Imperal College London

An Information-Theoretic Approach to Model Selection

Department of Life Sciences

Hovig Artinian

MSc CMEE

Word count: 1653

Contents

1 Introduction **2**

2 Materials & Methods **3**

2.1 Data Preparation 3

2.2 Data Analysis 4

2.2.1 Model Fitting 4

2.2.2 Model selection 4

2.3 Computing Tools 6

3 Results & Discussion **7**

3.1 Model Fitting 7

3.2 Model Selection 9

3.2.1 Case 1 9

3.2.2 Case 2 9

3.2.3 Case 3 11

4 Conclusion & Future Work **14**

5 Acknowledgements **14**

1 Introduction

Modelling nature has been the general interest in the field of ecology for over a century [1]. A crucial part of this process involves model selection. The basic approach is the null hypothesis testing, where biological inferences are made based on whether or not a suggested hypothesis is rejected [2]. This would be based on arbitrary criteria such as p-values, confidence intervals, and t-tests, that are set by statisticians as general rules of thumb. Amidst this mayhem, however, some scientists have started to shift their workflow towards more robust approaches that rely on advanced mathematical theories.

In this project, the information-theoretic approach is used for model selection. This methodology is based on information theory [3].

Preceding the model selection phase, however, is finding parameter estimates of models by fitting them to data. 7 models are used in this study: linear, quadratic, cubic, logistic [4], gompertz [5], baranyi [6], and buchanan [7]. The first three are phenomenological, which means their parameters have no biological significance. For the mechanistic models, however, 4 parameters are involved and are described in Table 1 below.

Parameter	Description	Models included
N_{max}	carrying capacity	logistic, gompertz, baranyi, buchanan
N_0	initial abundance value	logistic, gompertz, baranyi, buchanan
r_{max}	growth rate	logistic, gompertz, baranyi, buchanan
t_{lag}	time taken for lag phase	gompertz, baranyi, buchanan

Table 1: Parameters involved in the mechanistic models used in this study with their corresponding descriptions.

The Non-Linear Least Squares (NLLS) technique was performed for model fitting, which allows models to detect non-linear patterns found in data. Once parameter estimates are found, model selection is performed by calculating several information-theoretic criteria such as AIC, AIC_c, BIC, R^2, AIC differences, likelihood of models, Akaike weights, and evidence ratios.

The main objective of this project is to show how sometimes finding a "best" model is not always ideal; rather, a multimodel inference approach would be optimal instead.

2 Materials & Methods

2.1 Data Preparation

The starting dataset consisted of 4387 samples. No missing abundance values were detected. However, negative values were present. The smallest value was largely negative and, hence, removed. To deal with the rest, while still minimizing the amount of data points lost, the smallest value was added to the whole data, and then removed (to avoid having zero as an abundance value). The end result was a dataset with 4385 values. Next, each species/temperature/medium/citation/replicate was grouped together, resulting in 305 unique IDs. Finally, the new dataset was saved to be used for data analysis.

2.2 Data Analysis

2.2.1 Model Fitting

Non-linear least squares (NLLS) fitting was used to fit all 7 models to each unique group in the new dataset.

To work with this method, starting parameter values must first be provided. The better the starting values, the more precise the estimated parameter values will be.

For phenomenological models, finding the starting values was straightforward (they were set to 1). In the case of mechanistic models, on the other hand, more computation was needed. The starting values of N_{max} and N_0 were set to be the highest and lowest abundance values in the dataset, respectively. That of r_{max} was less direct. A straight-line was fit to the first 50% of the dataset, and its slope was assigned as the starting value of r_{max}. Lastly, the intersection point between the fitted tangent line and the horizontal line at y = N_0 was set to be the starting value of t_{lag}.

Next, the actual fitting was performed, where residuals for each models to be fit were provided using the newly found starting values. For each model, if the fit converged, the estimated parameters were saved in a variable; otherwise, the estimated parameter values were set to 0.

