Preamble

An ocean data assimilation system in miniature

The pages of this book are filled with the mathematics of oceanic and atmospheric circulation models, observing systems and variational calculus. It would only be natural to ask: What is going on here, and is it really new? The answers are "regression" and hence "no": almost every issue of any marine biology journal contains a variational ocean data assimilation system in miniature.

P.1 Linear regression in marine biology

The article "Repression of fecundity in the neritic copepod *Acartia clausi* exposed to the toxic dinoflagellate *Alexandrium lusitanicum*: relationship between feeding and egg production", by Jörg Dutz, appeared in *Marine Ecology Progress Series* in 1998. Dinoflagellates are a species of phytoplankton, or small plant-like creatures. The genus *Alexandrium* (`www.units.it/~abiolab/set_previous.htm`, click on 'Toxic microalgae') produces toxins which rise through the food web to produce paralytic shellfish poisoning in a variety of hydrographical regions, ranging from temperate to tropical. Zooplankton, or small animal-like creatures (`www.ios.bc.ca/ios/plankton/ios_tour/zoop_lab/copepod.htm`), graze on these dinoflagellates. The effect of the toxins on the grazers naturally arises. Dutz (1998) fed toxin-bearing *Alexandrium lusitanicum* and toxin-free *Rhodomonas baltica* (`bioloc.coas.oregonstate.edu/baltica.jpg`) to females of the copepod *Acartia clausi* in controlled amounts, and measured the fecundity or gross growth

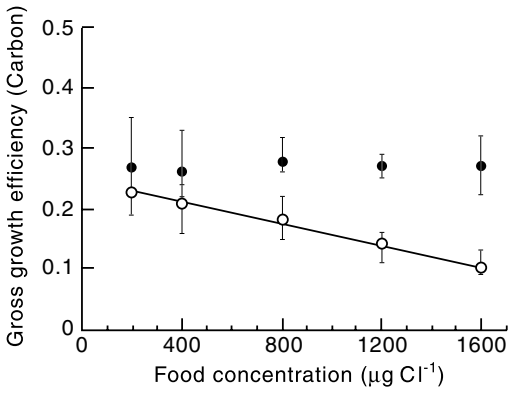

Figure P.1.1 Gross growth efficiency of *Acartia clausi* versus food supply. Solid circles: nontoxic *Rhodomonas baltica*; open circles: toxic *Alexandrium lusitanicum* (after Dutz, 1998).

efficiency in terms of total carbon production. He found that the grazers were not killed, and they continued to lay eggs. However, their fecundity was affected: see Fig. P.1.1. Note the controlled food concentration (abscissa x) with five values: 200, 400, 800, 1200, and 1600 $\mu g\, C\, l^{-1}$. Fecundity is not influenced by the supply of nontoxic *Rhodomonas* (solid circles), but is clearly reduced as the supply of toxic *Alexandrium* (open circles) increases. The gross growth efficiencies (ordinate y) in the latter case are respectively: 0.23, 0.21, 0.18, 0.14, 0.10 (Dutz, 1998; Table 2). The error bars indicate Dutz' maximum and minimum estimates. A straight line clearly fits the *Alexandrium* data well. The regression parameters are: $a = 0.25$, $b = 9.2 \times 10^{-5}$, $r^2 = 0.997$, $F_{1,3} = 355$, $P < 0.0005$.

A brief review of linear regression is in order. The data are M ordered pairs: $(x_m,\ y_m)$, $1 \le m \le M$. The model is

$$y_m = \alpha + \beta x_m + \epsilon_m, \tag{P.1.1}$$

where α and β are unknown constants, while ϵ_m is a random variable with mean and covariance

$$E\epsilon_m = 0, \quad E(\epsilon_m\, \epsilon_n) = \sigma^2 \delta_{nm} = \begin{cases} \sigma^2, & n = m \\ 0, & n \neq m. \end{cases} \tag{P.1.2}$$

