Inferring biological dynamics 101

- An introduction (for those new to the topic).
- A discussion of what should be in an introduction (for the experienced).

Syllabus

0. Pre-requisites.
1. Fundamental concepts.
   (a) Biological modeling.
   (b) Statistics.
   (c) Computation.
2. Research methods.
3. Case studies.
4. Individual projects.

0. Pre-requisites

These should be as minimal as possible, but no more. Nobody wants to say “you have to go back to basics before you can get started on this.”

- Statistics. An introductory course involving hypothesis testing and random variables.
- Computing. Familiarity with R is assumed, but not more advanced R topics such as S4 classes. Basic familiarity with C is necessary for many applications. Basic cluster computing (embarrasingly parallel) is essential for larger models, but prior experience is not assumed.
- Biological modeling. Some familiarity with differential equation and Markov chain models.
- Experimental biology. None!
- Persistence. Combining complex biological dynamic systems with complex nonlinear stochastic models is seldom routine!

Biological modeling

Here, a model is a quantitative connection between a scientific hypothesis (i.e., a question) and data.

The circularity of scientific progress (questions → experimentation → conclusions → questions) suggests that

question driven modeling

should be followed by

model driven questioning.

- A **qualitative model** will mean a conceptual framwork that gives useful abstract insight into a system, without providing a quantitative explanation for data. This is a common use of "model" in other biological settings [6].
- A **quantitative model** is a candidate data-generating process for the experiment. Such a model is appropriate for parameter estimation, forecasting, or evaluating potential interventions.
- These is a continuum: perhaps only some aspects of the model and/or data are to be subjected to quantitative analysis.
- A **full-information** analysis assesses the compatibility between all aspects of the data and the model.
- A **feature-based** approach compares only selected aspects.
- This distinction has gray areas. For example, data are usually processed in some way prior to analysis, even in a full-information approach.

Inference: A birds eye view

- Abstractly, a statistical model is a probability density function $f(\cdot \,|\, \theta)$ for a vector of potential observations given an unknown parameter vector, θ.

- We observe data, $\mathbf{y}$.

- If we reason directly about which values of θ would be likely to give observations similar to $\mathbf{y}$ then we are doing **frequentist** inference.

- If we augment the model with a prior $\pi(\theta)$ and compute the posterior $\pi(\theta \,|\, \mathbf{y})$ then we are doing **Bayesian** inference.

- Later, we assume an unobserved dynamic process which gives rise to $\mathbf{y}$. The statistical model is a distribution for potential observations, whether or not latent processes are specified.

- An **estimator** is a map from potential outcomes to parameter values. Evaluating this map at the data gives a **parameter estimate**.

Models vs. Methods vs. Data

- Ideally, there is a conceptual separation between choice of model and choice of inference approach. This is not always clear in practice!

- Examples confounding an estimating method with a model:

  1. "GEE (generalized estimating equation) model"

  2. Comparing a (Bayesian) hierarchical linear model to a (non-Bayesian) non-hierarchical linear model.

- **We must be careful not to confuse data with the abstractions we use to analyze them.** William James (1842–1910).

  Statistical modeling is a tool to aid understanding the data, not a substitute for understanding the data.

Information in the data

Minimum model complexity acceptable to scientists ≈ Maximum model complexity estimable from the data

- We often want to work at the limits of what the data can tell us.
- Some questions may have clear answers (in the context of given model assumptions and data) while others may not.
- Establishing the well-posed questions is part of the analysis.
- Strong model assumptions (i.e., few parameters to estimate) may lead to statistically stronger, but scientifically weaker, conclusions.
- It is possible, and sometimes advisable, to work with models for which some combinations of unknown parameters are not estimable from the data.

Don't shoot the messenger!

"The estimated parameter made no scientific sense, so we fixed it at a plausible value."

- If matching model to data gives uninterpretable results, there may be some unappreciated aspect of the model or data (unless there's a bug!).

