Short Course on Data Assimilation for Sea-Ice Modelers

P. Heimbach and C. Wunsch (MIT)

Organizers: D. Holland (NYU), D. Menemenlis (JPL),
A. Proshutinsky (WHOI), and J. Ukita (GSFC)

Sponsor: W. Abdalati (NASA Cryosphere Program)

Smith Laboratory Conference Room
Woods Hole Oceanographic Institute
Woods Hole, Massachusetts

May 10-11, 2003

Introduction

This short course is a follow-on to the Workshop on Sea Ice Data Assimilation that was held at the Naval Academy, Annapolis, MD on July 23-24, 2002. The purpose is to share the ocean-state-estimation experience of the ECCO consortium (http://www.ecco-group.org/) and to encourage a new generation of expertise in sea-ice data assimilation.

The first part of the course material is based on a revision of chapters 1, 3, and 6 of *The Ocean Circulation Inverse Problem* by Carl Wunsch (Cambridge University Press, 1996). It concerns **Discrete Inverse and State Estimation Problems**. The reader is advised that the material included herein is still a working draft.

The second part of the course material consists of a series of slides that concern **Adjoint Model Code Generation via Automatic Differentiation and its Application to Ocean / Sea Ice State Estimation**.

The course material is also available at http://ecco.jpl.nasa.gov/sea_ice/ in electronic (pdf) format.

Short Course Agenda

Saturday, May 10, 2003, 8:00am - 6:00pm

- Welcoming Statements, and Introductions
- Definition of inverse problems
- Basic machinery of discrete inverse methods:
  - Regression
  - Least squares
  - Singular vectors
  - Gauss-Markov Theorem
- The time-dependent inverse problem:
  - Green's function and representer methods
  - Kalman filter and optimal smoothers
  - Extended Kalman Filters
  - Approximate Kalman Filter
  - Monte-Carlo Methods and Ensemble Kalman Filter

Sunday, May 11, 2003, 8:00am - 6:00pm

- The time-dependent inverse problem (continued):
  - Pontryagin Principle
  - The adjoint method
- Automatic adjoint model compilers
- Some results from ECCO
- Development of sea-ice model adjoint
- Closing discussions

Short Course Particiants

1. David Barber, Faculty of Environment, University of Manitoba, Winnipeg MB R3T 2N2, Canada (204-474-9667, dbarber@Ms.UManitoba.CA).

2. Cecilia Bitz, Polar Science Center, Applied Physics Laboratory, University of Washington, Seattle WA 98105 (206-543-1339, bitz@apl.washington.edu).

3. Tom Carrieres, Canadian Ice Service, Environment Canada, LaSalle Academy, Ottawa ON K1A-0H3, Canada (613-996-5042, Tom.Carrieres@ec.gc.ca).

4. Gustavo Carrio, Dept.of Atmospheric Science, Colorado State University, Fort Collins CO 80523 (970-491-8508, carrio@atmos.colostate.edu).

5. Bin Cheng, Finnish Institute of Marine Research, Helsinki, Finland (358-9-61394427, bin@fimr.fi).

6. Valerie Duliere, Universite Catholique de Louvain, Institut d'Astronomie et de Geophysique Georges Lemaitre (ASTR), Louvain-la-Neuve, Belgium (010-473064, duliere@astr.ucl.ac.be).

7. Daniel Feltham, Centre for Polar Observation and Modelling, Department of Space and Climate Physics, University College London, UK, (020-7679-3017, Daniel.Feltham@cpom.ucl.ac.uk).

8. Charles Fowler, Colorado Center for Astrodynamics Research, University of Colorado at Boulder, CO 80309 (303-492-0975, cfowler@colorado.edu).

9. Thomas Haine, Department of Earth & Planetary Sciences, The Johns Hopkins University, Baltimore MD 21218 (410-516-7048, Thomas.Haine@jhu.edu).

10. Patrick Heimbach, Department of Earth Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge MA 02139 (617-253-5259, heimbach@mit.edu).

11. Petra Heil, TPAC/Antarctic CRC, University of Tasmania, Hobart Tasmania 7001, Australia (Petra.Heil@utas.edu.au).

12. Bill Hibler, International Arctic Research Center, University of Alaska, Fairbanks AK 99775 (907-474-7254, billh@iarc.uaf.edu).

13. David Holland, Courant Institute of Mathematical Sciences, New York University, New York, NY 10012 (212-998-3245, holland@cims.nyu.edu).

14. Greg Holloway, Institute of Ocean Sciences, Sidney BC V8L4B2 Canada (250-363-6564, hollowayg@dfo-mpo.gc.ca).

15. Elizabeth Hunke, T-3 Fluid Dynamics Group, Los Alamos National Laboratory, Los Alamos NM 87545 (505-665-9852, eclare@lanl.gov).

16. Jason Hyatt, Physical Oceanography, Woods Hole Oceanographic Institutution, Woods Hole MA 02543 (508-548 0129, jhyatt@whoi.edu).

17. Joong-Tae Kim, Texas A&M University, College Station TX 77843 (julian_kim@yahoo.com).

18. Igor Kulakov, Arctic and Antarctic Research Institute, St. Petersburg, Russia.

19. Ronald Kwok, Jet Propulsion Laboratory, California Institute of Technology, Pasadena CA 91109 (818-354-5614, ron.kwok@jpl.nasa.gov).

20. Douglas Lamb, National Ice Center, Suitland MD 20746 (301-394-3104, lambd@natice.noaa.gov).

21. Daniel Lea, Department of Earth & Planetary Sciences, The Johns Hopkins University, Baltimore MD 21218 (daniel.lea@jhuapl.edu).

22. Ron Lindsay, Polar Science Center, Applied Physics Laboratory, University of Washington, Seattle WA 98105 (lindsay@apl.washington.edu)

23. Nadja Lonnroth, Department of Oceanography, Texas A&M University, College Station TX 77843 (979-845-8216, nadjal@ocean.tamu.edu).

24. Joseph Lovick, International Arctic Research Center, University of Alaska, Fairbanks AK 99775 (907-474-6839, Joh3@anatexis.com).

25. Ted Maksym, National Ice Center, Suitland MD 20746 (Tmaksym@natice.noaa.gov).

26. Jin Meibing, School of Fisheries and Ocean Sciences, University of Alaska, Fairbanks AK 99775 (ffjm@uaf.edu).

27. Walt Meier, U.S. Naval Academy, Oceangraphy Dept., Annapolis MD 21402 (410-293-6563, meier@usna.edu).

28. Dimitris Menemenlis, Jet Propulsion Laboratory, California Institute of Technology, Pasadena CA 91109 (818-354-1656, menemenlis@jpl.nasa.gov).

29. Andrey Proshutinsky, Physical Oceanography, Woods Hole Oceanographic Institution, Woods Hole MA 02543 (508-289-2796, aproshutinsky@whoi.edu).

30. Anastasia ROMANOU, Center for Atmosphere-Ocean Science, New York University, New York, NY 10012 (romanou@cims.nyu.edu).

