Science Signaling
AAAS

` Biomedical Model Fitting and Error Analysis

Kevin D. Costa, Steven H. Kleinstein and Uri Hershberg (27 September 2011)
Science Signaling **4** (192), tr9. [DOI: 10.1126/scisignal.2001983]

Science Signaling (ISSN 1937-9145) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue, NW, Washington, DC 20005.

COMPUTATIONAL BIOLOGY

Biomedical Model Fitting and Error Analysis

Kevin D. Costa,[1,*] Steven H. Kleinstein,[2,3] Uri Hershberg[4]

This Teaching Resource introduces students to curve fitting and error analysis; it is the second of two lectures on developing mathematical models of biomedical systems. The first focused on identifying, extracting, and converting required constants—such as kinetic rate constants—from experimental literature. To understand how such constants are determined from experimental data, this lecture introduces the principles and practice of fitting a mathematical model to a series of measurements. We emphasize using nonlinear models for fitting nonlinear data, avoiding problems associated with linearization schemes that can distort and misrepresent the data. To help ensure proper interpretation of model parameters estimated by inverse modeling, we describe a rigorous six-step process: (i) selecting an appropriate mathematical model; (ii) defining a "figure-of-merit" function that quantifies the error between the model and data; (iii) adjusting model parameters to get a "best fit" to the data; (iv) examining the "goodness of fit" to the data; (v) determining whether a much better fit is possible; and (vi) evaluating the accuracy of the best-fit parameter values. Implementation of the computational methods is based on MATLAB, with example programs provided that can be modified for particular applications. The problem set allows students to use these programs to develop practical experience with the inverse-modeling process in the context of determining the rates of cell proliferation and death for B lymphocytes using data from BrdU-labeling experiments.

Lecture Notes

Summary

Biomedical modeling includes two powerful mathematical approaches to aid in understanding complex biological systems: namely, forward and inverse modeling (see Slides 2 to 7). This lecture is primarily focused on the latter, providing an introduction to the concepts, techniques, and criteria used to develop, implement, and evaluate an inverse model. The combined technique of forward-inverse modeling (see Slide 8) is also discussed in the context of estimating the uncertainty in resulting inverse model parameters (*1*).

Forward modeling, which includes data simulation (see Slide 5), involves a set of mathematical equations describing a biomedical system of interest, designed to incorporate a desired degree of anatomical, physical, or biological detail (*2*). Forward models are used for generating realistic synthetic data (including prescribed noise characteristics) under precisely defined conditions. This allows candidate hypotheses to be tested in silico by predicting outcomes to experimental states not easily achieved in living systems. Forward modeling can sometimes suggest improvements in experimental design and can potentially reduce the use of laboratory animals. Forward models can have arbitrary complexity as required by the problem at hand, with model parameter values typically prescribed based on published quantities.

Inverse modeling, which involves data fitting (see Slides 6 and 7), uses parameter estimation techniques applied to mathematical equations designed to provide a "best fit" to a set of experimental measurements, so as to extract values of desired model parameters often representing specific biophysical quantities (*3*). Data-fitting techniques generally involve an iterative process of adjusting model parameter values to minimize the average difference between the model-predicted and experimental data. Evaluating the quality of an inverse model requires a combination of established mathematical techniques, as well as intuition and experience, guided by a six-step process (see Slide 9), which is presented in detail in the remainder of the lecture.

Step 1: Select an appropriate mathematical model

Polynomial, exponential, and other standard functions (also called "trend lines" in spreadsheet software) are often used when a data set appears to follow a mathematical trend but the governing relation is not understood. Physically based models, on the other hand, can be derived from underlying theoretical principles when the governing physical process is known. With physically based modeling, unlike modeling using trend lines, the resulting parameter values have a specific biophysical interpretation (Slide 10).

Step 2: Define a "figure-of-merit" function

Also called an "error function," this provides a measure of the agreement between the data and the model fit for a given set of model parameters (see Slides 11 to 13). The form of the error function can be derived from probability theory (*4*, *5*) and is often based on a weighted sum of squared residuals in which each residual measures the difference between a measured data point and the corresponding model-predicted value. The weight reflects the variability of the measurement, so that the most reliable data points have the biggest influence on the error function. The process of minimizing the squared residuals error function is often called a "least-squares" model-fitting approach.