- Imposing a canonical biological interpretation on model parameters is problematic: when we fit parameters to data, we are letting the data choose their own interpretation.

- Constraining parameters to stop the model fitting the data, or even rejecting the parameter estimation paradigm, are messenger-shooting responses to avoid the hard work of aligning the biological and statistical aspects of model fitting.

The case against fixing parameters

- If the estimated parameter agrees with your preconceptions, there is no need to fix it.

- If an estimated parameter is noxious to you, fixing it may result in other biases: remaining parameters will twist and turn to find the region of model space that you tried to fence off.

- Consider re-interpreting parameters to include unmodeled phenomena. This is scientifically unpleasant, but may be what the data ask for.

Example: measles in small & large towns [4].

- If extra-demographic stochasticity is not modeled, estimated infectious periods go down (to increase demographic noise).

- The data prefer to accomodate for unmodeled spatial aspects by increasing the estimated duration of infection, rather than via the inhomogeneity exponent, α.

The case for fixing parameters

- Parameter values which are uncontroversial and/or inconsequential can be fixed to simplify the numerical analysis and model interpretation (e.g., life expectancy of humans in an SIR epidemic model).

- Fixing parameters is logically no different from fixing other aspects of model structure (e.g., the SIR structure for epidemic models).

- Fixing parameters can complement estimation. The extent to which the data agree quantitively with a particular biological story is part of the point of the modeling exercise!

The case against prior distributions

- A fairly restrictive prior has the same disadvantages as fixing parameters, but adds neither the logical clarity nor simplicity of fixing.
- Broad priors can lead to multiple posterior modes, or nonlinear ridges. Numerical issues then force practitioners toward fixing.
- Quantitative prior information on relationships between parameters is usually unavailable, even when marginal prior information exists.
- Asserting prior independence of parameters should not be considered a scientific justification.
- There is no such thing as an objectively flat prior: a “flat” prior is skewed on a log scale.
- Conclusions can be surprisingly sensitive to the prior: in the limit as the prior flattens, the Bayes factor selects a $\text{Normal}(0, 1)$ model over a $\text{Normal}(\theta, 1)$ whatever the data $\mathbf{y}$. In this case, a flat prior is fine for computing the posterior.

The case for prior distributions

- If you have quantitative prior information, you should use it.
- “If we knew the prior, we’d all be Bayesians!” (if you philosophically dispute the existence of a prior, you could still agree with this statement.)
- Computational advances have made Bayesian inference a flexible framework applicable to many situations.

A subjective view

- Parameter fixing is done too often (maybe for reasons of computational convenience).

- Bayesian inference is done too often (maybe for reasons of computational convenience).

- To obtain new insights about the relationship between the model and the data, keep as open-minded as possible about parameter values.

- We seek to make model-based conclusions under assumptions which are (i) minimal; (ii) scientifically justified. Augmenting the model with a fairly arbitrary prior distribution, for the purpose of accessing the Bayesian inference machinery, is inadvisable on both counts.

- Any model involving unobserved random processes has computational similarities to Bayesian inference, for which parameters are unobserved random processes. Bayes' identity, $\mathbb{P}(A \mid B) = \mathbb{P}(B \mid A)\mathbb{P}(A)\big/\mathbb{P}(B)$, is useful in both cases.

Full-information vs. Feature matching

- The **likelihood function** is $f(\mathbf{y} \,|\, \theta)$ viewed as a function of θ.
- Maximizing the likelihood, and Bayesian inference based on the likelihood plus a prior, are called **full-information** or **statistically efficient** methods.
- **Feature matching** methods are based on some function of the data other than the likelihood. This includes generalized method of moments, probe matching, and Bayesian method of moments (ABC).
- Potential motives for feature matching are:
  (i) computational convenience.
  (ii) interpretability.
  (iii) using only trustworthy aspects of the data.
  (iv) diagnosing model misspecification.

