `%>%` (Bache and Wickham, 2022) also works) which uses the output of a previous function as the **first** argument of the next.[2] We recommend retaining the output Stan fit object from `hmde_run` for diagnostic purposes.

We provide two additional functions for posterior analysis. The first is called `hmde_plot_de_pieces` which plots the fitted growth functions of all individuals, which can be useful for qualitative analysis. The second is `hmde_plot_obs_est_ind` which plots the observed and estimated sizes over time for different individuals, either a chosen number of them randomly selected from the data, or specified individuals based on a vector of ID values.

The following demonstrations align with the vignettes for each model, and demonstrate each included DE and associated provided dataset. We will go through the constant model in detail, summarise the most interesting results from the von Bertalanffy model, and give a brief overview of the Canham results, which are more extensively explored in O'Brien et al. (2024). Due to the random nature of MCMC sampling, re-running the code may get slightly different point estimates.

5.1 Constant growth fit to trout size data

In circumstances where the number of observations available per individual is very limited, average growth rates over time may be the only plausible model to fit. In particular, if there are individuals with only two size observations, than the best that can be done is a single estimate of growth rate based on that interval. Such a model behaves as constant growth, which we can think of as the average rate of change across the observation period and is given by Equation (2), β is the average growth rate across the observation period. The constant growth model corresponds to linear sizes over time, and is equivalent to a linear mixed model for size, where there is an individual effect when fit to multiple individuals.

Our example data for the constant model comes from Moe et al. (2020), a publicly available dataset of mark-recapture data for *Salmo trutta* in Norway. The time between observations is not controlled, nor is the number of observations per individual. As a result the data consists primarily of individuals with two observations of size, constituting a single observation of growth which limits the growth functions that can be fit to individuals as a single parameter model is the best that can be fitted to two sizes. The constant growth function in Equation (2) is the most appropriate of the functions we have in `hmde`, as we can use the single growth interval to estimate the average growth rate β.

To implement the workflow we fit the model and extract the estimates. We have already chosen the constant model for step 1, so to look at the required data structure we call `hmde_model("constant_multi_ind")`. As the provided trout data is already in the form required by the `hmde_assign_data` function we don't need to do any further pre-processing for step 2 and can pass the data frame directly to step 3 using `hmde_assign_data("constant_multi_ind", data = Trout_Size_Data)`. The following code includes command to run with multiple cores and demonstrates the use of the pipe operator to pass the required data structure from `hmde_model` to `hmde_assign_data`, then the correctly formatted list from

[2] Pipe is perhaps easiest thought of as a way to turn recursive function calls into an easier-to-read chain of them. From a mathematical perspective it is function composition where `f(x) |> g()` is the same as $g(f(x)) = g \circ f(x)$. Thankfully, the arrow points to the next function so the reading order is easier for English native speakers.

`hmde_assign_data` to `hmde_run` for step 4 that fits the model. The overall output of this part of the workflow is the model fit.

```
# Constant model chosen as Step 1
trout_constant_fit <-
hmde_model("constant_multi_ind") |> #Step 2
hmde_assign_data(data = Trout_Size_Data)  |> #Step 3
hmde_run(chains = 4, cores = 4, iter = 2000) #Step 4
```

Fitting the model is the most computationally expensive part, so we recommend saving the model fit at this point.

In step 5 we extract estimates for each level in the hierarchy using the function `hmde_extract_estimates`, which is set up to take the data as structured for the model fit.

```
trout_estimates <-
hmde_extract_estimates( #Step 5
fit = trout_constant_fit,
input_measurement_data = Trout_Size_Data)
```

We can directly compare the observed sizes over time to estimated values. For a quantitative fit metric we use R^2 calculated on $(y_{ij}, \hat{Y}_{ij})$. For qualitative analysis we look at scatter plots of observed and estimated sizes, and inspect plots of sizes over time, as in Figure 2(a) and (b). The R^2 statistic is a metric primarily used in linear regression that measures the proportion (ie. decimal value in the [0,1] interval) of variance in one coordinate that can be explained by the regression model. In this context, we interpret it as how strongly the fitted and observed values agree. We don't expect perfect agreement which would be $R^2 = 1$ because we don't get perfect agreement. O'Brien et al. (2024) showed that the change between observed and fitted values can actually correct for measurement errors in size, so disagreement is not a bad thing overall. In this case, $R^2 = 0.953$ and indicates strong agreement between the estimated and observed sizes even though we have chosen a very simplistic model in the constant function. We also look at the estimated and observed sizes directly using the `hmde_plot_obs_est_ind` function, and the fitted DEs using `hmde_plot_de_pieces`.