The error ϵ_m is an admission of measurement error, and of the unrepresentativeness of a linear relationship. Note that the model consists of an explicit functional form (here, a linear relationship), together with probabilistic statements (here, mean and covariance) about the error in the form. We seek an estimate (here, a regression line):

$$\hat{y} = a + bx, \tag{P.1.3}$$

where a and b are to be chosen. As an estimator, let us choose a uniformly weighted sum of squared errors:

$$WSSE = \sigma^{-2} \sum_{m=1}^{M} (y_m - a - bx_m)^2. \tag{P.1.4}$$

A value for σ may be inferred from the error bars in Fig. P.1.1. It is easily shown that $WSSE$ is minimal if a and b satisfy the normal equations:

$$\begin{pmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \overline{y} \\ \overline{xy} \end{pmatrix}, \tag{P.1.5}$$

where the overbar denotes the arithmetic mean, for example $\overline{x} = M^{-1} \sum_{m=1}^{M} x_m$. Note that (P.1.5) is independent of the uniform weight σ^{-2}. These equations are of course trivially solved for a and b. The following statements may be made about the first and second moments of the solution:

$$\begin{gathered} Ea = \alpha, \quad Eb = \beta, \\ E(a-\alpha)^2 = \frac{\overline{x^2}\sigma^2}{M(\overline{x^2} - (\overline{x})^2)}, \quad E(b-\beta)^2 = \frac{\sigma^2}{M(\overline{x^2} - (\overline{x})^2)}. \end{gathered} \tag{P.1.6}$$

Moreover, a, b and $\hat{y}_m$ are normally distributed around α, β and y_m respectively. Note that the error variances in (P.1.6) are $O(M^{-1})$. In addition to the posterior error estimates (P.1.6), there are significance test statistics such as the variance-ratio or F test:

$$F_{1,M-2} = \frac{\sum_{m=1}^{M} (y_m - \overline{y})^2}{\sum_{m=1}^{M} (y_m - \hat{y}_m)}, \tag{P.1.7}$$

where $\hat{y}_m \equiv a x_m + b$. The numerator is the total variance of the data; the denominator is the total variance of the residuals for the regression line (P.1.3). Note that (P.1.7) is independent of σ^2. The subscripts 1 and $M-2$ indicate the number of degrees of freedom in the denominator and the numerator, respectively. The value of F here is 355; accordingly the probability P of the null hypothesis ($\alpha = \beta = 0$) being true is less than 0.05%. In other words it is highly credible that grazing on *Alexandrium lusitanicum* does repress the fecundity of *Acartia clausi*.

Exercise P.1.1

An alternative test statistic is provided by the weighted denominator in (P.1.7):

$$\begin{aligned} resWSSE &= \sigma^{-2} \sum_{m=1}^{M} (y_m - \hat{y}_m)^2 \\ &\sim \chi^2_M, \quad \text{as} \quad M \to \infty. \end{aligned} \tag{P.1.8}$$

Verify that $E\chi^2_M = M$, $\mathrm{var}\chi^2_M = 2M$. Calculate (P.1.8) using Dutz' data, and draw conclusions. □

If the data had suggested it, Dutz could have considered quadratic regression:

$$\begin{gathered} y_m = \alpha + \beta x_m + \gamma x_m^2 + \epsilon_m, \\ E\epsilon_m = 0, \quad E(\epsilon_n \epsilon_m) = \sigma^2 \delta_{nm}. \end{gathered} \tag{P.1.9}$$

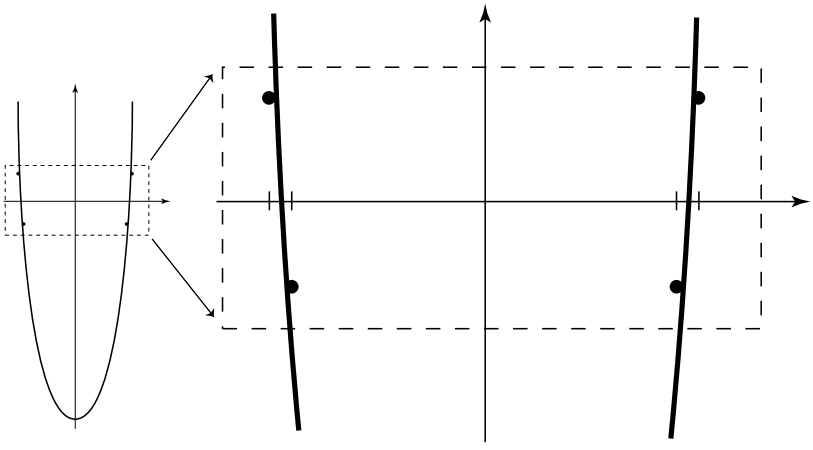

Figure P.1.2 On the left: the parabola of least-squares best fit to four data points, which are shown as solid circles. The abscissa values for the data (see the tick marks on the abscissa in the zoom on the right) are ill-chosen. As a result, the least-squares best fit is clearly ill-conditioned. The abscissa itself would be a more sensible fit to the data.

