Imperal College London

An Information-Theoretic Approach to Model Selection

Department of Life Sciences

Hovig Artinian

MSc CMEE

March 6, 2020

1 Introduction

Modelling nature has been the general interest in the field of ecology for over a century [1]. The study of bacterial population growth has been a major interest in various fields, mainly in food microbiology, for many decades. The main part of this area of research involves fitting models to bacterial growth curves generated when population abundance is plotted as a function of time. The reason for this is to be able to capture the behavior of the curve.

The general pattern of the curve is sigmoidal and composed of three main parts: lag phase, exponential phase, and stationary phase
This technique allows models to detect non-linear patterns found in data. (Figure 1). In earlier years, phenomenological models were used to fit such curves. However, they would either fail to capture the true behaviour of the curve, or would not have any biological significance. In more recent years, scientists in the field have developed mechanistic models to mitigate these issues.

In this project, seven models were used to fit various bacterial species abundance data gathered from lab experiments around the world. Three phenomenological and four mechanistic models were used (Table 1). The main objective was to compare the performance of these models.

Model Name	Model Type	Equation
Linear	Phenomenological	eq1
Quadratic	Phenomenological	eq2
Cubic	Phenomenological	eq3

Table 1: Models used

2 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 were 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.

In the case of phenomenological models, finding the starting values was straightforward (they were set to 1). On the other hand, the starting values for mechanistic models needed more computation. The starting values of N_{max}, also known as the carrying capacity, and N_0 were set to be the highest

and lowest abundance values in the dataset, respectively. That of r_{max},the growth rate, 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, AIC, AIC_c, BIC, and R^2 values were calculated.

2.3 Computing Tools

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

R - data exploration, data preparation, plotting

packages used: dplyr, ggplot2 (reference)

Python - heavy computation (NLLS fitting)

packages used: pandas, numpy, lmfit

LaTeX - writing the report

Bash - to stitch all the scripts together

Git - save all versions of code/scripts

include packages

https://github.com/alexandervdm/gummi.

3 Results

4 Discussion

Although Akaike's Information Criterion is recognized as a major measure for selecting models, it has one major drawback: The AIC values lack intuitivity 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.

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 under-

stand 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.

5 Conclusion & Future Work

studying the death phase

References

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