31. Helga SCHAFFRIN, Center for Atmosphere-Ocean Science, New York University, New York, NY 10012 (schaffri@cims.nyu.edu).

32. Abha Sood, Alfred Wegener Institute for Polar und Marine Research, Bremerhaven, Germany (0471-4831-1787, asood@awi-bremerhaven.de).

33. Donald Stark, Department of Oceanography, Naval Postgraduate School, Monterey CA 93943 (831-656-3130, drstark@nps.navy.mil).

34. Esteban TABAK, Center for Atmosphere-Ocean Science, New York University, New York University, New York, NY 10012 (tabak@cims.nyu.edu).

35. Dinh Hai Tran, Canadian Ice Service, Environment Canada, LaSalle Academy, Ottawa ON K1A-0H3, Canada (613-996-5042, Hai.Tran@ec.gc.ca).

36. Jinro Ukita, NASA Goddard Space Flight Center, Greenbelt MD 20771 (301-614-5919, jukita@fram.gsfc.nasa.gov).

37. Petteri Uotila, Courant Institute of Mathematical Sciences, New York University, (212-998-3234, uotila@cims.nyu.edu).

38. Mike Vanwoert, National Ice Center, Suitland MD 20746 (301-394-3105, mvanwoert@natice.noaa.gov).

39. Andreas Vieli, School of Geographical Sciences, University of Bristol, Bristol BS8 1SS, UK (117-928-5598, A.Vieli@bristol.ac.uk).

40. Xingren Wu, National Ice Center, Washington DC 20395 (301-394-3153, xwu@natice.noaa.gov).

41. Carl Wunsch, Department of Earth Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge MA 02139 (617-253-5937, cwunsch@mit.edu).

42. Nikolai G. Yakovlev, Institute of Numerical Mathematics, Russian Academy of Sciences, Moscow 119991, Russia (iakovlev@inm.ras.ru).

43. Jinlun Zhang, Polar Science Center, Applied Physics Laboratory, University of Washington, Seattle WA 98105 (206-543-5569, zhang@apl.washington.edu)

Discrete Inverse and State Estimation Problems.
With Fluid FlowApplications
30 April 2003 PARTIAL DRAFT

Carl Wunsch

Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology

E-mail address, C. Wunsch: cwunsch@mit.edu

Preface

This book is to a large extent the second edition of *The Ocean Circulation Inverse Problem.* It differs from the original version in a number of ways.

In teaching the basic material at MIT and elsewhere over the past 10 years, it became clear that it was of interest to many students outside of physical oceanography—the audience for whom the book had been written. The oceanographic material, instead of being a motivating factor, was in practice an obstacle to understanding for students with no oceanic background. In the revision therefore, I have tried to make the examples more generic and understandable, I hope, to anyone with a rudimentary experience with simple fluid flows.

Second, much of the oceanographic application of the methods, which still seemed novel and controversial at the time of writing, have become widespread and more sophisticated. The oceanographic applications are thus focussed less on explaining how the calculations were done, and more on summarizing what has been accomplished. Furthermore, the time-dependent problem (here called "state estimation" to distinguish it from meteorological practice) has evolved rapidly in the oceanographic community from a hypothethical methodology to one that is clearly practical and in ever-growing use.

A number of exercises, to make clear the basic concepts, is scattered throughout the book. Perhaps they will be of some pedagogical use. In the interests of keeping the book as short as possible, I have however, omitted some of the more interesting theoretical material of the original version, but which readers can find in the wider literature on control theory.

CHAPTER 1

Introduction

The most powerful insights into the behavior of the physical world are obtained when direct observations are well described by a theoretical framework that is then available for predicting new phenomena or new observations. An example is the observed behavior of radio signals and their extremely accurate description by the Maxwell equations of electromagnetic radiation. Other such examples include planetary motions as described by Newtonian mechanics, or the movement of the atmosphere and ocean as described by the equations of fluid mechanics, or the propagation of seismic waves in the earth as described by the elastic wave equations. To the degree that the theoretical framework supports, and is supported by, the observations one develops sufficient confidence to calculate similar phenomena either in previously unexplored places or to make predictions of future behavior (the position of the moon 1000 years, or the climate state of the earth, 100 years in the future).

Developing a coherent view of the physical world requires some mastery therefore, of both the structure of the necessary theoretical frameworks, and of the meaning and interpretation of real data. Conventional scientific education, at least in the physical sciences, puts a heavy emphasis on learning how to solve appropriate differential and partial differential equations (Maxwell, Schrodinger, Navier-Stokes, etc.). One learns which problems are "well-posed", how to construct solutions either exactly or approximately, and how to intepret the results. Much less emphasis is placed on the problems of understanding the implications of data, which are inevitably imperfect—containing noise of various types, often incomplete, and possibly inconsistent.

Many interesting problems arise in using observations in conjunction with theory. In particular, one is driven to conclude that there are no "well-posed" problems outside of textbooks, that stochastic elements are inevitably present and must be confronted, and that more generally, one must make inferences about the world from data that are necessarily always incomplete. The main purpose of this introductory chapter is to provide some comparatively simple examples of the type of problems one confronts in practice, and for which many interesting and useful tools exist for their solution. In an older context, this subject was called the "calculus of observations."[1]

[1] Whittaker and Robinson (1944)

1. Differential Equations

Differential equations are often used to describe natural processes. Many phenomena in the physical sciences are believed adequately explainable through some of the famous systems of equations of mathematical physics, including Schrodinger's, Maxwell's, the fluid equations of Navier-Stokes, or the equations of elastic solids. Consider the elementary problem of finding the temperature in a bar where one end, at $r = r_{A.}$ is held at constant temperature T_A, and at the other end, $r = r_B$, it is held at temperature $T_{B.}$ The only mechanism for heat transfer within the bar is by molecular diffusion, so that the governing equation is,

$$\kappa \frac{d^2 T}{dr^2} = 0 \tag{1.1}$$

subject to the boundary conditions

$$T(r_A) = T_A, \; T(r_B) = T_B. \tag{1.2}$$

Eq. (1.1) is so simple we can write the solution to (1.1) in a number of different ways. One form is,

$$T(r) = a + br \tag{1.3}$$

where a, b unknown parameters, until some additional information is provided. Here the additional information is contained in the boundary conditions (1.2), and with two parameters to be found, there is just sufficient information, and

$$T(r) = \frac{r_B T_A + r_A T_B}{r_B - r_A} + \frac{T_B - T_A}{r_B - r_A} r, \tag{1.4}$$

a straight line. Such problems, or analogues for much more complicated systems, are sometimes called "forward" or "direct" and they are "well-posed": exactly enough information is available to produce a unique solution (easily proved here, not so easily in other cases). If there are small perturbations in T_i, or r_i, then the solution changes only slightly—it is stable and differentiable. This sort of problem and its solution is what is generally taught beginning in elementary science courses.