Figure 2(b) demonstrates that at the individual level, the constant growth function produces linear sizes over time that are averaging out the observed behaviour. In Panel (d) we see that there is a large range of estimated average growth rates, and a possible downwards trend for large sizes that would be better estimated with a size-dependent model.

The last pieces of analysis for the constant model are to look at the distribution of β_i, which we do with a histogram, and extract the species-level parameter estimates. Figure 2(d) shows a right-skewed, unimodal distribution of average growth rates with a tail of high values. The extreme 7 cm per year growth may not be biologically plausible, and that individual could be further investigated by identifying them from having the highest β_i value. Table 1 gives the posterior estimates for the hyper-parameters, with both the raw mean for the log-transformed distribution, and the exponent of that value which can be more easily interpreted as an average growth rate for the species in cm/yr.

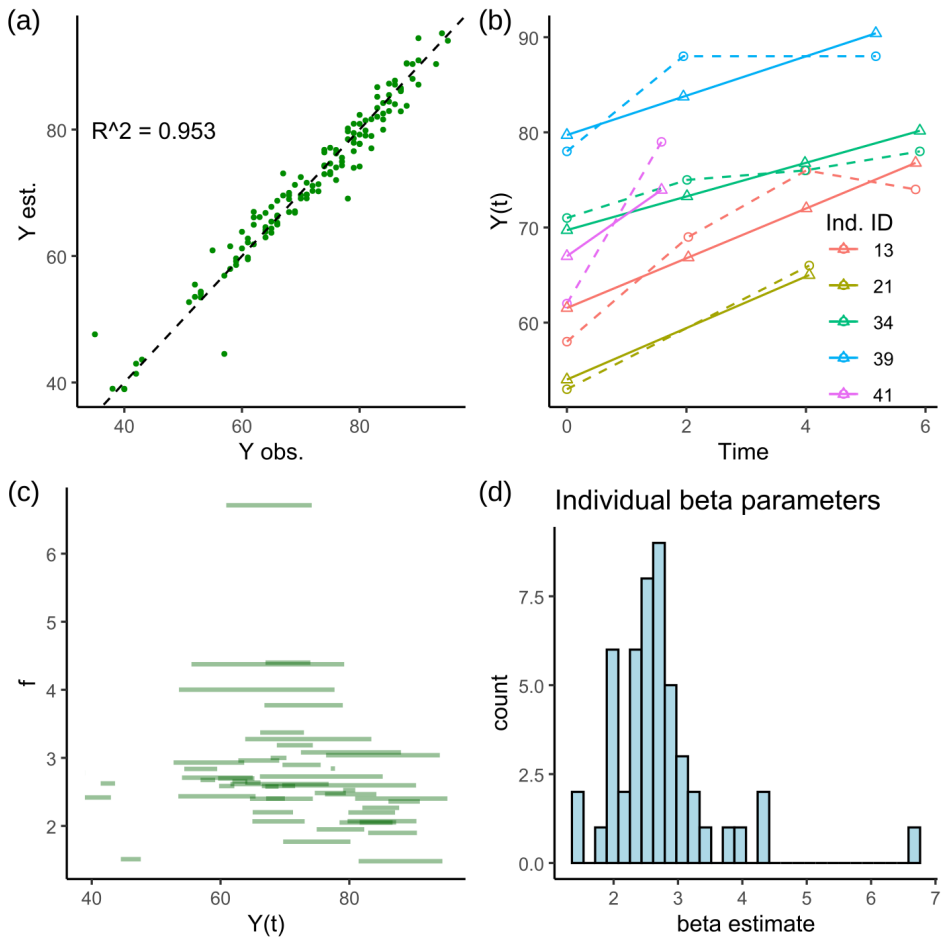

Figure 2: Analysis plots for *S. trutta*. **(a)** gives a comparison of estimated and observed sizes showing strong agreement. **(b)** gives observed and estimated sizes over time for five randomly selected individuals. **(c)** shows all fitted growth functions. **(d)** is a histogram of the $\hat{\beta}_i$s.