The estimate would be

$$\hat{y} = a + bx + cx^2. \tag{P.1.10}$$

The estimator would again be (P.1.4), for which the normal equations are

$$\begin{pmatrix} 1 & \overline{x} & \overline{x^2} \\ \overline{x} & \overline{x^2} & \overline{x^3} \\ \overline{x^2} & \overline{x^3} & \overline{x^4} \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \overline{y} \\ \overline{xy} \\ \overline{x^2 y} \end{pmatrix}. \tag{P.1.11}$$

Suppose for simplicity that $\overline{x} = \overline{x^3} = 0$ (these are at our disposal). Then the system (P.1.11) is ill-conditioned; that is, the solution (a, b, c) is highly sensitive to the inhomogenity on the right-hand side if $\overline{x^4}/(\overline{x^2})^2 \ll 1$. This ratio is also at our disposal. Just such a situation is sketched in Fig. P.1.2. The best fit to the four data points is a deep parabola, yet the most sensible fit would be the abscissa itself ($y = 0$). In conclusion, the stability of the estimate (P.1.10) is controlled by the choice of abscissa values x_m, $1 \le m \le M$.

P.2 Data assimilation checklist

The preceeding elementary application of linear regression in marine biology has every aspect of an "ocean data assimilation system": see the following checklist.

Data assimilation checklist

INPUTS

(i) **There is an observing system**, consisting of measurements of gross growth efficiency at selected food concentration levels.

(ii) **There are dynamics**, expressed here as (P.1.1), the explicit general solution of the differential equation

$$\frac{d^2y}{dx^2} = 0, \tag{P.2.1}$$

plus measurement errors ϵ_m, $1 \leq m \leq M$. The values α, β indicated in (P.1.1) for the regression constants a, b are the "true" values.

(iii) **There is an hypothesis** (P.1.2) about the distribution of errors ϵ_m around the true regression line.

(iv) **There is an estimator**, here the uniformly weighted sum of squared errors (P.1.4).

(v) **There is an optimization algorithm**, here the normal equations (P.1.5) which would, in the general case of N^{th}-order polynomial regression, be robustly solved using the singular value decomposition.

OUTPUTS

(vi) **There is an estimate of the state**, here the regression line (P.1.3) with values of a and b obtained from the normal equations (P.1.5).

(vii) **There are estimates of data residuals and dynamical residuals**. Here the two types of residual are indistinguishable; both are in fact given by $y_m - \hat{y}_m$.

(viii) **There are posterior error statistics**, here the means and variances (P.1.6) for $a - \alpha$ and $b - \beta$.

(ix) **There is an assessment of the array or observing system**. Here it is the conditioning of the normal matrix, and is determined by the choices of food concentrations x_m, $1 \leq m \leq M$.

(x) **There are test statistics**, here the F-variable (P.1.7) and χ^2-variable (P.1.8). These indicate the credibility of the hypothetical model, and thus the credibility of the derived posterior error statistics.

(xi) **There are indications for model improvement**. Here, however, the indication is that the linear model is so credible that a quadratic model (P.1.10) is unnecessary.

Variational assimilation of El Niño data from the tropical Pacific, into a coupled intermediate model of the ocean and atmosphere, is described in §5.5. The checklist reads as follows.

INPUTS

(i) The observations are monthly-mean and five-day mean values of Sea Surface Temperature (SST, or $T^{(1)}$), the depth of the 20° isotherm ($Z20$) and surface winds (u^a, v^a), at the TOGA–TAO moorings, from April 1994 to May 1998.

(ii) The dynamics are those of an intermediate coupled model after Zebiak and Cane (1987); the thermodynamics of the upper oceanic layer and the coupling through the wind stress are nonlinear. Otherwise the oceanic and atmospheric dynamics are those of linearized shallow-water waves.

(iii) The hypothesis consists of means and autocovariances of errors in the dynamics, in the initial conditions and in the data.
(iv) The estimator is the combined, space-integrated and time-integrated weighted squared error.
(v) The optimization algorithm is the iterated, indirect representer algorithm for solving the nonlinear Euler–Lagrange equations.