On the other hand, the problems one encounters in actually doing science differ significantly—both in the questions being asked, and in the information available. A very large number of possibilities presents itself:

(1) One or both of the boundary values T_A, T_B is known from measurements; they are thus given as $T_A = T_A^{(c)} \pm \Delta T_A$, $T_B = T_B^{(c)} \pm \Delta T_B$, where the $\Delta T_{A,B}$ are an estimate of the possible inaccuracies in the theoretical values $T_i^{(c)}$. (Exactly what that might mean is taken up later.)

(2) One or both of the positions, $r_{A,B}$ is also the result of measurement and are of the form $r_{A,B}^{(c)} \pm \Delta r_{A,B}$.

(3) T_B is missing altogether, but is known to be positive, e.g., $T_B > 0$

(4) One of the boundary values, e.g., T_B is unknown, but an interior value $T_{int} = T_{int}^{(c)} \pm \Delta T_{int}$ is provided instead. Perhaps many interior values are known instead, but none of them perfectly.

Other possibilities exist. But even this short list raises a number of interesting, practical problems. One of the themes of this book is that almost nothing in reality is known perfectly. It is possible that ΔT_A is very small; but as long as it is not actually zero, there is no longer any possibility of finding a single unique solution.

Many variations on this model and theme arise in practice. Suppose the problem is made slightly more interesting by introducing a "source" $q(r)$, so that the temperature field is thought to satisfy the equation,

$$\frac{d^2T(r)}{dr^2} = q(r), \tag{1.5}$$

along with its boundary conditions, producing another conventional forward problem. One can convert (1.5) into a different problem by supposing that one knows $T(r)$, and seeks $q(r)$. Such a problem is even easier to solve than the conventional one: differentiate T twice. Because convention dictates that the "forward problem" involves the determination of $T(r)$ from a known $q(r)$ and boundary data, this latter problem might be labelled as an "inverse" one—simply because it contrasts with the conventional formulation.

In practice, a whole series of new problems can be raised: suppose $q(r)$ is imperfectly known. How should one proceed? If one knows $q(r)$ and $T(r)$ at a series of positions $r_i \neq r_A, r_B$, could one nonetheless deduce the boundary conditions? Could one deduce $q(r)$ if it were not known at these interior values?

$T(r)$ has been supposed to satisfy the differential equation (1.1). For many purposes, it is helpful to reduce the problem to one that is intrinsically discrete. One way to do this would be to expand the solution in a system of polynomials,

$$T(r) = \alpha_0 r^0 + \alpha_1 r^1 + ... + \alpha_n r^n, \tag{1.6}$$

and

$$q(r) = \beta_0 r^0 + \beta_1 r^1 + ... + \beta_n r^n \tag{1.7}$$

where the β_i would conventionally be known, and the problem has been reduced from the need to find a function $T(r)$ defined for all values of r, to one in which only the finite number of parameters $\alpha_i,\ 0 \leq i \leq n$ must be found.

An alternative discretization is obtained by using the coordinate r. Divide the interval $r_A = 0 \leq r \leq r_B$ into $N-1$ intervals of length Δr, so that $r_B = (N-1)\Delta r$. Then, taking a simple two-sided difference:

$$\begin{aligned}
T(2\Delta r) - 2T\left(\Delta r\right) + T(0) &= \left(\Delta r\right)^2 q\left(0\right) \\
T(3\Delta r) - 2T\left(2\Delta r\right) + T(1) &= \left(\Delta r\right)^2 q(1\Delta r) \\
&\quad . \\
&\quad . \\
T(\left(N-1\right)\Delta r - 2T\left(\left(N-2\right)\Delta r\right) + T\left((N-3)\Delta r\right) &= \left(\Delta r\right)^2 q\left(\left(N-2\right)\Delta r\right)
\end{aligned} \tag{1.8}$$

If one counts the number of equations in (1.8) it is readily found that there are $N-2$ of them, but with a total of N unknown $T\left(p\Delta r\right)$. The two missing pieces of information are provided by the two boundary conditions $T\left(0\Delta r\right) = T_0,\ T\left(\left(N-1\right)\Delta r\right) = T_{N-1}$. Thus the problem of solving the differential equation has been reduced to finding the solution of a set of ordinary linear simultaneous equations, which we will write, in the notation of Chapter 2, as

$$\mathbf{Ax} = \mathbf{b}, \tag{1.9}$$

where $\mathbf{A}$ is a square matrix, $\mathbf{x}$ is the vector of unknowns $T\left(p\Delta t\right)$, and $\mathbf{b}$ is the vector of values $\mathbf{q}\left(p\Delta t\right)$, and of boundary values. The list above, of variations, e.g., where a boundary condition is missing, or where interior values are provided instead of boundary conditions, become statements then about having too few, or possibly too many, equations for the number of unknowns. Uncertainties in the T_i or in the $q\left(p\Delta r\right)$, become statements about having to solve simultaneous equations with uncertainties in some elements. That models, even nonlinear ones, can be reduced to sets of simultaneous equations is the unifying theme of this book. One might need truly vast numbers of grid points, $p\Delta r$, or polynomial terms, and ingenuity in the formulation to obtain adequate accuracy, but as long as the number of parameters, $N < \infty$, one has achieved a great, unifying simplification.

Consider a slightly more interesting ordinary differential equation, that for the simple mass-spring oscillator,

$$m\frac{d^2\xi\left(t\right)}{dt^2} + \gamma\frac{d\xi\left(t\right)}{dt} + k_0\xi\left(t\right) = q\left(t\right), \tag{1.10}$$

where m is mass, k_0 is a spring constant, and γ a dissipation parameter. Although the equation is slightly more complicated than is (1.5), and we have relabelled the independent variable as t (to suggest time), rather than as r, there really is no fundamental difference. This differential equation can also be solved in any number of ways. As a second order equation, it is well-known that one must provide two extra conditions to have enough information to have a unique solution. Typically, there are *initial* conditions, $\xi\left(0\right), d\xi\left(0\right)/dt$—a position and velocity, but there is nothing to prevent us from assigning two end conditions, $\xi\left(0\right), \xi\left(t = t_f\right)$, or even two velocity conditions $d\xi\left(0\right)/dt, d\xi\left(t_f\right)/dt$, etc.

If we naively discretize (1.10) as we did the straightline equation, we have,

$$\xi(p\Delta t+\Delta t)= \tag{1.11}$$
$$\left(2-\frac{r\Delta t}{m}-\frac{k(\Delta t)^2}{m}\right)\xi(p\Delta t)+\left(\frac{r\Delta t}{m}-1\right)\xi(p\Delta t-\Delta t)+(\Delta t)^2\,\frac{q\left(p\Delta t\right)}{m},$$
$$2\leq p\leq N-1$$

which is another set of simultaneous equations as in (1.9) in the unknown $\xi\left(p\Delta t\right)$; an equation count again would show that there are two fewer equations than unknowns—corresponding to the two boundary or two initial conditions. In Chapter 2, several methods will be developed for solving sets of simultaneous linear equations, even when there are apparently too few or too many of them. In the present case, if one were given $\xi\left(0\right),\xi\left(1\Delta t\right),$ Eq. (1.11) could be stepped forward in time, generating $\xi\left(3\Delta t\right),\xi\left(4\Delta t\right),...,\xi\left(\left(N-1\right)\Delta t\right)$. The result would be identical to the solution of the simultaneous equations—but with far less computation.