Feature matching is seductive

- Offers an opportunity to use “Expert scientific insights” to simplify the analysis.
- Apart from objections about objectivity, low-dimensional summary statistics can be surprisingly uninformative for complex systems [8].
- Full-information likelihood inference has tools to detect and correct model mispecification issues which are more problematic for feature matching [5].
- Feature matching can complement full-information methods. For example, one can identify the differing messages in the data at various frequency components.
- **If one really thinks that only certain aspects of the model or data are to be taken seriously, then one should restrict attention to those features.**

Tools for likelihood-based inference

- The **log likelihood** is $\ell(\theta) = \log f(\mathbf{y} \,|\, \theta)$.

- The **maximum likelihood estimate (MLE)** is $\hat{\theta} = \arg\max \ell(\theta)$.

- Writing $\theta = (\theta_1, \dots, \theta_d)$, the **observed Fisher information** matrix is $I = -\Big[(\partial^2/\partial\theta_i\partial\theta_j)\ell(\hat{\theta})\Big]$.

- Remarkably, in many situations $\hat{\theta}$ is approximately $\mathrm{Normal}(\theta, I^{-1})$. This is **statistically efficient** since it attains the Cramér-Rao bound.

- Exact finite sample properties are, in principle, available by simulation.

- Approximate confidence intervals can be constructed using I^{-1}. Finite sample properties can be improved using **profile likelihood** methods, which also avoid differentiating the likelihood function.

Profile likelihood

- Write $\theta = (\phi, \nu)$ where ϕ is a d_ϕ-dimensional component of θ.
- The **profile log likelihood** is $\ell_p(\phi) = \max_\nu \ell(\phi, \nu)$.
- The chi-square approximation for likelihood ratio tests gives a 95% confidence interval for ϕ,

$$\{\phi : 2[\ell(\hat{\theta}) - \ell_p(\phi)] < C\},$$

where C is the 0.95 quantile of the chi-square distribution on d_ϕ degrees of freedom.

- The cut-off, C, is has asymptotic justification but good finite sample properties. It could be refined by a simulation experiment.
- The **sliced log likelihood** is $\ell_s(\phi) = \ell(\phi, \hat{\nu})$ where $\hat{\theta} = (\hat{\phi}, \hat{\nu})$. Computing $\ell_s(\phi)$ is easy, and it has uses, but it must not be confused with $\ell_p(\phi)$.

Factorizing the likelihood

- Write $\mathbf{y} = (y_1, \ldots, y_N)$. The joint density can be factored in terms of one-step prediction densities, $f(\mathbf{y} \,|\, \theta) = \prod_{n=1}^{N} f(y_n \,|\, y_1, \ldots, y_{n-1}, \theta)$.
- Many other factorizations exist. Likelihood is not synonymous with one-step prediction!

Interpreting units of log likelihood

- $f(\mathbf{y} \,|\, \theta)$ has dimension $(\text{units of } \mathbf{y})^{-1}$. Ratios, or differences of logs, are dimensionless.
- $[f(y_n \,|\, y_1, \ldots, y_{n-1}, \theta)]^{-1}$ is the width of a (uniform) one-step prediction window for y_n.
- The log likelihood from simple statistical models (linear regression, ARMA, iid Normal, etc) gives a benchmark of predictability.
- **A flag is raised if a mechanistic model has much lower likelihood than benchmarks.**
- “Much lower” means $\gg 1$ log unit. Chance variation in likelihood ratios is ≈ 1 log unit.

Likelihood-based model selection

- **Likelihood ratio test (LRT)**. Let $\Theta_0 \subset \Theta_1$ be two nested subsets of parameter space, with dimensions $d_0 < d_1$. If the true parameter is in Θ_0 then, under standard conditions, $2\big[\max_{\Theta_1} \ell(\theta) - \max_{\Theta_0} \ell(\theta)\big] \approx \chi^2_{d_1 - d_0}$.

- **Akaike's information criterion (AIC)**. Minimizing $AIC = -2 \max \ell(\theta) + 2d$ seeks to minimize prediction error for the fitted model.

