

VU Research Portal

Comparing loss functions and interval estimates for survival data

Augustine, Starrlight; Lika, Konstadia; Kooijman, Sebastiaan A.L.M.

published in

Ecological Modelling
2020

DOI (link to publisher)

[10.1016/j.ecolmodel.2020.109077](https://doi.org/10.1016/j.ecolmodel.2020.109077)

document version

Publisher's PDF, also known as Version of record

document license

Article 25fa Dutch Copyright Act

[Link to publication in VU Research Portal](#)

citation for published version (APA)

Augustine, S., Lika, K., & Kooijman, S. A. L. M. (2020). Comparing loss functions and interval estimates for survival data. *Ecological Modelling*, 430, 1-4. Article 109077. <https://doi.org/10.1016/j.ecolmodel.2020.109077>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:

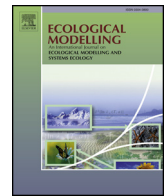
vuresearchportal.ub@vu.nl



ELSEVIER

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Comparing loss functions and interval estimates for survival data

Starrlight Augustine^{*,a}, Konstadia Lika^{b,c}, Sebastiaan A.L.M. Kooijman^d^a Akvaplan-niva, Fram High North Research Centre for Climate and the Environment, Tromsø, 9296, Norway^b Department of Biology, University of Crete, Heraklion, 70013, Greece^c Institute of Computational Mathematics, Foundation of Research and Technology Hellas, Heraklion, 70013, Greece^d Department of Theoretical Biology, VU University Amsterdam, de Boelelaan 1087, Amsterdam, 1081 HV, the Netherlands

ARTICLE INFO

Keywords:

Dynamic energy budget theory
Add-my-Pet
Survival data
Aging
Maximum likelihood
Symmetric bounded loss function
Profile-based interval estimates

ABSTRACT

We compare parameter point and interval estimates based on the symmetric bounded loss function, as used in the Add-my-Pet collection on animal energetics, with the maximum likelihood method for number of surviving individuals as function of time. The aging module of Dynamic Energy Budget theory is used to generate Monte Carlo data sets. The simulations show that estimates based on the symmetric loss function give almost the same results in terms of point as well as interval estimates, compared to maximum likelihood estimation, while this loss function avoids the need to model the stochastic component of data sets. For most data types on energetics, we don't have such stochastic models, so maximum likelihood methods cannot be used. Our findings support the view that model plasticity dominates interval estimates, rather than the detailed structure of the stochastic component.

1. Introduction

The Add-my-Pet (AmP) collection (AmP, 0000; Marques et al., 2018) is about referenced data on animal energetics, parameter estimates of Dynamic Energy Budget (DEB) models and implied properties. The data concern many aspects of energy budgets, such as growth, reproduction, development, respiration, feeding, excretion, and of life history, such as age and weight at birth and puberty, lifespan etc. The simultaneous estimation of parameters from a variety of data sets require a special type of estimation methods (Lika et al., 2020). The large size of the AmP collection, presently it has over 2000 animal species, motivates for research on the best estimation procedures. Morais et al. (2018), for instance, compared the Nead Melder simplex (NM) method with a Directional Direct-Search method, concluding that they don't differ much provided that the NM method is done with a series of a limited number of steps, using continuation.

The present study focuses on the loss function. Can the present choices be improved? Marques et al. (2019) explains why the default AmP parameter estimation is based on the (parameter-free) symmetric bounded loss function (Marques et al., 2019), and not on the maximum likelihood (ML) method. Important arguments are that, generally, models are lacking for the stochastic aspect in energetics data and to minimize the number of parameters that are involved. On top of that, AmP uses zero-variate data sets, i.e. data sets that consist of a single

number, such as an incubation time or a maximum body weight. It is impossible to meaningfully include a stochastic component in such a single number, while such a number contains valuable information about an aspect of energetics Lika et al. (2019), in combination with other data.

Data on survival of individuals as function of time is an exception to the problem that ML methods cannot be used, since this data is dimensionless and a natural and parameter-free model exists for the stochastic component in the case that individuals die independently. Since the ML criterion is typically superior for large samples, we here compare the performance of both loss functions.