But if one were given $\xi\left(\left(N-1\right)\Delta t\right)$ instead of $\xi\left(1\Delta t\right),$ such a simple time-stepping rule could no longer be used. One would have a similar difficulty if $q\left(j\Delta t\right)$ were missing for some j, but instead one had knowledge of $\xi\left(p\Delta t\right),$ for some p. Looked at as a set of simultaneous equations, there is no conceptual problem. There *is* a problem only if one did seek to time-step the equation forward, but without the required second condition at the starting point—there would be inadequate information to go forward. Many of the methods explored in this book are ways to solve simultaneous equations while avoiding the need for all-at-once brute force solution. Nonetheless, one is urged to always recall that most of the interesting algorithms are nonetheless nothing but clever ways of solving large sets of such equations.

2. Partial Differential Equations

Finding the solutions of linear differential equations is equivalent, when discretized, to solving sets of simultaneous linear algebraic equations. Unsurprisingly, the same is true of partial differential equations. As an example, consider a very familiar problem:

Solve

$$\nabla^2\phi=\rho, \tag{2.1}$$

for ϕ, given ρ, in the domain $\mathbf{r}\in D$, subject to the boundary conditions $\phi=\phi_0$ on the boundary ∂D, where $\mathbf{r}$ is a spatial coordinate of dimension greater than 1.

This statement is the Dirichlet problem for the Laplace-Poisson equation, whose solution is well-behaved, unique, and stable to perturbations in the boundary data, ϕ_0, and the source or forcing, ρ. Because it is the familiar boundary value problem, it is by convention again labeled a forward or direct problem. Now consider a different version of the above:

Solve (2.1) for ρ given ϕ in the domain D.

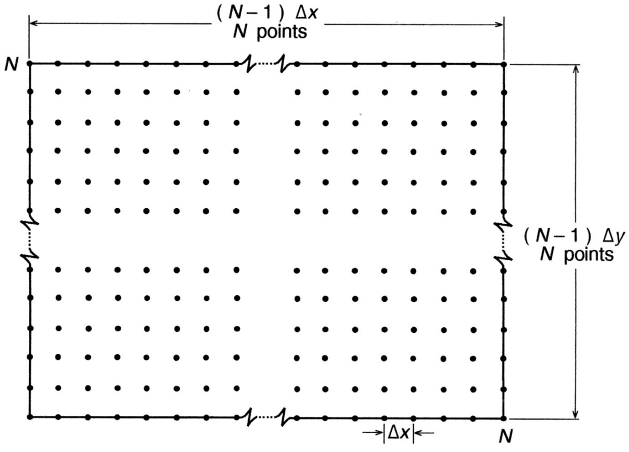


FIGURE 1. Simple square, homogeneous grid used for discretizing the Laplacian, thus reducing the partial differential equation to a set of linear simultaneous equations.

This latter problem is even easier to solve than the forward problem: merely differentiate ϕ twice to obtain the Laplacian, and ρ is obtained directly from (2.1). Because the problem as stated is inverse to the conventional forward one, it is labeled, as with the ordinary differential equation, an *inverse problem.* It is inverse to a more familiar boundary value problem in the sense that the usual unknowns ϕ have been inverted or interchanged with (some of) the usual knowns ρ. Notice that both problems, as posed, are well-behaved and produce uniquely determined answers (ruling out mathematical pathologies in any of ρ, ϕ_0, ∂D, or ϕ). Again, there are many variations possible: one could, for example, demand computation of the boundary conditions, ϕ_0, from given information about some or all of ϕ, ρ.

Write the Laplace-Poisson equation in finite difference form for two Cartesian dimensions:

$$\phi_{i+1,j} - 2\phi_{i,j} + \phi_{i-1,j} + \phi_{i,j+1} - 2\phi_{i,j} + \phi_{i,j-1} = \rho_{ij}, \quad i,\, j \in D\,. \tag{2.2}$$

To make the bookkeeping as simple as possible, suppose the domain D is the square $N \times N$ grid displayed in Figure 1, so that ∂D is the four line segments shown. There are $(N-2) \times (N-2)$ interior grid points, and Equations (2.2) are then $(N-2) \times (N-2)$ equations in N^2 of the ϕ_{ij}. If this is the forward problem with ρ_{ij} specified, there are fewer equations than unknowns. But if we append to (2.2) the set of boundary conditions:

$$\phi_{ij} = \phi_{ij}^0\,, \quad i,\, j \in \partial D\,, \tag{2.3}$$

there are precisely $4N-4$ of these conditions, and thus the combined set (2.2) plus (2.3), which we write again as (1.9),

$$\mathbf{x} = \mathrm{vec}\{\phi_{ij}\} = \begin{bmatrix} \phi_{11} \\ \phi_{12} \\ \cdot \\ \cdot \\ \phi_{NN} \end{bmatrix}, \qquad \mathbf{b} = \mathrm{vec}\left\{\rho_{ij}, \phi^0_{ij}\right\} = \begin{bmatrix} \rho_{11} \\ \rho_{12} \\ \cdot \\ \cdot \\ \rho_{N-2,N-2} \\ \phi^0_{11} \\ \cdot \\ \phi^0_{ij} \end{bmatrix},$$

a set of $M = N^2$ equations in $M = N^2$ unknowns. (The operator vec, defined formally later, is used to form a column vector out of the two-dimensional array ϕ_{ij}, etc.) The nice properties of the Dirichlet problem can be deduced from the well-behaved character of the matrix $\mathbf{A}$. Thus the forward problem corresponds directly with the solution of an ordinary set of simultaneous algebraic equations.[2] One complementary inverse problem says: "Using (1.9)compute ρ_{ij} and the boundary conditions, given ϕ_{ij}," an even simpler computation—it involves just multiplying the known $\mathbf{x}$ by the known matrix $\mathbf{A}$.

But now let us make one small change in the forward problem, changing it to the Neumann problem:

Solve

$$\nabla^2 \phi = \rho \tag{2.4}$$

for ϕ, given ρ, in the domain $\mathbf{r} \in D$ subject to the boundary conditions $\partial\phi/\partial\hat{\mathbf{m}} = \phi'_0$ on the boundary ∂D, where $\mathbf{r}$ is again the spatial coordinate and $\hat{\mathbf{m}}$ is the unit normal to the boundary.