2.2.2 Model selection

For model selection, first, Akaike's Information Criterion (AIC) [8] was calculated for each model. The AIC value gives the quality of each model relative to the other models in the set used for fitting the data. Hence, the model with the lowest AIC score is preferred. For models where the sample size (n) to number of parameters (K) ratio was less than 40 (arbitrary suggestion), AIC_c was calculated instead. It is known to be the second-order variant of AIC derived by Sugiura in 1978 [9]. The difference is in the bias-correction term added to the AIC value of the model [10]. This makes sure that when sample size and number of parameters are close, the model gets heavily penalized.

Next, AIC[1] differences (Δ_i) were calculated to get a better idea of the empirical support provided by each model in the set; thus, obtaining a relative ranking of all the models in the set. The formula is given as follows:

$$\Delta_i = AIC_i - AIC_{min} \tag{1}$$

where AIC_i is the AIC value of the ith model and AIC_{min} is the AIC value of the "best" model in the set. The larger the difference, the less likely it is for that model to be the best. This likelihood was quantified and calculated by the following formula:

$$\mathcal{L}(g_i|x) \propto e^{-\frac{1}{2}\Delta_i} \tag{2}$$

Contrary to Δ_i, the larger the relative likelihood of a model, the better chances it has to be the preferred model. For better interpretation, the likelihood of the models were normalized, adding to 1. This normalized

[1]To avoid repetition, the term "AIC" will be used to mean both AIC and AIC_c. There is, however, a distinction between them, as explained, and the reader should bear that in mind.

form is known as the Akaike weight, w_i of a model:

$$w_i = \frac{e^{-\frac{1}{2}\Delta_i}}{\sum\limits_{r=1}^{R} e^{-\frac{1}{2}\Delta_r}} \tag{3}$$

where R is the number of models.

Lastly, evidence ratios were calculated to show how reliable it is to believe that the best model is actually a good fit for the data [10]. To put it in another way, evidence ratios reveal whether or not the best model in a set of models works best alone, or within a group consisting of one or more of the other models in the set. It can be found by simply calculating the ratio of Akaike weights:

$$Evidence\ ratio = \frac{w_1}{w_j} \tag{4}$$

where w_1 is the Akaike weight of the model with the lowest AIC value, and w_j is the Akaike weight of the jth model.

2.3 Computing Tools

Several programming languages were used to create the different aspects of this project.

R [11] was used for: (1) data exploration/preparation - playing around with data in R is fast and intuitive, (2) plotting - really good packages that produce nice-looking plots, (3) finding information-theoretic criteria - since calculating statistical measures is fast and straighforward in R. Packages used: *dplyr* [12] for data manipulation and *ggplot2* [13] for plotting. **Python** [14] was used to perform heavy computation, especially model fitting since the package for NLLS fitting in Python is much more robust than the one in R. Packages used: *Pandas* [15] for using dataframes, *NumPy* [16] for scientific and numeric computing, and *LMFIT* [17] for NLLS fitting. **LaTeX** was used for typesetting. **Bash** was used to glue the R and Python scripts together. **Git** was used for version control of all codes/scripts/workflow.

3 Results & Discussion

3.1 Model Fitting

Model fitting convergence success rate was 100% for 5 out of the 7 models used for NLLS fitting. The baranyi model converged 80.98% of the time when fitted to all 305 groups, while the the logistic model did not converge with over half the total number of groups in the dataset (44.92%). We would generally expect to see all phenomenological models, as well as the buchanan model, to converge because of their low level of complexity.

Figure 1 below gives a general idea of what some of the data look like and how the models were able to be overlaid.

However, visualizing the plots is not a reliable way of determining good fits. Looking at the model selection criteria can be more helpful in this case, which is shown in the next section.