2. Methods

2.1. Loss functions

The symmetric bounded loss function that is used as a default in the AmP procedure is

$$F_{sb} = \sum_{i=1}^n \sum_{j=1}^{n_i} \frac{w_{ij} (d_{ij} - p_{ij})^2}{n_i (d_i^2 + p_i^2)} \quad \text{with } d_i = \sum_{j=1}^{n_i} \frac{d_{ij}}{n_i} \quad \text{and } p_i = \sum_{j=1}^{n_i} \frac{p_{ij}}{n_i}$$

where w 's are weights, d 's data, p 's predictions, j scans data-points with a data-set of n_i points ($n_i = 1$ is allowed), i scans the n data-sets. The motivation for this choice is given in Marques et al. (2019), and

* Corresponding author.

E-mail address: sta@akvaplan.niva.no (S. Augustine).

<https://doi.org/10.1016/j.ecolmodel.2020.109077>

Received 23 January 2020; Received in revised form 6 April 2020; Accepted 7 April 2020

Available online 27 May 2020

0304-3800/ © 2020 Elsevier B.V. All rights reserved.

alternatives are discussed as well.

In the situation where dataset i represents number of surviving individuals as function of time, the loss function can naturally be based on the ML criterion for the multinomial distribution, see Kooijman (1981, 1983). In view of the fact that minus the log likelihood counts as a loss function, assuming that all individuals die independently with the same (time-dependent) probability, the loss function for the ML criterion becomes

$$F_{ML} = - \sum_{i=1}^n \frac{1}{n_i} \left(d_{i,n_i} \ln p_{i,n_i} + \sum_{j=1}^{n_i} (d_{ij} - d_{i,j-1}) \ln(p_{ij} - p_{i,j-1}) \right)$$

where d_{ij} is the number of surviving individuals at time t_{ij} and $p_{ij} = d_{i0} S_{ij}$ the predicted number, where survival probability S_{ij} depends on model parameters, and d_{i0} is the number of individuals at time $t_{i0} = 0$. Terms that do not depend on parameters are omitted. The initial number of individuals, d_{i0} , has no stochasticity and is in fact a design parameter. Notice that, while data-point j of data-set i can have its own weight coefficient w_{ij} in F_{sb} , this is not the case using the maximum likelihood criterion where all data-points have equal weight.

In the case of multiple data sets, deviations from predictions can be quantified by the mean relative error (MRE), i.e., the mean of the relative error (RE_i) for each data set i : $MRE = \frac{1}{n} \sum_{i=1}^n RE_i$, where $RE_i = \sum_{j=1}^{n_i} \frac{w_{ij} |p_{ij} - d_{ij}|}{w_i |d_{ij}|}$ and $w_i = \sum_{j=1}^{n_i} w_{ij}$, where, generally, $w_{ij} \geq 0$ for the AmP procedure and $w_{ij} = 1$ for the maximum likelihood approach; in this paper we use $w_{ij} = 1$ for both F_{sb} and F_{ML} .

2.2. Survival data

To compare the performance of both loss functions, we use the simplified DEB module for aging (Kooijman, 2010, Eq. (6.5)), where the survival probability S for age t amounts to

$$S(t) = \exp \left(\frac{6\dot{h}_W^3}{\dot{h}_G^3} \left(1 - \exp(\dot{h}_G t) + \dot{h}_G t + \frac{\dot{h}_G^2 t^2}{2} \right) \right) \xrightarrow{\dot{h}_G \rightarrow 0} \exp(-(\dot{h}_W t)^3)$$

with cubed Weibull aging rate $\dot{h}_W^3 = \dot{h}_a \dot{k}_E / 6$, Gompertz aging rate $\dot{h}_G = s_G \dot{k}_E$, while Weibull aging acceleration \dot{h}_a and Gompertz stress coefficient s_G are parameters of the aging module, and length-specific energy conductance \dot{k}_E is a parameter of the reserve mobilisation module of DEB theory. The simplifications are that food availability is assumed to be large and the growth period is short, relative to the mean life span. This (mechanistic) model, which has the famous empirical models by Weibull (1951) and Gompertz (1825) as special cases, results from the idea that respiration involves damage inducing compounds, which generate damage compounds that affect the hazard rate (Leeuwen et al., 2002).