This new problem is another classical, much analyzed forward problem. It is, however, well-known that the solution to (2.4) with these new boundary conditions is indeterminate up to an additive constant. This indeterminacy is clear in the discrete form: Equations (2.3) are now replaced by

$$\phi_{i+1,j} - \phi_{i,j} = \phi^{0\prime\prime}_{ij}, \qquad i,\, j \in \partial D' \tag{2.5}$$

etc., where $\partial D'$ represents the set of boundary indices necessary to compute the local normal derivative. There is a new combined set:

$$\mathbf{A}\mathbf{x} = \mathbf{b}_1, \;\; \mathbf{x} = \mathrm{vec}\left\{\phi_{ij}\right\}, \;\; \mathbf{b}_1 = \mathrm{vec}\left\{\rho_{ij}, \phi^{0\prime}_{ij}\right\} \tag{2.6}$$

Because only *differences* of the ϕ_{ij} are specified, there is no information concerning the mean value of $\mathbf{x}$. When we obtain some proper machinery in Chapter 2, we will be able to demonstrate that even though (2.6) appears to be M equations in M unknowns, in fact only $M-1$ of the

[2] Lanczos (1961) has a much fuller discussion of this correspondence.

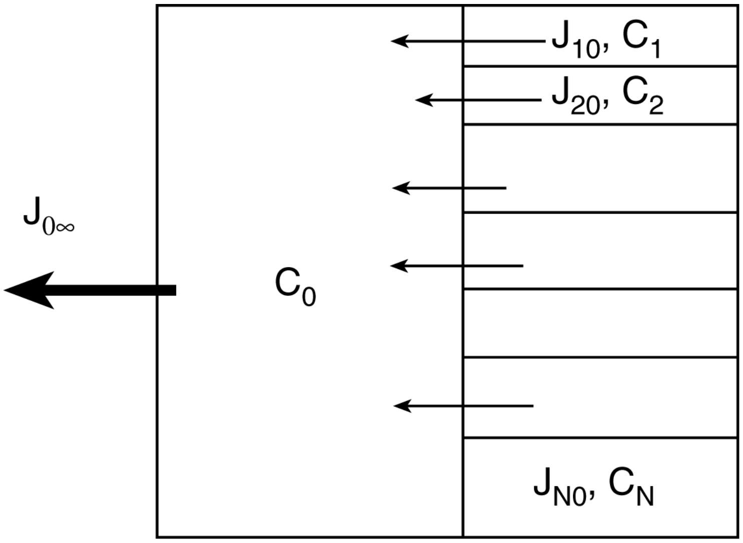


FIGURE 2. A simple reservoir problem in which there are multiple sources of flow, at rates J_{i0}, each carrying an identifiable property C_i, perhaps a chemical concentration. In the forward problem, given J_{i0}, C_i one could calculate C_0. One form of inverse problem provides C_0 and the C_i and seeks the values of J_{i0}.

equations are independent, and thus the Neumann problem is an underdetermined one. This property of the Neumann problem is well-known, and there are many ways of handling it, either in the continuous or discrete forms. In the discrete form, a simple way is to add one equation setting the value at any point to zero (or anything else). Notice however, that in all cases, the inverse problem of determining $\mathbf{b}_1$ from $\mathbf{x}$ remains simple and well-posed.

3. More Examples

A Tracer Box Model

In scientific practice, one often has observations of elements being described by the differential system or other model. Such situations vary enormously in their complexity and sophistication of both the data and the model. A useful and interesting example of a simple system, with applications in many fields is one in which there is a large reservoir (Figure 2) connected to a number of source regions which provide fluid to the reservoir. One would like to determine the rate of mass transfer from each source region to the reservoir.

We suppose that some chemical tracer or dye, C_0 is measured in the reservoir, and that the concentrations of the dye, C_i, in each source region are known. Let the unknown transfer rates be J_{i0} (transfer from source i to reservoir 0). Then we must have

$$C_1 J_{10} + C_2 J_{20} + ... + C_N J_{N0} = C_0 J_{0\infty} \tag{3.1}$$

which states that for a steady-state, the rate of transfer in, must equal the rate of transfer out (written $J_{0\infty}$). To conserve mass, one must have

$$J_{10} + J_{20} + ... + J_{N0} = J_{0\infty}. \tag{3.2}$$

This model has produced for us two equations in $N + 1$ unknowns, $[J_{i0,}J_{0\infty}]$ which evidently is insufficient information if $N > 1$. The equations have also been written as though everything were perfect. If for example, the tracer concentrations C_i were measured with finite precision and accuracy (they always are), one might try to accomodate the resulting inaccuracy as

$$C_1J_{10} + C_2J_{20} + ... + C_NJ_{N0} + n = C_0J_{0\infty} \tag{3.3}$$

where n represents the resulting error in the equation. Its introduction of course, produces one more unknown. If the reservoir were capable of some degree of storage or fluctuation in level, one might want to introduce an error term into (3.2) as well. One should also notice, that as formulated, one of the apparently infinite number of solutions to Eqs. (3.1, 3.2) includes $J_{i0} = J_{0\infty} = 0$—no flow at all. Clearly something more is required if this null solution is to be excluded.

To make the problem slightly more interesting, let us suppose that the tracer C is radioactive, and decays with a decay constant λ. We must then modify (3.1) to

$$C_1J_{10} + C_2J_{20} + ... + C_NJ_{N0} - C_0J_{0\infty} = -\lambda C_0 \tag{3.4}$$

Now the zero solution for J_{ij} is no longer possible if $C_0 > 0$, but we still have potentially many more unknowns than equations. These equations are once again in the canonical form $\mathbf{Ax} = \mathbf{b}$.

A Tomographic Problem

So-called tomographic problems occur in many fields, most notably in medicine, but also in materials testing, oceanography, meteorology and geophysics. Generically, they arise when one is faced with the problem of inferring the distribution of properties inside an area or volume based upon a series of integrals through the region. Consider Fig. 3. To be specific, we suppose we are looking at the top of the head of a patient lying supine in a so-called CAT-scanner. The two external shell sectors represent in (a) a source of x-rays and in (b) set of x-ray detectors. X-rays are emitted from the source and travel through the patient along the indicated lines where the intensity of the received beam is measured. Let the absorbtivity/unit length within the patient be a function, $c(\mathbf{r})$, where $\mathbf{r}$ is the vector position within the patient's head. Consider one source at $\mathbf{r}_s$ and a receptor at $\mathbf{r}_r$ connected by the path as indicated. Then the intensity