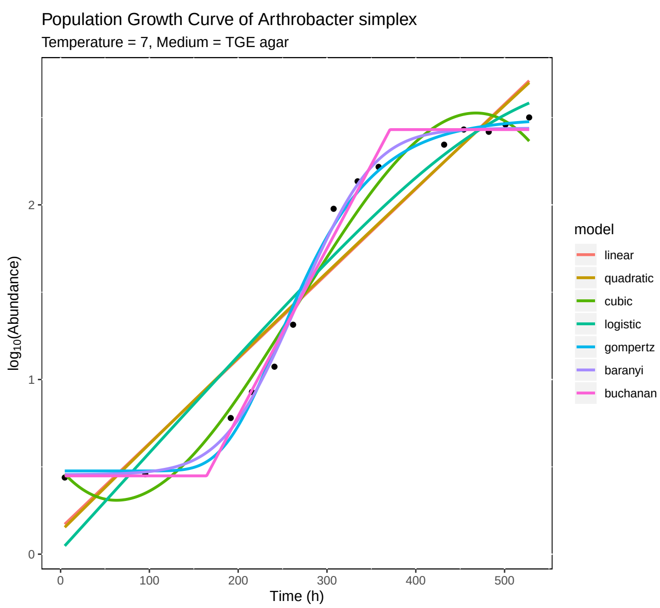
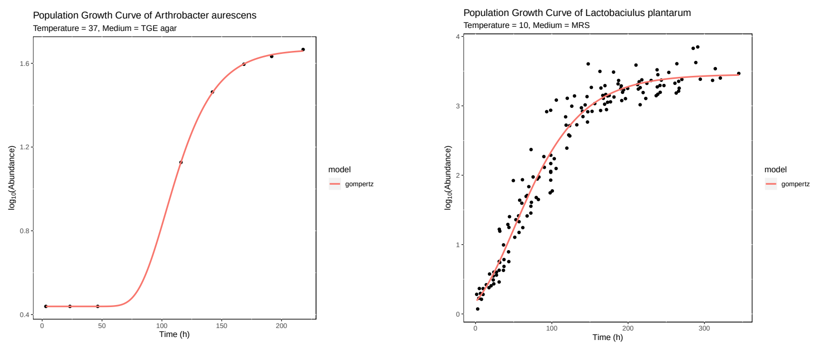


Figure 1: A general example of a bacterial population growth curve with all 7 models overlaid.



(a) (b)

Figure 2: Population growth curve with 1 best model (gompertz in both cases)

3.2 Model Selection

3.2.1 Case 1

In Figure 2, two plots are shown where the best model was deemed to be enough to predict the data. Looking at the model selection criteria in the dataset of both Figure 2(a) and 2(b), the gompertz model had the lowest AIC value, making it the best fit model relative to the others in the set. The AIC differences are all > 10, meaning that they don't have any real effect in explaining the data in the presence of the gompertz model. This is confirmed by the likelihoods, akaike weights, and evidence ratios of all other models compared to gompertz.

3.2.2 Case 2

In Figure 3, the baranyi model outperforms the rest of them based on having the minimum AIC value. However, looking at the AIC differences,

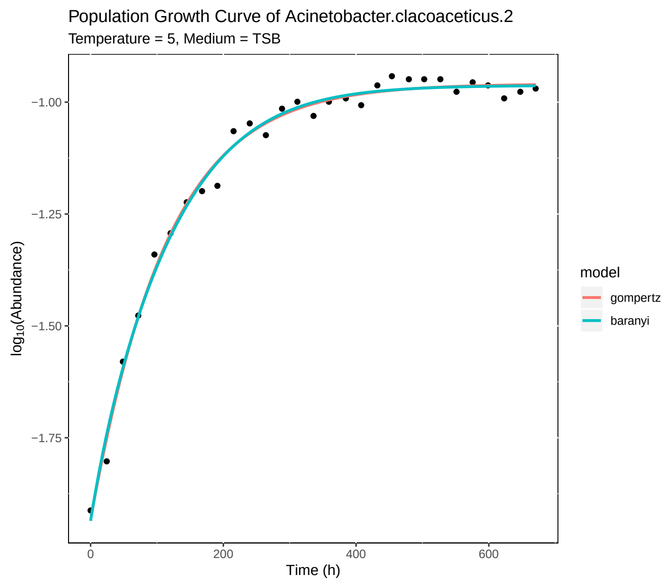


Figure 3: Population growth curve with 2 best models (in order of decreasing likelihood: baranyi, gompertz)