Step 3: Adjust model parameters to get a "best fit" to the data

This step involves several nuances and is therefore treated in detail (Slides 14 to 20). A relatively simple solution exists for the values of slope and intercept that minimize the least-squares error function to provide the best fit of a straight line to a given set of data (see Slides 15 to 17). For this reason, it has historically been a common approach to "linearize" a given data set by graphing in terms of a suitable change of variables, such as the Lineweaver-Burk plot for enzymatic reactions (*6*), and then perform a linear regression (Slide 18). However, such linearization often distorts the data error structure, violates key assumptions, and affects resulting model parameter values (*3*, *7*), which may lead to incorrect conclusions.

With ready access to computers, it is preferable to fit nonlinear data using an appropriate nonlinear inverse model. Obtaining the best-fit model involves computing

[1]Department of Medicine (Cardiology), Mount Sinai School of Medicine, New York, NY 10029, USA. [2]Department of Pathology, Yale School of Medicine, New Haven, CT 06520, USA. [3]Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA. [4]School of Biomedical Engineering, Science and Health Systems, Drexel University, Philadelphia, PA 19104, USA.

*Corresponding author. E-mail, kevin.costa@mssm.edu

partial derivatives of the error function with respect to each model parameter (analytically or using numerical approximation), which must approach zero when the error function in minimized. The derivatives are then used to iteratively update an initial set of model parameters until the error function stops decreasing (see Slides 19 and 20). The popular Levenberg-Marquardt method and alternative numerical methods for implementing these nonlinear minimization procedures are described in detail elsewhere (*4*, *8*).

Step 4: Examine "goodness of fit" to the data

A correlation coefficient is often used to characterize the goodness of fit between the model and data. However, a high correlation can exist even for a model that systematically differs from the data. Therefore, it is also important to examine a plot of the residuals. A good model fit should yield residuals that are uniformly spaced along the abscissa and normally distributed around zero with no systematic trends (Slide 21).

Step 5: Determine whether a much better fit is possible

This is challenging. One difficulty with nonlinear minimization, particularly with increasing model complexity, is the potential to get stuck in a local minimum of the error function without finding the global minimum (Slide 22). Although some minimization algorithms are more robust than others (*4*), none can guarantee global convergence for an arbitrarily complex nonlinear error function. The only test is to repeat the process using a different set of initial model parameter guesses and determine whether an equivalent set of best-fit parameters is obtained (see Slides 23 and 24).

Another way to improve a model fit is to increase the number of model parameters. When doing so, a statistical *F*-test should be used to determine whether the increase in model degrees of freedom is justified by the decrease in fitting error (Slides 25 and 26).

Step 6: Evaluate accuracy of best-fit parameter values

The final step is to determine the error in the estimated model parameter values. This involves fitting the model to multiple data sets that differ only because of random variability and then examining the variation in the individual model parameters, typically expressed as a confidence interval for each parameter value (*4*, *5*). It is often impractical to do this using multiple data sets acquired from the same sample under a single set of conditions. Therefore, Monte Carlo simulation of synthetic data sets using known parameter values, but including noise representative of the actual measurement noise, can be used to estimate the error in the parameter values obtained by inverse modeling (Slides 27 and 28). The simulations can be performed by using the inverse model itself or by using more complex forward models that represent the experimental system. The bootstrap technique (Slide 29) is another method for generating synthetic data by shuffling and substituting data points from the original data set. These methods are powerful tools for the critical task of estimating the error in fitted model parameters (Slide 30). In physically based inverse models, a further assessment is performed to determine whether the estimated parameter values fall within a range that is reasonable and not physically impossible, to help ensure that the model parameters are properly interpreted in a real-world context (Slide 31).

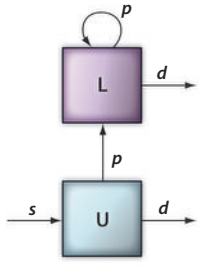
The above inverse-modeling process offers a powerful technique for maximizing the information that can be extracted from biomedical measurements and for improving the understanding of underlying biophysical processes. The problem set below was designed to reinforce key concepts by providing a real-world example from the biomedical literature, in which inverse modeling is used to determine the rates of cell proliferation and death for B lymphocytes (*9*), which are key players in the immune response to infection (*10*). The solution requires formulating a physically based model of bromodeoxyuridine (BrdU) labeling, which is used to assess cell division, fitting the model to published experimental data using both unweighted and weighted least-squares error functions, using bootstrapping to determine confidence intervals for model parameters, comparing resulting model parameters with values from the literature, and using the *F*-test to evaluate the effect of eliminating one parameter from the model. The slides conclude with a general introduction to the problem set (Slides 32 and 33) and a list of online resources and references for those who wish to explore the topic of model fitting and error analysis in greater detail (Slide 34).