Sp. Par.	**Post. Mean**	**95% CI**
$\mu_{\ln(\beta)}$	0.881	(0.562, 1.121)
μ_β (cm/yr)	2.414	(1.754, 3.069)
$\sigma_{\ln(\beta)}$	0.467	(0.161, 0.816)

Table 1: Posterior estimates for species-level hyper-parameters in the constant model for trout.

5.2 von Bertalanffy fit to lizard size data

Our second example uses size-dependent growth based on the von Bertalanffy function given in Equation (3). The key behaviour of the von Bertalanffy model is a high growth rate at small sizes that declines linearly as the size approaches S_{max}. This manifests as growth slowing as an individual matures, with an asymptotic final size. We restrict β and S_{max} to be positive, and use the maximum of observed sizes as the prior mean for S_{max} to avoid model pathologies. As a result the growth rate is non-negative.

The data for the von Bertalanffy demonstration is the `Lizard_Size_Data` object provided with the package. As with the constant model, we pass the `Lizard_Size_Data` directly through the workflow to fit the model. The following code covers the steps in the workflow in two concise parts.

```
# von Bertalanffy model chosen for Step 1
lizard_fit <-
hmde_model("vb_multi_ind") |> #Step 2
hmde_assign_data(data = Lizard_Size_Data)  |> #Step 3
hmde_run(chains = 4, cores = 1, iter = 2000) #Step 4

lizard_estimates <-
hmde_extract_estimates( #Step 5
fit = lizard_fit,
input_measurement_data = Lizard_Size_Data)
```

We look at the plots of sizes over time and estimated growth functions to get a feel for plausibility based on how well the models fit the data. As we have two individual parameters, we can also look at the distribution of estimates in a scatter plot. If a user wishes to test for relationships between parameters across individuals, these estimates allow that to be done.

One interesting result of fitting individual trajectories is we can see that the von Bertalanffy model underestimates the largest sizes in Figure 3(a). The mathematical interpretation of this is that the straight line decay to 0 growth, and there may actually be an asymptotic decline instead. A problem for other work.

Finally, we look at the population-level parameter estimates. As in the constant growth case we will give both the raw and exponentiated values for the mean of the log-normal distribution to help with interpretation. The estimates and CIs are provided in Table 2. The estimate of $\mu_{Y_{max}} = 24.5\,\text{mm}$, the average value of maximum size across the population, is quite reasonable given what is known about the species *L. delicata*.

5.3 Canham fit to tree size data

Our final demonstration implements Equation (4) to model size-dependent growth in trees. We use the Barro Colorado Island data included with the package, which has six observations five years apart for each of 50 individuals from *G. recondita*. The Canham function is used as O'Brien et al. (2024) demonstrated that it performs well for the same data.

The following code runs the Canham model sampling and extracts samples. As sampling takes a few hours for this example we recommend using the provided set of estimates in the `Tree_Size_Ests`