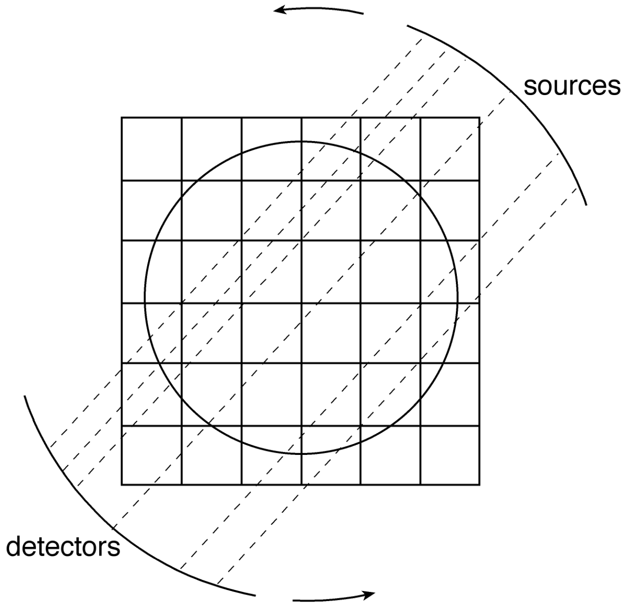


FIGURE 3. Generic tomographic problem in two dimensions. Integrals through an otherwise impenetrable solid are obtained by measuring the change in properties between the transmitted sources and receivers. Properties can be anything measurable, including travel times, intensities, group velocities etc. The tomographic problem is to reconstruct the interior from these integrals. In the particular configuration shown, the source and receiver and supposed to revolve so that a very large number of paths can be built up. It is also supposed that the division into small rectangles is an adequate representation. In principle, one can have many more integrals than the number of squares defining the unknowns.

measured at the receptor is,

$$I\left(\mathbf{r}_s, \mathbf{r}_r\right) = \int_{\mathbf{r}_s}^{\mathbf{r}_s} c\left(\mathbf{r}\left(s\right)\right) ds, \tag{3.5}$$

where s is the arc-length along the path. The basic tomographic problem is to determine $c\left(\mathbf{r}\right)$ for all $\mathbf{r}$ in the patient, from measurements of I. In the medical problem, the shell sectors rotate around the patient, and an enormous number of integrals along (almost) all possible paths are obtained. An analytical solution to this problem, as the number of paths becomes infinite, is produced by the Radon transform[3]. Given that tumors and the like have a different absorbtivity than does normal tissue, the recreated image of $c\left(\mathbf{r}\right)$ permits physicians to "see" inside the patient. In most other situations however, the number of paths tends to be much smaller than the formal number of unknowns and other solution methods must be found.

[3] Herman (1980). "Radon" is pronounced not as in radon gas, but with the 'a' sounding as in "mad."

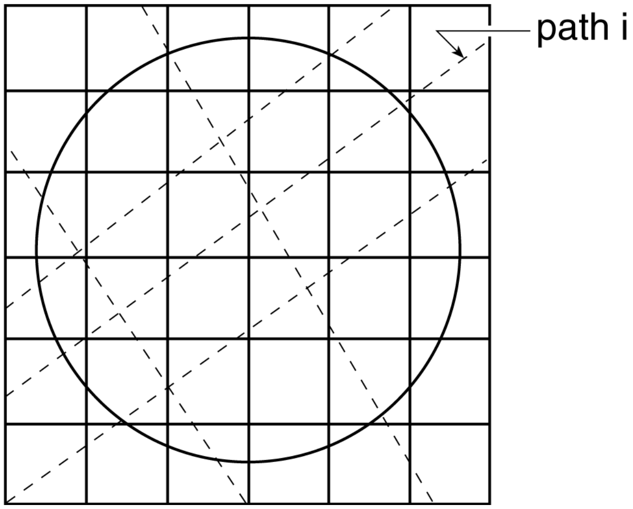


FIGURE 4. Simplified geometry for defining a tomographic problem. Some squares may have no integrals passing through them; others may be multiply-covered. Boxes outside the physical body can be handled in a number of ways, including the addition of constraints setting the corresponding $c_j = 0$.

Note first however, that we should modify Eq. (3.5) to reflect the inability of any system to produce a perfect measurement of the integral, and so more realistically we write

$$I(\mathbf{r}_s, \mathbf{r}_r) = \int_{\mathbf{r}_s}^{\mathbf{r}_s} c(\mathbf{r}(s))\, ds + n(\mathbf{r}_s, \mathbf{r}_r) \tag{3.6}$$

where n is the measurement noise.

To proceed, let us surround the patient with a bounding square (Fig. 4)–simply to produce a simple geometry, and subdivide the area into subsquares as indicated, each numbered in sequence, $1 \leq j \leq N$. These squares are supposed sufficiently small that $c(\mathbf{r})$ is effectively constant within them. Also number the paths, $1 \leq i \leq M$. Then Eq. (3.6) can be approximated with arbitrary accuracy (by letting the number of subsquares become arbitrarily large), as

$$I_i = \sum_{j=1}^{N} c_j \Delta r_{ij} + n_i \tag{3.7}$$

Here Δr_{ij} is the arc length of path i within square j (most of them will vanish for any particular path). Once again, these last equations are of the form

$$\mathbf{Ex} + \mathbf{n} = \mathbf{y}, \tag{3.8}$$

where here, $\mathbf{E} = \{\Delta r_{ij}\}$, $\mathbf{x} = [c_j]$, $\mathbf{n} = [n_i]$. Quite commonly there are many more unknown c_j than there are integrals I_i. (In the present context, there is no distinction between writing

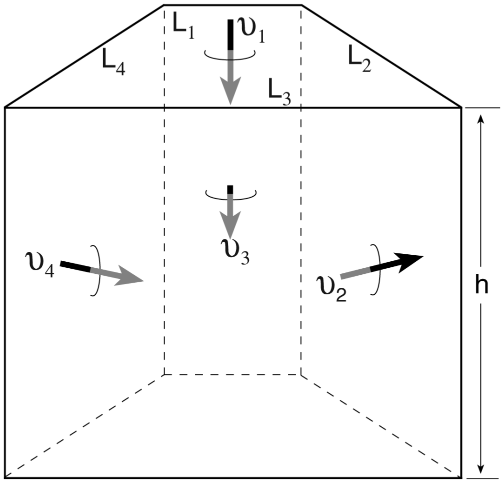


FIGURE 5. Volume of fluid bounded on four open sides across which fluid is supposed to flow. Mass is conserved, giving one relationship among the fluid transports v_i; conservation of one or more other tracers C_i leads to additional useful relationships.

matrices $\mathbf{A}, \mathbf{E}$. I will however, generally use $\mathbf{E}$ where noise elements are present, and $\mathbf{A}$ where none are intended.)

Outside of medicine measurements are not always of x-ray intensities. In seismology or oceanography, for example, c_j is commonly $1/v_j$ where v_j is the speed of sound or seismic waves within the area; I is then a travel time rather than an intensity. The methodology also works in three-dimensions, the paths need not be straight lines and there are many generalizations.[4] A probem of great practical importance is determining what one can say about the solutions to Eqs. (3.3) even where many more unknowns exist than formal pieces of information y_i.

As with all these problems, many other forms of discretization are possible. For example, we could have expanded the continuous function

$$c\left(\mathbf{r}\right)=\sum_q\sum_p a_{nm}T_n\left(r_x\right)T_m\left(r_y\right) \tag{3.9}$$

where $\mathbf{r}=(r_x,r_y)$ and the T_n are any suitable expansion functions (sines and cosines, Chebyschev polynomials, etc.). The linear equations (3.3) then represent constraints leading to the determination of the set of a_{nm}.