the gompertz model is very likely as well to be the best model to predict the data. The rest, however, have large AIC differences; therefore, they can be neglected. The evidence ratio of the gompertz model to the baranyi model is found to be 1.2398. That tells us there is a 1.2398/(1.2398+1) x 100 = 55.35% chance that the baranyi model is the best fit for the data. That surely is not good enough; the baranyi model alone will not be a good enough predictor compared to when it is considered together with the gompertz model. Therefore, in such cases, both models should be considered to make biological inferences.

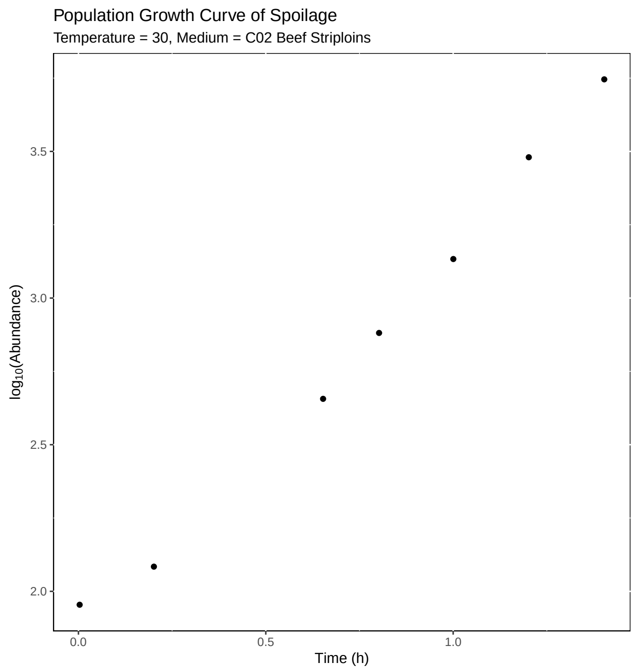


Figure 4: Population growth curve with 5 best models (in order of decreasing likelihood: logistic, gompertz, buchanan, baranyi, cubic)

3.2.3 Case 3

In Figure 4, the plot may not be visually appealing, but it conveys the true objective of this project. The cubic model is considered to be the best one based on its relative AIC value. However, looking at the AIC differences, all, but one (logistic model), models are almost as likely to be the best model. The evidence ratios suggest there are 46.71%, 43.73%, 40.37%, 33.78%, 28.82% chances for the quadratic, gompertz, baranyi, linear, or buchanan models to be the best fit for the data, respectively.

Although Akaike's Information Criterion is recognized as a major measure for selecting models, it has one major drawback: The AIC values lack intuitivity (easy to use and understand) despite higher values meaning less goodness-of-fit. For this purpose, Akaike weights come to hand for calculating the weights in a regime of several models. Additional measures can be derived, such as (AIC) and relative likelihoods that demonstrate the probability of one model being in favor over the other.

AIC:

good thing: it accounts for overfitting - it has a penalty that increases as more parameters are added. Adding parameters will almost always improve the goodness of the fit.

An individual AIC value, by itself, is not interpretable due to the unknown constant (interval scale). AIC is only comparative, relative to other AIC values in the model set; thus such differences Δ_i are very important and useful.

It is important to note here that AIC values do not give any information about the goodness-of-fit of a model to the data. Rather, they show how each model performs relative to the other ones in the set. Hence, a single AIC value of a model has no meaning.

BOOK:

Ambivalence:

The inability to ferret out a single best model is not a defect of AIC or any other selection criterion. Rather, it is an indication that the data are simply inadequate to reach such a strong inference. That is, the data are ambivalent concerning some effect or parametrization or structure.

In such cases, all the models in the set can be used to make robust inferences: multimodel inference.