OUTPUTS

(vi) There are estimates of space-time fields of surface temperature, currents, thermocline depths and surface winds.
(vii) There are corresponding space–time fields of minimal residuals in the dynamics, initial conditions and data.
(viii) There are space–time covariances of errors in the optimal estimates of the coupled circulation.
(ix) These are assessments of the efficiency of the monthly-mean TOGA–TAO system for observing the "weak" dynamics of the coupled model, that is, observing the intermediate dynamics subject to the hypothesized error statistics.
(x) The reduced estimator is a χ^2-variable for testing the hypothesized error moments (they were found to lack credibility).
(xi) The dominance of the minimal residual in the upper-ocean thermodynamic balance indicates that it would serve no purpose to hypothesize increased variances for the dynamical errors: the low-resolution intermediate dynamics should be abandoned in favor of a fully-stratified, high-resolution, Primitive Equation model.

Variational data assimilation, or generalized inversion of dynamical models and observations, is really no more than regression analysis. The novelty lies in the mathematical and physical subtlety of realistic dynamics, in the complexity of the hypotheses about the multivariate random error fields, and in the sheer size of modern data sets. The novelty also lies in the emergence of powerful and efficient optimization algorithms, which allow us to test our models in the same way that all other scientists test theirs.

Chapter 1

Variational assimilation

Chapter 1 is a minimal course on assimilating data into models using the calculus of variations. The theory is introduced with a "toy" model in the form of a single linear partial differential equation of first order. The independent variables are a spatial coordinate, and time. The well-posedness of the mixed initial-boundary value problem or "forward model" is established, and the solution is expressed explicitly with the Green's function. The introduction of additional data renders the problem ill-posed. This difficulty is resolved by seeking a weighted least-squares best fit to all the information. The fitting criterion is a penalty functional that is quadratic in all the misfits to the various pieces of information, integrated over space and time as appropriate. The best-fit or "generalized inverse" is expressed explicitly with the representers for the penalty functional, and with the Green's function for the forward model. The behavior of the generalized inverse is examined for various limiting choices of weights. The smoothness of the inverse is seen to depend upon the nature of the weights, which will be subsequently identified as kernel inverses of error covariances. After reading Chapter 1, it is possible to carry out the first four computing exercises in Appendix A.

1.1 Forward models

1.1.1 Well-posed problems

Mechanics is captured mathematically by "well-posed problems". The mechanical laws for particles, rigid bodies and fields are with few exceptions expressed as ordinary or partial differential equations; data about the state of the mechanical system are provided

in initial conditions or boundary conditions or both. The collection of general equations and ancillary conditions constitute a "well-posed problem" if, according to Hadamard (1952; Book I) or Courant and Hilbert (1962; Ch. III, §6):

(i) a solution exists,

which

(ii) is uniquely determined by the inputs (forcing, initial conditions, boundary conditions),

and which

(iii) depends continuously upon the inputs.

Classical particles and bodies move smoothly, while classical fields vary smoothly so only differentiable functions qualify as solutions. The repeatability of classical mechanics argues for determinism. The classical perception of only finite changes in a finite time argues for continuous dependence.

Ill-posed problems fail to satisfy at least one of conditions (i)–(iii). They cannot be solved satisfactorily but can be resolved by generalized inversion, which is the subject of this chapter. Inevitably, well-posed problems are also known as "forward models": given the dynamics (the mechanical laws) and the inputs (any initial values, boundary values or sources), find the state of the system. In this first chapter, an example of a forward model is given; the uniqueness of solutions is proved, and an explicit solution is constructed using the Green's function. That is, the well-posedness of the forward model is established.

1.1.2 A "toy" example

The following "toy" example involves an unknown "ocean circulation" $u = u(x, t)$, where x, t and u are real variables. The "ocean basin" is the interval $0 \leq x \leq L$, while the time of interest is $0 \leq t \leq T$: see Fig. 1.1.1.

The "ocean dynamics" are expressed as a linear, first-order partial differential equation:

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = F \tag{1.1.1}$$

for $0 \leq x \leq L$ and $0 \leq t \leq T$, where c is a known, constant, positive phase speed. The inhomogeneity $F = F(x, t)$ is a specified forcing field; later it will become known as the prior estimate of the forcing. An initial condition is

$$u(x, 0) = I(x) \tag{1.1.2}$$

for $0 \leq x \leq L$, where I is specified. A boundary condition is

$$u(0, t) = B(t) \tag{1.1.3}$$

for $0 \leq t \leq T$, where B is specified.