Energy parameter \dot{k}_E is kept fixed, since the number of surviving individuals as function of time, as stand-alone data, does not provide information about this parameter. This leaves parameters \dot{h}_a and s_G to be estimated.

We generated data sets of surviving individuals as function of time, using a random number generator, estimated parameters \dot{h}_a and s_G , and evaluated the mean relative error.

If the different times of counting surviving individuals belong to different experiments, the number of surviving individuals is not non-increasing in time by necessity. We here assume, however, that we follow a cohort of individuals in time, so the number of surviving individuals cannot increase in time. The Monte Carlo data sets took the number of surviving individuals N_i at time t_i equal to that at time t_{i-1} minus the number of dead individuals between t_{i-1} and t_i . The probability of death for each of the N_{i-1} individuals, conditionally that they were alive at that time, is $(S_{i-1} - S_i) / S_{i-1}$, where S_i is the survival probability at time t_i . Fig. 1 presents an example of a simulated data set.

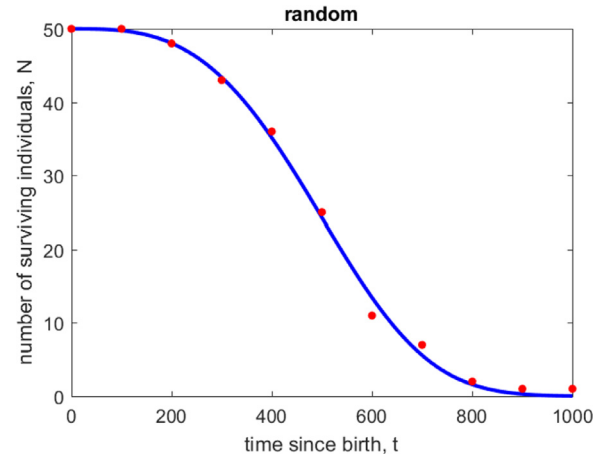


Fig. 1. A typical example of a simulated data set (red dots) generated from the DEB aging model (blue curve), starting from 50 individuals. Model parameters: length-specific energy conductance (kept fixed) $\dot{k}_E = 0.025 \text{ d}^{-1}$, Weibull aging acceleration $\dot{h}_a = 10^{-6} \text{ d}^{-1}$, Gompertz stress coefficient $s_G = 10^{-2}$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.3. Point estimates and confidence intervals

The point estimates are obtained from the minimization of the loss functions, which is done with a Nelder Mead simplex method (McCullagh and Nelder, 1983). The interval estimates are calculated on the basis of loss function profiles, starting from the value for a particular parameter, θ , that minimizes the loss function, F^{\min} , moving the parameter step-wise up or down, while estimating the other parameters (Marques et al., 2019; Stavrakidis-Zachou et al., 2019).

For the AmP symmetric bounded loss function, F_{sb} , a $100(1 - \alpha)\%$ confidence interval for θ is given by

$$\{\theta: F_{sb}(\theta) - F_{sb}^{\min} < F_c\}$$

Since the distribution of the symmetric bounded function is generally unknown, we obtain the empirical cumulative distribution function of the global minima that correspond with many Monte Carlo simulations of data (e.g., 500 or 1000), generated by adding a scaled log-normally distributed error to the predicted value with zero mean and the same coefficient of variation (cv). The coefficient of variation is computed as the mean of the absolute difference between observed and predicted values divided by the predicted value; $cv = \frac{1}{n} \sum_{i=1}^n \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{|d_{ij} - p_{ij}|}{p_{ij}}$. The reason to define the cv as a relative measure comes from the need to solve the dimension problem when more than one type of data set is used. The cumulative distribution is then used to find the critical value F_c that corresponds to a given confidence level $1 - \alpha$, i.e., $P(F_{sb} \leq F_c) = 1 - \alpha$.