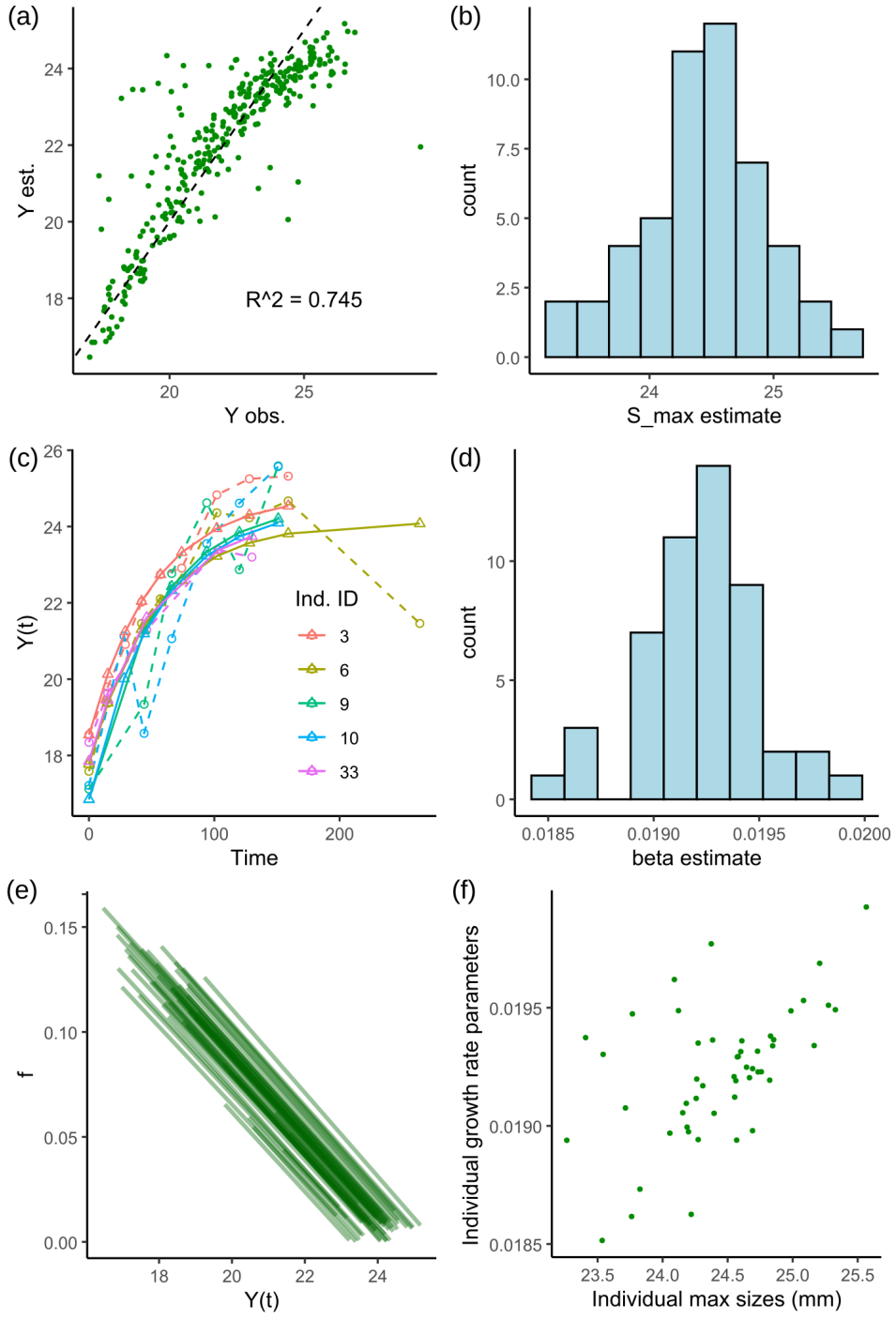

Figure 3: Plots for *L. delicata*. **(a)** gives a plot of observed and estimated $Y(t)$ values showing reasonable agreement, with some under-estimation of large sizes. **(c)** shows sizes over time for 5 randomly selected individuals. **(e)** gives the fitted von Bertalanffy functions for all individuals and shows that the slopes controlled by β are very similar, but there is more variation in the max size. **(b)** and **(d)** are histograms of individual parameters. **(f)** is a scatter plot of individual parameters.

Sp. Par.	Post. Mean	95% CI
$\mu_{\ln(Y_{max})}$	3.199	(3.180, 3.219)
$\mu_{Y_{max}}$ (mm)	24.51	(24.04, 25.01)
$\sigma_{\ln(Y_{max})}$	0.0342	(0.0135, 0.0495)
$\mu_{\ln(\beta)}$	-4.015	(-4.178, -3.808)
μ_{β}	0.0180	(0.0153, 0.0222)
$\sigma_{\ln(\beta)}$	0.0915	(0.0127, 0.285)

Table 2: Posterior estimates for species-level hyper-parameters in the von Bertalanffy model for lizards.

data object that comes with the package, and takes the place of the `tree_estimates` object in this code.

```
# Canham function chosen for Step 1
tree_fit <-
hmde_model("canham_multi_ind") |> #Step 2
hmde_assign_data(data = Tree_Size_Data)  |> #Step 3
hmde_run(chains = 4, cores = 4, iter = 2000) #Step 4

tree_estimates <-
hmde_extract_estimates( #Step 5
fit = tree_fit,
input_measurement_data = Tree_Size_Data)
```

