# Fitting Models to Data in Ecology and Evolution

CMEE Masters

Samraat Pawar

Imperial College London

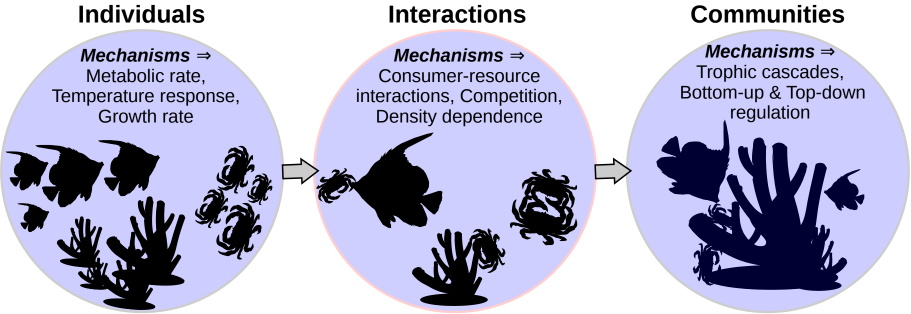
May 16, 2018

# WHAT IS MODEL SELECTION?

- Several competing hypotheses (Mathematical models) are fitted to data and compared statistical theory
- This is an advance over the traditional “null hypothesis” approach in Biology
- Necessary for developing the advancement of Biology from from an observational and axiomatic discipline to one with general theories.
- Necessary for understanding the mechanisms underlying Biological patterns and phenomena

# MECHANISTIC VS. PHENOMENOLOGICAL MODELS

- Mechanistic models aim to explain the PROCESSES underlying observed patterns
- Empirical or phenomenological models show relationships between observed data (e.g. population size as a function of temperature or rainfall), but provide no insights into why they are related



**Individuals**

***Mechanisms*** ⇒ Metabolic rate, Temperature response, Growth rate

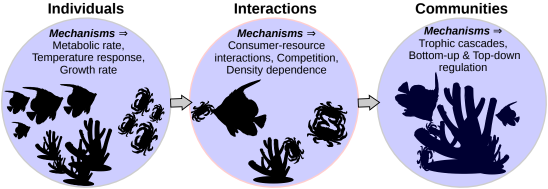
**Interactions**

***Mechanisms*** ⇒ Consumer-resource interactions, Competition, Density dependence

**Communities**

***Mechanisms*** ⇒ Trophic cascades, Bottom-up & Top-down regulation

# WHAT ARE MECHANISMS?



- Ecological studies often focus on explaining phenomena using somewhat phenomenological models.
- For example, insect invasions, outbreaks and spread (`http://www.sandyliebhold.com/pubs/science_DC1/`) — papers in your `Readings` directory.
  - Why the cycles?, Why the travelling waves? What mechanisms operate? (budmoth/parasitoid interaction? (budmoth/food quality interaction?) Are these truly mechanisms?
- Another example, disease outbreaks (Papers in your `Readings` directory)

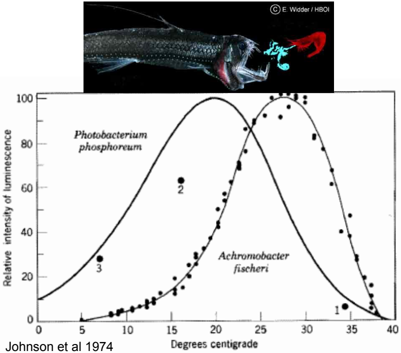
# WHAT ARE MECHANISMS?

- Somewhat subjective!
- For example, the Ricker model can be thought of as mechanistic:

$$N_{t+1} = N_t e^{r\left(1-\frac{N_t}{k}\right)} \tag{1}$$

- What is the mechanism? — Density dependence through scramble competition (Brannstrom & Sumpter 2005)
- If the Ricker model and another model with contest competition were compared with data — some would call it mechanistic modelling because one is trying to get at the underlying mechanism, scramble or contest competition
- But is this REALLY mechanistic? What are *r* and *k* really?
- Many (including yours truly!) now argue that we have not progressed far enough because the first level has been ignored!

# AN EXAMPLE OF A FUNDAMENTAL MECHANISM: METABOLISM



Johnson et al 1974

$$B = B_0 \boxed{e^{-\frac{E}{kT}}} f(T, T_{pk}, E_D)$$

$T$ = temperature (K)
$k$ = Boltzmann constant (eV $K^{-1}$)
$E$ = Activation energy (eV)
$T_{pk}$ = Temperature of peak performance
$E_D$ = Deactivation energy (eV)
(J H van't Hoff 1884, S Arrhenius 1889)

- Surely there is more to thermal responses?
  - Oxygen limitation
  - Complexity of metabolic network
  - Hormonal regulation
- *What about alternative models?*

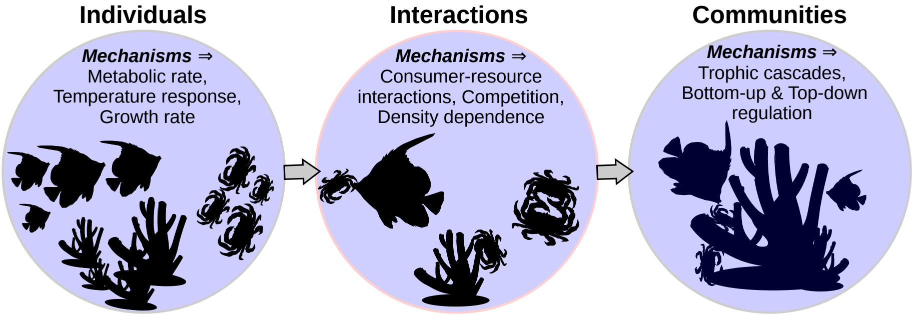
# MODELLING, AND FITTING MODELS TO DATA: WHAT'S THE BIG IDEA?