For the loss function based on maximum likelihood criterion, F_{ML} a $100(1 - \alpha)\%$ confidence interval for θ is given by

$$\{\theta: 2(F_{ML}(\theta) - F_{ML}^{\min}) < \chi_{\nu, 1-\alpha}^2\}$$

In this case the confidence interval is obtained from the theoretical results that, asymptotically, $-2 \ln LR \sim \chi_{\nu}^2$, where LR is the ratio of two likelihood functions and ν is the difference in the number of parameters to be estimated. The critical value, $\chi_{\nu, 1-\alpha}^2$ is defined by $P(X \leq \chi_{\nu, 1-\alpha}^2) = 1 - \alpha$, where $X \sim \chi_{\nu}^2$.

3. Results

To evaluate the performance of the two loss functions F_{ML} and F_{sb} , we generated 500 Monte Carlo data sets, as in Fig. 1. For each data set the aging acceleration \dot{h}_a and the Gompertz stress coefficient s_G were

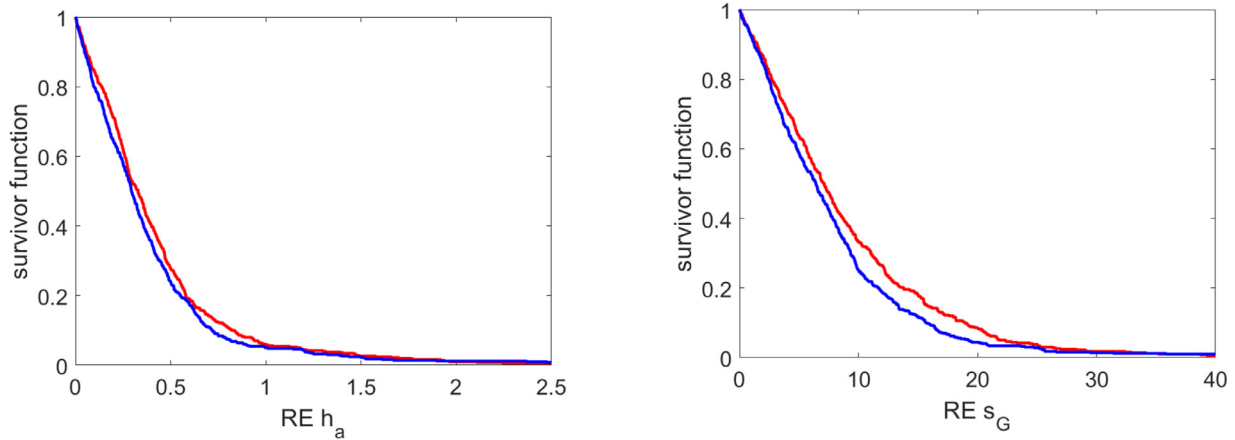


Fig. 2. The survivor functions (i.e. the fraction that exceeds the value of the argument) for the relative error of the aging acceleration \hat{h}_a (left) and the Gompertz stress coefficient s_G (right) for the AmP's symmetric bounded loss function (red) and the Maximum Likelihood criterion (blue). The Monte Carlo study involved 500 trial data sets. Observation times and model parameters as in Fig. 1. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