Figure 4 demonstrates a strong alignment between the fitted growth functions and observed growth behaviour, both in Panel (a) with the randomly selected size trajectories and the observed and estimated sizes in Panel (c). Among the individual-level parameter estimates, the Spearman's rank correlation coefficients were 0.195 for g_{max} and S_{max}, -0.511 for g_{max} and k, and -0.257 for S_{max} and k, so we have some evidence of a moderate negative relationship between g_{max} and k, which in practice means that a spike to a higher growth rate is not sustained for a long period, while the lower growth rates with the higher k values are more sustained.

In Table 3 we see higher estimates for the g_{max} mean parameters compared to the estimates from O'Brien et al. (2024). Bias is to be expected, because we specifically chose individuals with more than 3 cm difference between max and min observed sizes as the sampling frame for `Tree_Size_Data`, which excludes a lot of very low-growth individuals that would pull down the distribution of g_{max}. The other parameters show strong agreement with overlap in the 95% CIs.

6 Discussion

`hmde` provides a baseline implementation of three hierarchical Bayesian longitudinal models in order to demonstrate the application of the method to different biological situations. The package offers an

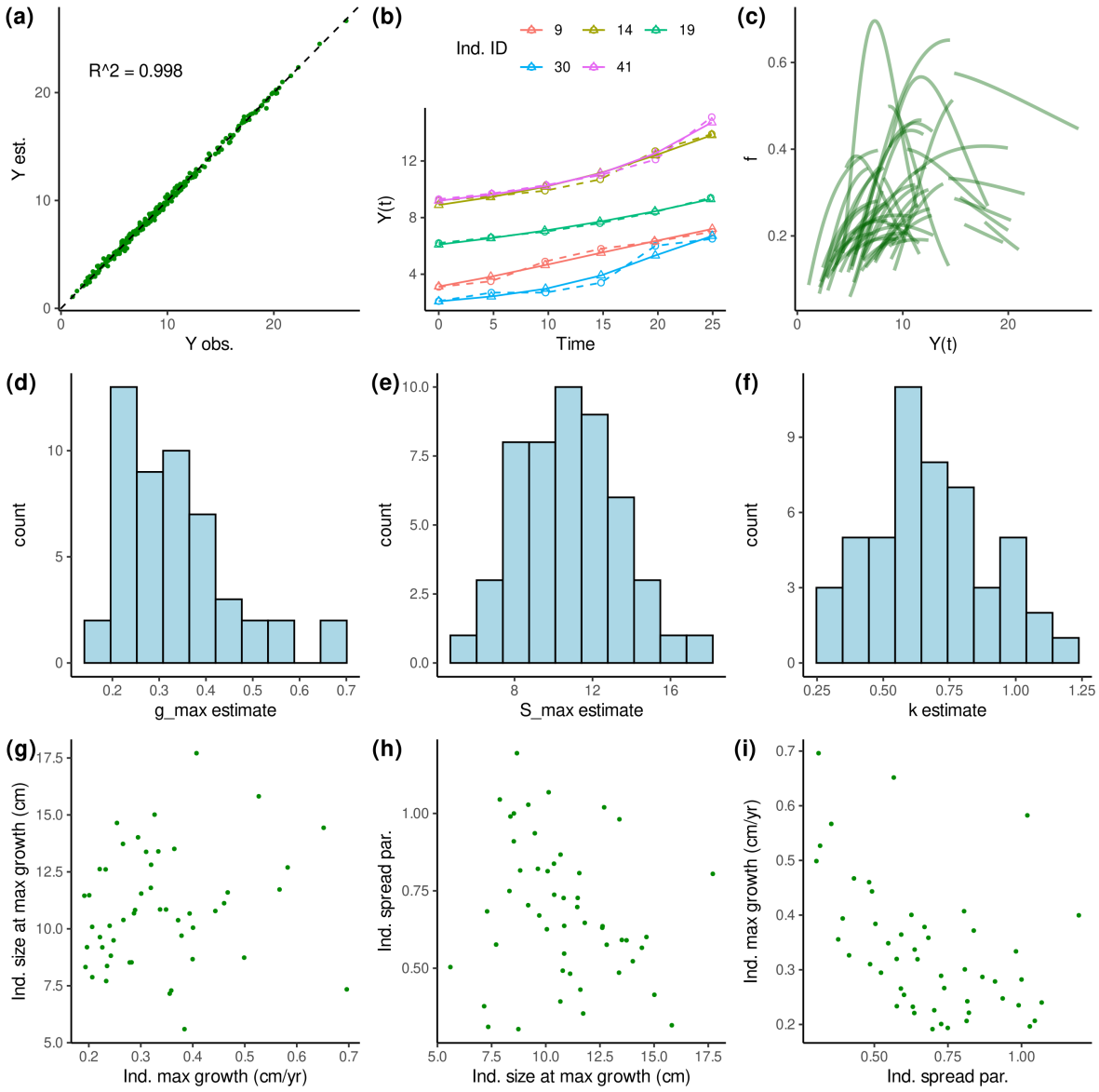

Figure 4: Analysis plots for *G. recondita*. **(a)** is a scatter plot of observed and estimated sizes showing very strong agreement. **(b)** shows sizes over time for 5 randomly selected individuals. **(c)** gives the fitted Canham functions for all individuals. **(d)**, **(e)**, and **(f)** are histograms of individual parameters, with the horizontal axis aligned to the pairwise scatter plots in **(g)**, **(h)**, and **(i)**.

Table 3: Posterior estimates for species-level hyper-parameters in the Canham model for trees, and the large sample (L. Samp.) estimates from O'Brien et al. (2024) for the same parameters on the larger sample.

Sp. Par.	Post. Mean	95% CI	L. Samp. est.	L. Samp. 95% CI
$\mu_{\ln(g_{max})}$	-1.178	(-1.39, -0.998)	-1.85	(-1.96,-1.74)