- If possible, use biological knowledge to construct models
- See if the models “agree well” with data
- Whichever model “agrees best” is most likely to have the right mechanisms
- That's the one that's best for predictions (e.g. population cycles), estimating rates (e.g. population or individual growth rates), etc
- Don't use models you already know have the wrong mechanisms just because they are popular!
- Phenomenological models often perform better than mechanistic ones

# MODELS: HOW TO BUILD THEM?

- It's an art, take practice (look at Levins' paper on the strategy of model building in biology)
- Build models one mechanism at a time — in biology, it means start at the right level of organization!
- Always consider a alternative that is more parsimonous, even if it is phenomenological (the TPC example: Sharpe-Schoolfield, or Polynomial?)!

# MODELS: HOW TO BUILD THEM?



**Individuals**

***Mechanisms*** ⇒
Metabolic rate,
Temperature response,
Growth rate

**Interactions**

***Mechanisms*** ⇒
Consumer-resource
interactions, Competition,
Density dependence

**Communities**

***Mechanisms*** ⇒
Trophic cascades,
Bottom-up & Top-down
regulation

- For example, the Boltzmann-Arrhenius model is a good first try describe and uncover mechanisms underlying individual level rates
- The next step would be to include species interactions with temperature dependence of individuals (or go in an evolutionary direction!)

# FITTING MODELS TO DATA

Two main ways to do it:

- One-step forecasting and machine learning (appropriate for discrete models) and time series data — focus in on maximizing ability to predict at the cost of mechanistic insights
- Ensemble fitting (appropriate for full time series or responses)
  - Least Squares methods
    - Linear
    - Non-linear
  - Likelihood based methods
    - Maximum Likelihood Estimation (MLE)
    - Bayesian

# ENSEMBLE FITTING

- These include MLE, Bayesian methods, and least squares optimization or fitting.
- MLE/Bayesian methods will be taught in Term II
- But you can go far with least squares methods.
- Non-linear least squares (NLLS) fitting is a particularly versatile and powerful approach, because many mechanisms in biology and inherently non-linear (Read paper by Bo).

# A quick reminder: Hypothesis testing and linear vs. and nonlinear models

# NLLS FITTING

Many of you will use NLLS. Basically, this is how it works:

1. Start with an initial value for each parameter in the model
2. Generate the curve defined by the initial values
3. Calculate the residual sum-of-squares (rss)
4. Adjust the parameters to make the curve come closer to the data points.
   - This the tricky part — you will use the Levenberg-Marquardt algorithm in the `lmfit` package in `python` or the equivalent in `R`
5. Adjust the parameters again so that the curve comes even closer to the points (RSS decreases)
6. Repeat 4–5
7. Stop simulations when the adjustments make virtually no difference to the RSS

# NLLS FITTING

Once the algorithm as converged (hopefully – but you may be surprised how well it usually works),

- Report the best-fit results, including sums of deviations of the data from the final model fit
- Then compare multiple models (e.g., Schoolfield vs. cubic)

The precise parameter values you obtain will depend in part on the initial values chosen and the stopping criteria – so different programs will not always give exactly the same results

# Comparing and selecting models

- It's all about the "Likelihood" of a model:
  the set of parameter values of the model ($\theta$) given outcomes ($x$), equals the probability of those observed outcomes given those parameter values, that is,

$$\mathcal{L}(\theta|x) = P(x|\theta)$$

- The easiest thing to do for you is to use information theory (including AIC and BIC) to compare models.
- Both AIC and BIC use the *estimated likelihoods of a model*:
  AIC: $-2\ln[\mathcal{L}(\theta|x)] + 2p$
  Small sample AIC (AICc): $-2\ln[\mathcal{L}(\theta|x)] + 2p$
  BIC (Schwartz criterion): $-2ln[\mathcal{L}(\theta|x)] + p\ln(n)$
  (where $n =$ sample size, $p$ number of free parameters)
- The lower the AIC or BIC, the better.

# Comparing and selecting models

This is how you calculate AIC and BIC (using python syntax):

- residuals = Observations - Predictions
- rss = sum(residuals ** 2)
- Then, AIC is n * log((2 * pi) / n) + n + 2 + n * log(rss) + 2 * p (*note n and p!*)
- And BIC is n + n * log(2 * pi) + n * log(rss / n) + (log(n)) * (p + 1)
- That is, $\mathcal{L}(\theta|x) = -\frac{n}{2/ln(RSS/n)}$
- For both AIC and BIC, If model **A** has AIC lower by 2-3 or more than model **B**, it's better — Differences of less than 2-3 don't really matter

Also note that:

- R$^2$ = 1 - (rss$/$tss), where tss is total sum of squares:
  tss = sum((Observations - mean(Predictions)) ** 2)
  (a useful measure of goodness of fit – you should report it)

# Comparing and selecting models

- Likelihood-Ratio test (LRT) and Adjusted $R^2$ are two other options.
- There are functions in `R` that allow you to perform model selection/simplification *for linear least squares model fitting*.

# READINGS

- Levins, R. (1966) The strategy of model building in population biology. Am. Sci. 54, 421–431.
- Johnson, J. B. & Omland, K. S. (2004) Model selection in ecology and evolution. Trends Ecol. Evol. 19, 101–108.
- Bolker, B. M. et al. (2013) Strategies for fitting nonlinear ecological models in R, AD Model Builder, and BUGS. Methods Ecol. Evol. 4, 501–512 .
- Some illustrative examples of (non-linear) model fitting to ecological/evolutionary data `https://groups.nceas.ucsb.edu/non-linear-modeling/projects`
- Additional readings at the end of Miniproject Chapter of your CMEE Notes