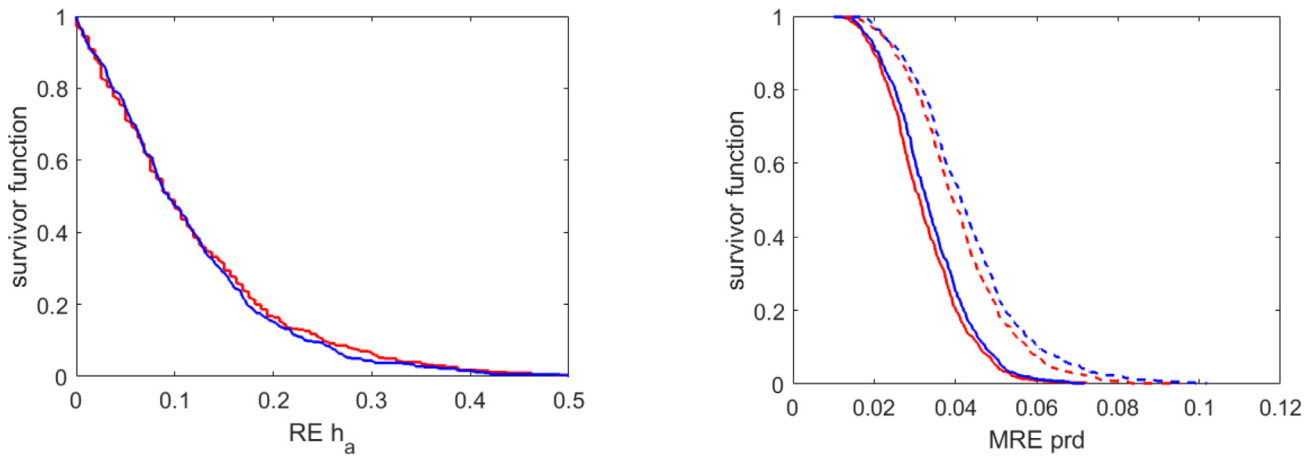


Fig. 3. The survivor functions for the relative error of the aging acceleration \hat{h}_a given the Gompertz stress coefficient s_G (fixed at the true value) for the AmP's symmetric bounded loss function (red) and the Maximum Likelihood criterion (blue). Simulation and model parameters as in Fig. 1. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 4. Survivor functions of the mean relative error of the survival predictions (MRE_{prd}) for the AmP's symmetric bounded loss function (red) and the Maximum Likelihood criterion (blue) when both the aging acceleration \hat{h}_a and the Gompertz stress coefficient s_G are estimated (solid lines) and when \hat{h}_a is estimated given s_G (fixed at the true value) (dotted lines). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

estimated. The relative deviations between the estimated and the true value was computed and their distribution is presented in Fig. 2. This figure shows that loss function F_{ML} and F_{sb} perform very similarly. However, we observed a strong negative correlation between estimated values of \hat{h}_a and s_G . Therefore, we also present the results in Fig. 3, where s_G has been kept fixed at its true value. Notice that for both loss functions the relative errors are now much smaller and the survivor functions (i.e. the fraction that exceeds the value of the argument) very similar.

Fig. 4 confirms the results of Fig. 3 that the high REs for the parameters is not the result of bad fits, but of the high negative correlation between the estimates of \hat{h}_a and s_G . The median MRE for the predictions is 0.03 (\hat{h}_a and s_G are estimated) and 0.04 (\hat{h}_a estimated, s_G fixed), while the median MRE for all the 70 thousand predictions in the AmP collection is about 0.06, which is still very low. Our case study is thus representative for the practical situation in AmP. Fig. 1 shows that the scatter for 50 initial individuals is small indeed. Fig. 4 also shows that the predictions based on F_{sb} and F_{ML} are also very similar.

We changed parameters settings, initial number of individuals and observation times, but the reported results are typical. E.g. with 30 initial individuals, rather than 50, the median MRE for the predictions

is 0.04, rather than 0.03, which is consistent with our expectations that if the number of initial individuals does down, the scatter goes up, so does the MRE.

Figs. 5 and 6 compare the profile-based confidence intervals. The target-parameter is fixed at a range of values, while the other is estimated, and the values of the loss function are plotted. The intervals at a particular uncertainty level are the set of values lower a cut-off, as defined in Section 2.3. The cut-offs for a 90% marginal confidence interval for F_{ML} is $\chi_{1,0.1}^2 = 2.71$ and for F_{sb} is $F_c = 0.0019$ (both \hat{h}_a and s_G are estimated) and $F_c = 0.002$ (\hat{h}_a is estimated and s_G is fixed). To obtain the cut-off values of F_c , 500 Monte Carlo data sets were generated by adding centered log-normally distributed error to the predicted value and scatter with $cv = 0.04$. The value of the cv is the mean of cv -values from a number of data sets after omitting the last 2–3 points (that correspond to probability of surviving beyond that point < 0.1) to avoid division by zero.