$\mu_{g_{max}}$ (cm/yr)	0.308	(0.248, 0.369)	0.16	(0.141, 0.176)
$\sigma_{\ln(g_{max})}$	0.434	(0.319, 0.574)	0.64	(0.56,0.71)
$\mu_{\ln(Y_{max})}$	2.30	(2.12, 2.48)	2.08	(1.96,2.20)
$\mu_{Y_{max}}$ (cm)	10.0	(8.30, 12.0)	8.0	(7.099, 9.025)
$\sigma_{\ln(Y_{max})}$	0.391	(0.225, 0.575)	0.47	(0.36,0.58)
$\mu_{\ln(k)}$	-0.552	(-0.861, -0.0662)	-0.24	(-0.40,-0.061)
μ_k	0.576	(0.423, 0.936)	0.787	(0.670, 0.9408)
$\sigma_{\ln(k)}$	0.565	(0.308, 0.842)	0.64	(0.51, 0.78)

interface to the methods used in O'Brien et al. (2024) which does not otherwise exist. We expand the application to non-tree taxa in this package with two animal demonstrations on *S. trutta* data from Moe et al. (2020) and *L. delicata* data from Kar et al. (2024).

While the method we implement is new, there are comparisons to existing ones. The constant model has a direct analogy as the equivalent is a linear model for size depending on time with an individual effect. The mathematical expression is

$$\begin{aligned} Y(t) &= \beta_0 + \beta_1 t && \text{(general model)}, \\ &= Y_{0,i} + \beta_{1,i} t && \text{(individual model)}, \end{aligned}$$

which linear mixed effects and hierarchical linear models can fit. The explicit inclusion of time enables a longitudinal structure. We include the constant model as the simplest use-case with minimal data requirements as some longitudinal surveys such as the mark-recapture surveys including the SUSTAIN trout data (Moe et al., 2020) can be dominated by individuals with only 2 observations. If a distribution of average growth rates is what the user wants, the constant model will provide it.

For the von Bertalanffy model, the analytic solution in Equation (5) is exponential and requires a non-linear form of function fitting. Ramírez-Bautista et al. (2016) used a non-linear model fit to pairwise difference data, and fit the von Bertalanffy model at the population level with a sex effect rather than fitting individual effects. In contrast we fit an individual longitudinal model where the population-level behaviour is encoded in the distribution of individual parameters.