A Second Tracer Problem

Consider the closed volume in Fig. 5 enclosed by four boundaries as shown. There are steady flows, $v_i\left(z\right),\ 1\leq i\leq 4$ either into or out of the volume, each carrying a corresponding fluid of constant density ρ_0. z is the vertical coordinate. If the width of each boundary is l_i, the

[4] Herman (1980); Munk et al. (1995).

statement that mass is conserved within the volume is simply

$$\sum_{i=1}^{r} l_i \rho_0 \int_{-h}^{0} v_i\left(z\right) dz = 0, \tag{3.10}$$

where the convention is made that flows into the box are positive, and flows out are negative. $z = -h$ is the lower boundary of the volume and $z = 0$ is the top one. If the v_i are unknown, Eq. (3.10) represents one equation (constraint) in four unknowns,

$$\int_{-h}^{0} v_i\left(z\right) dz. \tag{3.11}$$

One possible, if boring, solution is $v_i\left(z\right) = 0$. To make the problem somewhat more interesting, we now suppose that for some (mysterious) reason, the vertical derivatives, $v_i'\left(z\right) = dv_i/dz$, are known, so that

$$v_i(z) = \int_{-z_o}^{z} v_i'\left(z\right) dz + b_i, \tag{3.12}$$

where z_0 is a convenient place to start the integration (but can be any value). b_i are integration constants ($b_i = v_i\left(z_0\right)$) which remain unknown. Constraint (3.10) becomes,

$$\sum_{i=1}^{4} l_i \rho_0 \int_{-h}^{0} \left[\int_{-z_o}^{z} v_i'\left(z'\right) dz' + b_i\right] dz = 0, \tag{3.13}$$

or,

$$\sum_{i=1}^{4} h l_i b_i = -\sum_{i=1}^{4} l_i \int_{-h}^{0} dz \int_{-z_o}^{z} v_i'\left(z'\right) dz' \tag{3.14}$$

where the right-hand side is supposed known. Eq. (3.14) is still one equation in four unknown b_i, but the zero-solution is no longer possible, assuming the right-hand side does not vanish. Eq. (3.14) is a statement that the weighted average of the b_i on the left-hand-side is known. If one seeks to obtain estimates of the b_i separately, more information is required.

Suppose that information pertains to a tracer, perhaps a red-dye, known to be conservative, and that the box concentration of red-dye, C, is known to be in a steady-state. Then conservation of C becomes,

$$\sum_{i=1}^{4} \left[h l_i \int_{-h}^{0} C_i\left(z\right) dz\right] b_i = -\sum_{i=1}^{4} l_i \int_{-h}^{0} dz \int_{-z_o}^{z} C_i\left(z'\right) v_i'\left(z'\right) dz'^{,} \tag{3.15}$$

where $C_i\left(z\right)$ is the concentration of red-dye on each boundary. Eq. (3.15) provides a second relationship for the four unknown b_i. One might try to measure another dye concentration, perhaps green dye, and write an equation for this second tracer, exactly analogous to (3.15). With enough such dyes, one might obtain more constraint equations than unknown b_i. In any case, no matter how many dyes are measured, the resulting equation set is of the form (3.3). The number of boundaries is not limited to four, but can be either fewer, or many more.[5]

Vibrating String

Consider a uniform vibrating string anchored at its ends $r_x = 0$, $r_x = L$. The free motion of the string is governed by the wave equation

$$\frac{\partial^2 \eta}{\partial r_x^2} - \frac{1}{c^2}\frac{\partial^2 \eta}{\partial t^2} = 0, \; c^2 = T/\rho, \tag{3.16}$$

where T is the tension, and ρ the density. Free modes of vibration (eigen-frequencies) are found to exist at discrete frequencies, s_q,

$$2\pi s_q = \frac{q\pi c}{L}, \; q = 1, 2, 3, ..., \tag{3.17}$$

and which is the solution to a classical forward problem. A number of interesting and useful inverse problems can be formulated. For example, given $s_q \pm \Delta s_q$, $q = 1, ..., M$, to determine L, or c. These are particularly simple problems, because there is only one parameter, either c or L to determine. More generally, it is obvious from Eq. (3.17) that one has information only about c/L—they could not be determined separately.

Suppose however, that the density varies along the string, $\rho = \rho(r_x)$, so that $c = c(r_x)$. Then (it may be confirmed) that the observed frequencies are no longer given by Eq. (3.17), but by expressions involving the integral of c over the length of the string. An important problem is then to infer $c(r_x)$, and hence $\rho(r_x)$. One might wonder whether, under these new circumstances L can be determined independently of c?

A host of such problems, in which the observed frequencies of free modes are used to infer properties of media in one to three dimensions exists. The most elaborate applications are in the geophysics, where the normal mode frequencies of the vibrating whole earth are used to infer the interior properties (density and elastic parameters).[6] A good exercise is to render the variable string problem in discrete form.

4. Importance of the Forward Model

Inference about the physical world from data requires assertions about the structure of the data and its own relationships. One sometimes hears claims from people who are expert in

[5] Oceanographers will recognize this apparently highly artificicial problem as being a slightly simplified version of the so-called geostrophic inverse problem, and which is of great practical importance. It is a central subject in Chapter 5.

[6] Aki and Richards (1980). A famous two-dimensional version of the problem is described by Kac (1966); see also Gordon and Webb (1996).

measurements that “I don’t use models.” Such a claim is almost always completely vacuous. What the speaker usually means is that he doesn’t use equations, but is manipulating his data in some simple way (e.g., forming an average) that seems to be so unsophisticated that no model is present. Consider a simple problem faced by someone trying to determine the average temperature in a room. A thermometer is successively placed at different three-dimensional locations, $\mathbf{r}_i$, at times t_i. Let the measurements be y_i and the value of interest is

$$\tilde{m} = \frac{1}{M}\sum_{i=1}^{M} y_i. \tag{4.1}$$

In deciding to compute, and use $\tilde{m}$, the observer has probably made a long list of very sophisticated, but implicit, model assumptions. Among them we might suggest: (1) that the temperature in the room is sufficiently slowly changing that all of the t_i can be regarded as identical. A different observer might suggest that the temperature in the room is governed by shock waves bouncing between the walls at intervals of seconds or less. Should that be true, $\tilde{m}$ might prove completely meaningless. It might be objected that such an hypothesis is far-fetched. But the assumption that the room temperature is governed, e.g., by a slowly evolving diffusion process, is a rigid, and perhaps incorrect model. (2) That the errors in the thermometer are such that the best estimate of the room mean temperature is obtained by the simple sum in Eq. (4.1). There are many measurement devices for which this assumption is a very poor one (perhaps the instrument is drifting, or has a calibration that varies with temperature), and we will discuss how to determine averages in Chapter 2. But the assumption that property $\tilde{m}$ is useful, is a strong model assumption concerning the instrument being used and the physical process it is measuring.