The figures show that the shapes of the profiles are very similar again, despite the fact that the formulas for the loss function look rather different, and that the resulting intervals are also very similar. Notice that the confidence intervals for the aging acceleration \hat{h}_a are reduced

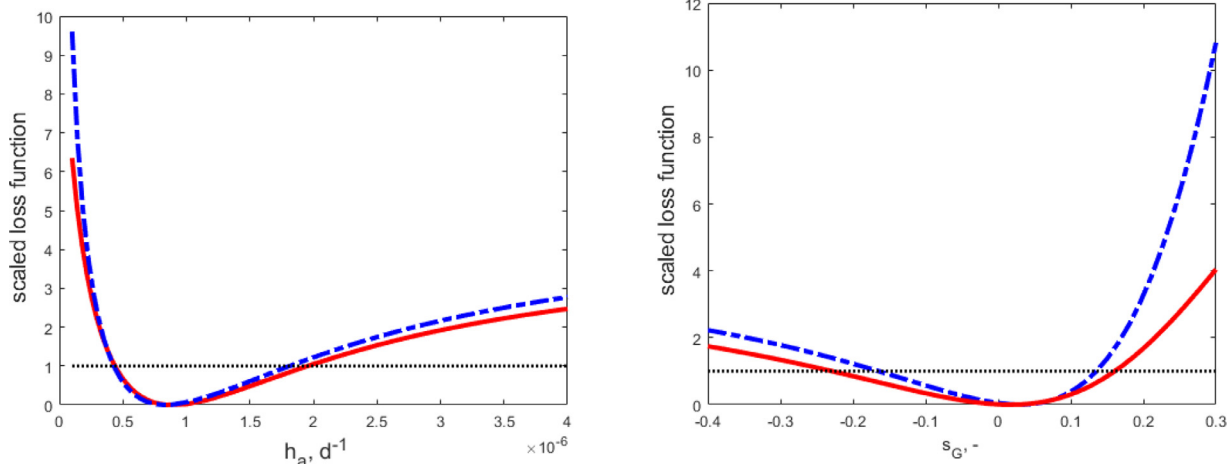


Fig. 5. Profiles of the scaled AMP's symmetric bounded loss function $(F_{sb}(\theta) - F_{sb}^{min})/F_c$ (red solid line) and the loss function based on maximum likelihood criterion $2(F_{ML}(\theta) - F_{ML}^{min})/\chi^2_{v;1-\alpha}$ (blue dashed-dotted line) for \hat{h}_a (left) and s_G (right). The 90% marginal confidence intervals for the aging acceleration \hat{h}_a are: $(0.0431 \cdot 10^{-5}, 0.1920 \cdot 10^{-5})$ using F_{sb} and $(0.0435 \cdot 10^{-5}, 0.1795 \cdot 10^{-5})$ using F_{ML} . The CIs for s_G are: $(-0.2304, 0.1577)$ using F_{sb} and $(-0.1699, 0.1358)$ using F_{ML} . $F_c = 0.0019$ and $\chi^2_{1,0.1} = 2.71$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

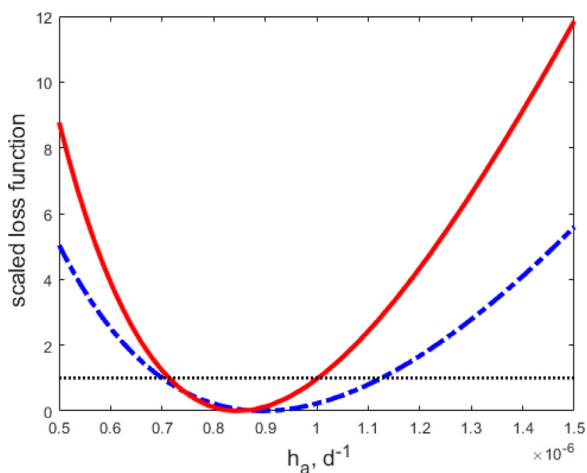


Fig. 6. Same as in Fig. 5 but with the Gompertz stress coefficient s_G fixed at the true value and $F_c = 0.002$. The 90% marginal confidence intervals for the aging acceleration \hat{h}_a are: $(0.0714 \cdot 10^{-5}, 0.1000 \cdot 10^{-5})$ using F_{sb} and $(0.0698 \cdot 10^{-5}, 0.1119 \cdot 10^{-5})$ using F_{ML} .

when the Gompertz stress coefficient s_G is fixed at the true value. This is again an implication of the strong negative correlation between the estimates of \hat{h}_a and s_G .