The Canham model is the stand-out of the three implemented in `hmde`, as the absence of an analytic solution to Equation (4) means we have to use numerical methods to encode the longitudinal structure. Fitting the solution through the longitudinal model given in Equation (1) contrasts to population-level average trajectory models fit as functions to pair-wise difference data such as Herault et al. (2011). The population model operates at the level of the DE and does not attempt to estimate a size over

time trajectory for individuals. We get access to the individual trajectories directly, enabling deeper analysis of behaviour within a population.

By providing demonstrations from three different taxa we show some of the applicability of a very general underlying method for longitudinal data. We chose to implement three models: constant (average) growth, von Bertalanffy (linear first order ODE), and the Canham (Canham et al., 2004) function (non-linear first order ODE). These are a quite restricted set of models and do not investigate dynamics in more than one dimension for Y which is a possible expansion. Even within tree growth modeling there are many other functions for dynamic processes (Herault et al., 2011).

7 Conclusions

The `hmde` package provides a relatively user-friendly structure for implementing the chosen set of models with a hierarchical Bayesian longitudinal method. We have given demonstrations of each model matched to an included dataset of repeat survey or experimental measurements that suit the chosen function.

References

Bache, S. M. and Wickham, H. (2022). *magrittr: A Forward-Pipe Operator for R*. https://magrittr.tidyverse.org, https://github.com/tidyverse/magrittr.

Canham, C. D., LePage, P. T., and Coates, K. D. (2004). A neighborhood analysis of canopy tree competition: effects of shading versus crowding. *Canadian Journal of Forest Research*, 34(4):778–787.

Condit, R., Pérez, R., Aguilar, S., Lao, S., Foster, R., and Hubbell, S. (2019). Complete data from the barro colorado 50-ha plot: 423617 trees, 35 years, 2019 version.

de Miguel, S., Guzmán, G., and Pukkala, T. (2013). A comparison of fixed-and mixed-effects modeling in tree growth and yield prediction of an indigenous neotropical species (centrolobium tomentosum) in a plantation system. *Forest Ecology and Management*, 291:249–258.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2021). *Bayesian data analysis*. CRC Press, 3rd april 2021 edition.

Herault, B., Bachelot, B., Poorter, L., Rossi, V., Bongers, F., Chave, J., Paine, C. T., Wagner, F., and Baraloto, C. (2011). Functional traits shape ontogenetic growth trajectories of rain forest tree species. *Journal of ecology*, 99(6):1431–1440.

Kar, F., Nakagawa, S., and Noble, D. W. (2024). Heritability and developmental plasticity of growth in an oviparous lizard. *Heredity*, 132(2):67–76.

Latz, J. (2023). Bayesian inverse problems are usually well-posed. *SIAM Review*, 65(3):831–865.

Moe, S. J., Nater, C. R., Rustadbakken, A., Vøllestad, L. A., Lund, E., Qvenild, T., Hegge, O., and Aass, P. (2020). Long-term mark-recapture and growth data for large-sized migratory brown trout (salmo trutta) from lake mjøsa, norway. *Biodiversity Data Journal*, 8.

Müller, K. and Wickham, H. (2024). *tibble: Simple Data Frames*. R package version 3.2.1, https://github.com/tidyverse/tibble.

O'Brien, T., Warton, D., and Falster, D. (2024). Yes, they're all individuals: Hierarchical models for repeat survey data improve estimates of tree growth and sie. *Methods in Ecology and Evolution*, 16(1).

Ramírez-Bautista, A., Hernández-Salinas, U., and Zamora-Abrego, J. G. (2016). Growth ecology of the tree lizard urosaurus bicarinatus (squamata: Phrynosomatidae), in a tropical dry forest of the chamela region, mexico. *Animal Biology*, 66(2):189–199.

Shine, R. and Charnov, E. L. (1992). Patterns of survival, growth, and maturation in snakes and lizards. *The American Naturalist*, 139(6):1257–1269.

Stan Development Team (2019). RStan: the R interface to Stan. R package version 2.19.2.

Stan Development Team (2024). Stan modeling language users guide and reference manual. Version 2.34.

Von Bertalanffy, L. (1938). A quantitative theory of organic growth (inquiries on growth laws. ii). *Human biology*, 10(2):181–213.