Problem Set

Following infection, B cells undergo affinity maturation, a process involving cycles of proliferation, mutation, and selection, which results in a population of B cells carrying receptors with higher affinity for the pathogen than those found in naïve B cells. To understand this process, we wish to measure how the rates of proliferation and death of B cells are influenced by the affinity of the B cell receptor for antigen. One experimental means of measuring proliferation is the tracking of BrdU incorporation by B cells. BrdU is a synthetic thymidine analog that gets incorporated into a cell's DNA during the S phase of the cell cycle. Antibodies against BrdU that are conjugated to fluorescent markers can be used to label these cells, so that this evidence of cell division can be quantified using flow cytometry.

In this problem set, you will develop and apply an ordinary differential equation (ODE) inverse model of BrdU labeling to estimate proliferation rate. The experimental data are adapted from a mouse study (*9*). In this study, a mouse was immunized and BrdU dynamics were observed at 4, 8, 24, and 72 hours after the start of BrdU injection on day 13 post infection. We will assume that BrdU labeling is continuous over the course of the experiment. The data are given in Table 1.

The model to describe BrdU labeling (Fig. 1) is defined by two cell populations:



Populations:
U–Number of unlabeled B cells
L–Number of BrdU–labeled B cells

Dynamic parameters:
p–Rate of proliferation (per hour)
d–Rate of death (per hour)
s–Rate of inflow of cells from their source in the primary lymphatic organs (cells per hour)

Fig. 1. Schematic model of BrdU cell labeling experiment. The model has two cell populations: unlabeled (U) and BrdU-labeled (L). The proliferation rate (*p*), and the death rate (*d*) are the same for unlabeled as for labeled cells. There is an influx of unlabeled cells (*s*) originating from lymphatic sources. The effect of neglecting this cell source can be examined by enforcing $s = 0$ in the model (see exercise 6 for details).

Table 1. Original data set for BrdU-labeling experiment.

	Time (hours)			
	4	8	24	72
Fraction labeled	0.900	0.905	0.912	0.961
SD	0.0083	0.0426	0.0028	0.0319
Number of samples	4	4	2	4

Table 2. Revised data set with corrected mean values at 4- and 8-hour time points (standard deviations and sample sizes as in Table 1).

	Time (hours)			
	4	8	24	72
Fraction labeled	0.354	0.650	0.912	0.961

unlabeled (U) and BrdU-labeled (L). When an unlabeled cell or a labeled cell proliferates, it becomes two labeled cells. The proliferation rate (p), and the death rate (d) are assumed to be the same for unlabeled as for labeled cells. There is also an influx of unlabeled cells (s) originating from lymphatic sources.

You will be editing several MATLAB functions, but can call "run_excercise.m" to run all the parts of this lab.

1. Translate the model above into a set of ordinary differential equations. This model should be inserted into the MATLAB file "BrdUlabel.m" and replace the current simulation of asexual reproduction. To make this model work, you will also need to update the vector *Yin* and uncomment the lines following the call to *ode45*. Note that we assume that labeling is at steady state. We have illustrated this by forcing the rate of death (d) to be equal to the two source rates (i.e., $d = p + s$).

2. Fit the model to the experimental data above, using parameters from the literature as your starting point. This can be accomplished by uncommenting the lines related to exercise 2 in "run_excercise.m," and editing the initial parameter values (*xo*) in "BrdUfit.m."

3. We will next modify the fit by accounting for the varying experimental error at different time points. First, uncomment the lines related to exercise 3 in "run_excercise.m." Next, create a copy of the file "BrdUfit.m" and call it "BrdUfit_weighted.m." In this new file, change the error calculation (e) so that it now fits the data using a weighted least-squares error function. The standard deviations for each point are given in Table 1 and will be passed to the new function in the second column of the variable *fl*.

Consider your parameter estimates from the optimization in step 3. How does this compare with the ranges you found in the literature?

4. We double-checked our lab notebooks and discovered a clerical error. The actual fraction of B cells labeled is as in Table 2 (SD and sample size are unchanged).

What are the new parameter estimates? Are these more in line with rates found in the literature?

5. In the MATLAB file "run_exercise.m," uncomment the lines related to exercise 5 on bootstrapping and run them for one hundred rounds ($B = 100$). Suggestion: Start with $B = 10$ while you are debugging. Calculate and plot confidence intervals for the model parameters by using the percentile method and by editing the assignments to the variables *sc* and *pc* in "run_exercise.m."