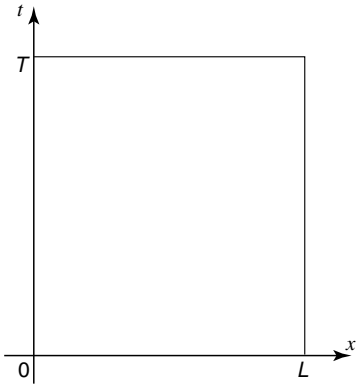

Figure 1.1.1 Toy ocean basin.

1.1.3 Uniqueness of solutions

To determine the uniqueness of solutions (Courant and Hilbert, 1962) for (1.1.1), (1.1.2) and (1.1.3), let u_1 and u_2 be two solutions for the same choices of F, I and B. Define the difference

$$v \equiv u_1 - u_2. \tag{1.1.4}$$

Then

$$\frac{\partial v}{\partial t} + c\frac{\partial v}{\partial x} = 0 \tag{1.1.5}$$

for $0 \le x \le L$ and $0 \le t \le T$;

$$v(x, 0) = 0 \tag{1.1.6}$$

for $0 \le x \le L$, and

$$v(0, t) = 0 \tag{1.1.7}$$

for $0 \le t \le T$.

Multiplying (1.1.5) by v and integrating over all x yields

$$\frac{d}{dt}\frac{1}{2}\int_0^L v^2\,dx = -c\left[\frac{1}{2}v^2\right]_{x=0}^{x=L} = -\frac{c}{2}v(L, t)^2, \tag{1.1.8}$$

using the boundary condition (1.1.7). Integrating (1.1.8) over time from 0 to t yields

$$\frac{1}{2}\int_0^L v^2(x, t)\,dx = \frac{1}{2}\int_0^L v^2(x, 0)\,dx - \frac{c}{2}\int_0^t v^2(L, s)\,ds. \tag{1.1.9}$$

The right-hand side (rhs) of (1.1.9) is nonpositive, as a consequence of the initial condition (1.1.6). Hence

$$v(x, t) = 0, \tag{1.1.10}$$

that is,

$$u_1(x,t) = u_2(x,t) \tag{1.1.11}$$

for $0 \le x \le L$ and $0 \le t \le T$. So we have established that (1.1.1), (1.1.2) and (1.1.3) have a unique solution for each choice of F, I and B.

1.1.4 Explicit solutions: Green's functions

We may construct the solution explicitly, using the Green's function (Courant and Hilbert, 1953) or fundamental solution γ for (1.1.1)–(1.1.3).

Let $\gamma = \gamma(x,t,\xi,\tau)$ satisfy

$$-\frac{\partial \gamma}{\partial t} - c\frac{\partial \gamma}{\partial x} = \delta(x-\xi)\delta(t-\tau), \tag{1.1.12}$$

where the δs are Dirac delta functions, and $0 \le \xi \le L$, $0 \le \tau \le T$. Also,

$$\gamma(L,t,\xi,\tau) = 0 \tag{1.1.13}$$

for $0 \le t \le T$, and

$$\gamma(x,T,\xi,\tau) = 0 \tag{1.1.14}$$

for $0 \le x \le L$.

Exercise 1.1.1

(a) Verify that

$$\gamma(x,t,\xi,\tau) = \delta(x-\xi-c(t-\tau))H(\tau-t) \tag{1.1.15}$$

for $0 \le x < L$, $0 \le t \le T$, where H is the Heaviside unit step function.

(b) Show that

$$u(\xi,\tau) = u_F(\xi,\tau) \equiv \int_0^T dt \int_0^L dx\, \gamma(x,t,\xi,\tau)F(x,t) + \int_0^L dx\, \gamma(x,0,\xi,\tau)I(x) + c\int_0^T dt\, \gamma(0,t,\xi,\tau)B(t). \tag{1.1.16}$$

□

Relabeling (1.1.16) yields

$$u_F(x,t) = \int_0^T d\tau \int_0^L d\xi \gamma(\xi,\tau,x,t)F(\xi,\tau) + \int_0^L d\xi \gamma(\xi,0,x,t)I(\xi) + c\int_0^T d\tau \gamma(0,\tau,x,t)B(\tau), \tag{1.1.17}$$