- AIC is not a formal statistical test, but is applicable for non-nested models.

- Non-standard nesting is common [7]. For example, let's add a new compartment to a dynamic model with individuals entering at rate λ and leaving at rate μ. When $\lambda = 0$, note that μ becomes undefined. The chi-square LRT is typically conservative in such situations [1].

Comparing transformations of the data

- Likelihoods can be compared between different models for the same data, but not between models for different data (or between models for different subsets of the data).
- Care is required when comparing likelihoods between a model for the original data and a model for a transformation of the data.
- Likelihoods for transformed data can be ported back to the original scale using the Jacobian.

Example: A log-SARMA benchmark

Standard software will give the log likelihood for a SARMA model fitted to the log of the data.

Check that subtracting $\sum_{n=1}^{N} \log y_n$ makes this comparable to log likelihoods fitted to the data.

Plug-and-play methodology

- An **implicit** model is for which we have an algorithm to generate realizations, without having a closed form model specification [3, 2].
- Statistical methods which can operate with implicit models are **plug-and-play** [2, 4].
- **Plug-and-play methods greatly reduce the gap between model development and inference. Simulation code for a new model can be "plugged in" to existing software.**
- In the context of dynamic systems, plug-and-play is defined via the dynamic process model. Measurement error is required to follow a convenient distribution.

Partially observed Markov process (POMP) models

- A **Markov process** is a time-indexed stochastic process for which the past and future are conditionally independent given the present.
- We allow discrete-time, continuous-time, discrete-valued, continuous-valued, vector-valued, function-valued, etc.
- If any variable that affects the future evolution of a system is modeled in the current state, then the Markov property holds tautologously.
- Delays cannot usually be modeled in a finite dimensional Markov process. In specific cases (e.g., gamma-distributed delays) this is possible.
- **Partial observations** are noisy functions of the process observed at a discrete set of times.
- Each observation is conditionally independent of past and future process values and other observations, given the current process value.

Motivations for the POMP framework

- POMP models have repeatedly been proposed (or assumed without discussion) as a general framework for modeling biological systems.
- A reasonable tradeoff between generality and tractability.
- Computationally practical algorithms exist for reconstructing unobserved variables from data (filtering and smoothing) and for evaluating the likelihood function.
- Difficulties arise for large state spaces (spatio-temporal POMPs).
- Theoretical properties of Markov processes and POMPs are well studied.

Inference methods for POMPs

Frequentist or Bayesian

Full-information or Feature-based

Plug-and-play or not

Plug-and-play

	Frequentist	Bayesian
Full-information	iterated filtering	particle MCMC
Feature-based	simulated moments	ABC

Not plug-and-play

	Frequentist	Bayesian
Full-information	EM algorithm	MCMC
Feature-based	Yule-Walker*	???

*Yule-Walker is the method of moments for ARMA, a linear Gaussian POMP.

References

[1] Anisimova, M., Bielawski, J. P., and Yang, Z. (2001). Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Molecular Biology and Evolution*, 18:1585–1592.

[2] Bretó, C., He, D., Ionides, E. L., and King, A. A. (2009). Time series analysis via mechanistic models. *Annals of Applied Statistics*, 3:319–348.

[3] Diggle, P. J. and Gratton, R. J. (1984). Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 46:193–227.

[4] He, D., Ionides, E. L., and King, A. A. (2010). Plug-and-play inference for disease dynamics: Measles in large and small towns as a case study. *Journal of the Royal Society Interface*, 7:271–283.

[5] Ionides, E. L. (2011). Discussion on “Feature matching in time series modeling” by Y. Xia and H. Tong. *Statistical Science*, 26:49–52.

[6] May, R. M. (2004). Uses and abuses of mathematics in biology. *Science*, 303(5659):790–793.

[7] Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82:605–610.

[8] Shrestha, S., King, A. A., and Rohani, P. (2011). Statistical inference for multi-pathogen systems. *PLoS Comput Biol*, 7(8):e1002135.