6. Does including the influx of cells (parameter s in the model) provide a statistically significant better fit to the data? To check this, you will create a version of "BrdUlabel.m" without influx of cells (the parameter s will be 0) and compare the two models (model 1 with influx as above, model 2 without influx using $s = 0$) using an F-test. To carry this out, first uncomment the lines related to exercise 6 in "run_excercise.m." Next, create a copy of the file "BrdUfit_weighted.m" and call it "BrdUfit_weighted_s0.m." In this new file, fix the value of s to be 0. Next, edit "run_excercise.m" to use the correct degrees of freedom (dF).

7. Print out the final version of your five pieces of code ("run_exercise.m," "BrdUlabel.m," "BrdUfit.m," "BrdUfit_weighted.m," and "BrdUfit_weighted_s0.m") and the output of the full running of all of the stages in "run_exercise.m."

Educational Details

Learning Resource Type: Lecture, problem set, digital presentation
Context: Graduate
Intended Users: Teacher, learner
Intended Educational Use: Learn, plan, teach
Discipline: Biochemistry, bioengineering, biophysics, cell biology, education
Keywords: mathematical modeling, simulation, computer programming, MATLAB

Technical Details

Format: PowerPoint (.ppt)
Size: 7.5 MB
Requirements: Microsoft PowerPoint
Format: MATLAB (.m)
Size: bootstrapBrdU.m (2 KB), BrdUfit.m (2 KB), BrdUlabel.m (2 KB), run_exercise.m (3 KB)
Requirements: Mathworks MATLAB

Supplementary Materials

http://stke.sciencemag.org/cgi/content/full/sigtrans;4/192/tr9/DC1

Slides. Development of Models II: Model Fitting and Error Estimation

Problem set. MATLAB code. Four MATLAB programs for analyzing BrdU labeling to determine the rates of cell proliferation and death for B lymphocytes.

Problem set answer key is available upon request. MATLAB code. Six MATLAB programs.

References and Notes

1. K. R. Lutchen, K. D. Costa, Physiological interpretations based on lumped element models fit to respiratory impedance data: Use of forward-inverse modeling. *IEEE Trans. Biomed. Eng.* **37**, 1076–1086 (1990).
2. N. A. Trayanova, B. M. Tice, Integrative computational models of cardiac arrhythmias—Simulating the structurally realistic heart. *Drug Discov. Today Dis. Models* **6**, 85–91 (2009).
3. GraphPad PRISM online guide to curve fitting (http://www.graphpad.com/manuals/prism4/regressionbook.pdf).
4. W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing* (Cambridge Univ. Press, Cambridge, ed. 3, 2007).
5. Numerical Recipes in Fortran 77 online (http://www.nrbook.com/a/bookfpdf.php).
6. H. Lineweaver, D. Burk, The determination of enzyme dissociation constants. *J. Am. Chem. Soc.* **56**, 658–666 (1934).

7. M. L. Lobemeier, Linearization plots: Time for progress in regression. *HMS Beagle* (BiomedNet, **73**, March 3, 2000).
8. MATLAB online help (http://www.mathworks.com/access/helpdesk/help/techdoc/)
9. S. M. Anderson, A. Khalil, M. Uduman, U. Hershberg, Y. Louzoun, A. M. Haberman, S. H. Kleinstein, M. J. Shlomchik, Taking advantage: High-affinity B cells in the germinal center have lower death rates, but similar rates of division, compared to low-affinity cells. *J. Immunol.* **183**, 7314–7325 (2009).
10. M. M. Harnett, B cells spread and gather. *Science* **312**, 709–710 (2006).
11. **Funding**: The development of this course was supported by a Systems Biology Center grant (P50 GM071558). This work was supported in part by National Heart, Lung, and Blood Institute (NHLBI), National Institutes of Health (NIH), Program of Excellence in Nanotechnology (PEN) Award, Contract HHSN268201000045C (K.D.C.), National Institute of Allergy and Infectious Diseases (NIAID), NIH, contract HHSN2662000500021C (U.H. and S.H.K.) and NIAID, NIH, contract HHSN272201000054C (S.H.K.).

10.1126/scisignal.2001983

Citation: K. D. Costa, S. H. Kleinstein, U. Hershberg, Biomedical model fitting and error analysis. *Sci. Signal.* **4**, tr9 (2011).