The AIC differences (Δ_i) and Akaike weights (w_i) are important in ranking and scaling the hypotheses, represented by models. The evidence ratios (e.g., w_i/w_j) help sharpen the evidence for or against the various alternative hypotheses. All of these values are easy to compute and simple to understand and interpret.

The principle of parsimony provides a philosophical basis for model selection, K-L information provides an objective target based on deep theory, and AIC, AIC_c, $QAIC_c$, and TIC provide estimators of relative, expected K-L information. Objective model selection is rigorously based on these principles. These methods are applicable across a very wide range of scientific hypotheses and statistical models. We recommend presentation of $\log(\mathcal{L}(\hat{\theta}))$, K, the appropriate information criterion (AIC, AIC_c, $QAIC_c$, or TIC), Δ_i, and w_i for various models in research papers to provide full information concerning the evidence for each of the models.

Do not mix null hypothesis testing with information-theoretic criteria:

Some authors state that the best model (say g_1) is *significantly* better than another model (say g_6 based on a Δ value of 4-7. Alternatively, sometimes one sees that model g_6 is rejected relative to the best model. These statements are poor and misleading. It seems best not to associate the words significant or rejected with results under an information-theoretic paradigm.

Questions concerning the strength of evidence for the models in the set are best addressed using the evidence ratio (Section 2.10), as well as an analysis of residuals, adjusted R2, and other model diagnostics or descriptive statistics.

4 Conclusion & Future Work

studying the death phase

5 Acknowledgements

I would like to extend my gratitude to my fellow colleagues Pablo Lechon and Sam Turner for engaging in insightful discussions with me about the miniproject, as well as pulling an all-nighter the night before the submission deadline.

References

[1] Sharon E. Kingsland. *Modeling Nature: Episodes in the History of Population Ecology.* University of Chicago Press, 1995.

[2] Jerald B. Johnson and Kristian S. Omland. Model selection in ecology and evolution, 2004.

[3] S. Guiasu. *Information theory with applications.* McGraw-Hill, New York, NY, 1977.

[4] R. Pearl and L. J. Reed. On the Rate of Growth of the Population of the United States since 1790 and Its Mathematical Representation. *Proceedings of the National Academy of Sciences*, 1920.

[5] M. H. Zwietering, I. Jongenburger, F. M. Rombouts, and K. Van't Riet. Modeling of the bacterial growth curve. *Applied and Environmental Microbiology*, 1990.

[6] József Baranyi and Terry A. Roberts. A dynamic approach to predicting bacterial growth in food. *International Journal of Food Microbiology*, 1994.

[7] R. L. Buchanan, R. C. Whiting, and W. C. Damert. When is simple good enough: A comparison of the Gompertz, Baranyi, and three-phase linear models for fitting bacterial growth curves. *Food Microbiology*, 1997.

[8] H Akaike. Information theory and an extension of the maximum likelihood principle. Proceedings of the 2nd international symposium on information theory. *Second International Symposium on Information Theory*, 1973.

[9] Nariaki Sugiura. Further Analysis of the Data by Anaike' S Information Criterion and the Finite Corrections. *Communications in Statistics - Theory and Methods*, 1978.

[10] D. R. Burnham, K. P. and Anderson. Model Selection and Inference: a Practical Informationtheoretic Approach. *Model selection and multimodel inference, 2nd ed. Springer, New York*, 2002.

[11] R R Development Core Team. *R: A Language and Environment for Statistical Computing*. 2011.

[12] Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *Package 'dplyr'*. 2018.

[13] Hadley Wickham. *ggplot2*. 2009.

[14] Python Software Foundation. Python Language Reference.

[15] Wes McKinney and PyData Development Team. Pandas - Powerful Python Data Analysis Toolkit. *Pandas - Powerful Python Data Analysis Toolkit*, 2015.

[16] © Copyright 2017 NumPy developers. NumPy NumPy, 2017.

[17] Matthew Newville, Antonino Ingargiola, Till Stensitzki, and Daniel B. Allen. LMFIT: Non-Linear Least-Square Minimization and Curve-Fitting for Python. *Zenodo*, 2014.