4. Discussion and conclusions

We used Monte Carlo studies to compare the behaviour of the loss functions, rather than real data, to be sure that we know the model as well as the true parameter values. While the Maximum Likelihood criterion is asymptotically most efficient, its performance for small samples is not always optimal. The most importance handicap of ML methods in the context of estimating DEB parameters is that the stochastic component in data needs to be modelled. Marques et al. (2019) summarizes the main components of scatter in energetic data, which shows that modelling the stochastic component is far from easy or straightforward, and likely to involve quite a few scatter parameters, while data typically hardly allow such extra parameters to be estimated.

We here find that the interval estimates of parameters based on the F_{ML} and F_{sb} profiles are very similar. This supports the view that intervals are dominated by model plasticity, rather than the detailed distribution of the scatter.

In principle it is possible to mix F_{ML} for survival data with F_{sb} for other data in a total loss function for all data sets, since each data set contributes with an additive term. Since the values of the loss functions differ considerably, also in the minimum, such mixed loss functions have to solve the problem of weighing survival relative to other data. Our present comparison for survival data shows, however, that the loss functions F_{sb} and F_{ML} perform equally well for survival data, at least in our example, so we see no need to construct such complex mixed loss functions.

Acknowledgements

SA was supported by the Norwegian Science Council (NFR 255295).

References

Am, P.. Add-my-Pet collectiononline database of DEB parameters, implied properties and referenced underlying data. http://www.bio.vu.nl/thb/deb/deblab/add_my_pet/ Last accessed: 2019/12/14.

Gompertz, B., 1825. On the nature of the function expressive of the law of mortality, and on a new method of determining the value of life contingencies. *Phil. Trans. R. Soc. Lond. B Biol. Sci.* 27, 513–585.

Kooijman, S.A.L.M., 1981. Parametric analyses of mortality rates in bioassays. *Water Res.* 15, 107–119.

Kooijman, S.A.L.M., 1983. Statistical aspects of the determination of mortality rates in bioassays. *Water Res.* 17, 749–759.

Kooijman, S.A.L.M., 2010. *Dynamic Energy Budget Theory for Metabolic Organisation*. Cambridge University Press.

Leeuwen, I.M.M.v., Kelpin, F.D.L., Kooijman, S.A.L.M., 2002. A mathematical model that accounts for the effects of caloric restriction on body weight and longevity. *BioGerontology* 3, 373–381.

Lika, K., Augustine, S., Kooijman, S.A.L.M., 2019. Body size as emergent property of metabolism. *J. Sea Res.* 143, 8–17. <https://doi.org/10.1016/j.seares.2018.04.005>.

Lika, K., Augustine, S., Kooijman, S.A.L.M., 2020. The use of augmented loss functions for estimating dynamic energy budget parameters. *Ecol. Mod.*, this volume, submitted.

Marques, G.M., Augustine, S., Lika, K., Pecquerie, L., Domingos, T., Kooijman, S., 2018. The AMP project: Comparing species on the basis of Dynamic Energy Budget parameters. *PLoS Computational Biology* 14(5): e1006100. <https://doi.org/10.1371/journal.pcbi.1006100>

Marques, G.M., Lika, K., Augustine, S., Pecquerie, L., Kooijman, S.A.L.M., 2019. Fitting multiple models to multiple data sets. *J. Sea Res.* 143, 48–56. <https://doi.org/10.1016/j.seares.2018.07.004>.

McCullagh, P., Nelder, J.A., 1983. *Generalized Linear Models*. Monographs on Statistics and Applied Probability Chapman & Hall, London.

Morais, J.V., Custdio, A.L., Marques, G.M., 2018. Calibration of parameters in dynamic energy budget models using direct-search methods. *J. Math. Biol.* 17. <https://doi.org/10.1007/s00285-018-1315-x>.

Stavrakidis-Zachou, O., Papandroulakis, N., Lika, K., 2019. A deb model for european sea bass (*Dicentrarchus labrax*): parameterisation and application in aquaculture. *J. Sea Res.* 143, 262–271.

Weibull, W., 1951. A statistical distribution of wide applicability. *J. Appl. Mech.* 18, 293–297.