This list can be extended (the interpretation of the mean is itself model-dependent), but more generally, all of the inverse problems listed in this chapter only make sense to the degree that the underlying forward model is likely to be an adequate physical description of the observations. For example, if one is attempting to determine ρ in Eq. (2.4) by taking the Laplacian $\nabla^2\phi$, (analytically or numerically), this solution to the inverse problem is only sensible if this equation really represents the correct governing physics.. If the correct equation to use were,

$$\frac{\partial^2\phi}{\partial r_x^2} + \frac{1}{2}\frac{\partial\phi}{\partial r_y} = \rho, \tag{4.2}$$

the calculated value of ρ would be incorrect. One might however, have good reason to use Eq. (2.4) as the most likely hypothesis, but nonetheless remain open to the possibility that it is not an adequate descriptor of the required field, ρ. A good methodology, of the type we will develop in susbsequent chapters, permits one to ask the question: is my model consistent with the data? If the answer to the question is “yes,” a careful investigator would *never* claim that the resulting answer is the correct one and that the model has been “validated” or “verified.” One claims only that the answer and the model are consistent with the observations, and remains open to the possibility that some new piece of information will be obtained that completely *invalidates* the model (e.g., some direct measurements of ρ showing that the inferred value is simply wrong). One can never validate or verify a model, one can only show consistency with existing observations.[7]

[7] Oreskes et al. (1994).

CHAPTER 2

Basic Machinery

1. Background

The purpose of this chapter is to record a number of results which are useful in finding and understanding the solutions to sets of usually noisy simultaneous linear equations and in which formally there may be too much or too little information. A lot of this material is elementary; good textbooks exist, to which the reader will be referred. Some of what follows is discussed here primarily so as to produce a consistent notation for later use. But some topics are given what may be an unfamiliar interpretation, and I urge everyone to at least skim the chapter.

Our basic tools are those of matrix and vector algebra as they relate to the solution of linear simultaneous equations, and some elementary statistical ideas mainly concerning covariance, correlation, and dispersion. Least-squares is reviewed, with an emphasis placed upon the arbitrariness of the distinction between knowns, unknowns and noise. The singular-value decomposition is a central building block, producing the clearest understanding of least-squares and related formulations. We introduce minimum variance estimation through the Gauss-Markov theorem as an alternative method for obtaining solutions to simultaneous equations, and show its relation to and distinction from least-squares. The Chapter ends with a brief discussion of recursive least-squares and estimation; this part is essential background for the study of time-dependent problems in Chapter 4.

2. Matrix and Vector Algebra

This subject is very large and well-developed and it is not my intention to repeat material better found elsewhere[1]. Only a brief survey of essential results is provided.

A matrix is an $M \times N$ array of elements of the form

$$\mathbf{A} = \{A_{ij}\}, \quad 1 \leq i \leq M, \ 1 \leq j \leq N\,.$$

Normally a matrix is denoted by a bold-faced capital letter. A vector is a special case of an $M \times 1$ matrix, written as a bold-face lower case letter, for example, $\mathbf{q}$. Corresponding capital or lower case letters for Greek symbols are also indicated in bold-face. Unless otherwise stipulated, vectors are understood to be columnar. The transpose of a matrix interchanges its rows and

[1] Noble & Daniel (1977); Strang (1988).

columns. Transposition applied to vectors is sometimes used to save space in printing, for example, $\mathbf{q} = [q_1, q_{2,}..., q_N]^T$ is the same as

$$\mathbf{q} = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_N \end{bmatrix} .$$

Matrices and Vectors

The inner, or dot, product between two $L \times 1$ vectors $\mathbf{a}$, $\mathbf{b}$ is written $\mathbf{a}^T\mathbf{b} \equiv \mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{L} a_i b_i$ and is a scalar. Such an inner product is the "projection" of $\mathbf{a}$ onto $\mathbf{b}$ (or vice-versa). A conventional measure of length of a vector is $\sqrt{\mathbf{a}^T\mathbf{a}} = \sqrt{\sum_i^N a_i^2} = \|\mathbf{a}\|$ .It is readily shown that $|\mathbf{a}^T\mathbf{b}| \le \|\mathbf{a}\| \, \|\mathbf{b}\|$; the magnitude of this projection can be measured as,

$$\mathbf{a}^T\mathbf{b} = \|\mathbf{a}\| \, \|\mathbf{b}\| \cos\phi \, ,$$

where $\cos\phi$ ranges between zero, when the vectors are orthogonal, and one, when they are parallel.

Suppose we have a collection of N vectors, $\mathbf{e}_i$, each of dimension N. If it is possible to represent perfectly an arbitrary N–dimensional vector $\mathbf{f}$ as the linear sum,

$$\mathbf{f} = \sum_{i=1}^{N} \alpha_i \mathbf{e}_i \, , \tag{2.1}$$

then $\mathbf{e}_i$ are said to be a "spanning set" or a "basis." A necessary and sufficient condition for them to have that property is that they should be "independent," that is, no one of them should be perfectly representable by the others:

$$\mathbf{e}_{j_0} - \sum_{i=1,\, i \ne j_0}^{N} \beta_i \mathbf{e}_i \ne 0, \quad 1 \le j_0 \le N \, . \tag{2.2}$$

A subset of the $\mathbf{e}_j$ are said to span a subspace (all vectors perfectly representable by the subset). For example, $[1, -1, 0]^T$, $[1, 1, 0]^T$ span the subspace of all vectors $[v_1, v_2, 0]$.

The expansion coefficients α_i in (2.1) are obtained by taking the dot product of (2.1) with each of the vectors in turn:

$$\sum_{i=1}^{N} \alpha_i \mathbf{e}_k^T \mathbf{e}_i = \mathbf{e}_k^T \mathbf{f}, \quad 1 \le k \le N \, , \tag{2.3}$$

a system of N equations in N unknowns. The α_i are most readily found if the $\mathbf{e}_i$ are a mutually orthonormal set, that is, if

$$\mathbf{e}_i^T \mathbf{e}_j = \delta_{ij} \, ,$$

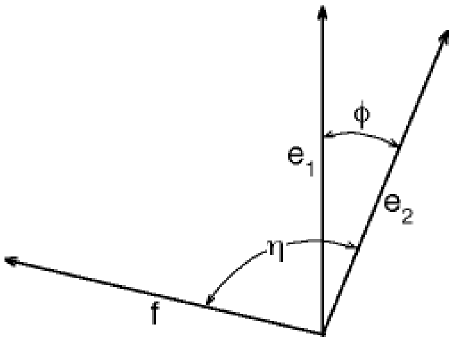


FIGURE 1

but this requirement is not a necessary one. With a spanning set, the information contained in the set of projections, $\mathbf{e}_i^T\mathbf{f} = \mathbf{f}^T\mathbf{e}_i$, is adequate then to determine the α_i and thus all the information required to reconstruct $\mathbf{f}$.

The concept of "nearly-dependent" vectors is helpful and can be understood heuristically. Consider figure 1, in which the space is two-dimensional. Then the two vectors $\mathbf{e}_1$, $\mathbf{e}_2$, as depicted there, are independent and can be used to expand an arbitrary two-dimensional vector $\mathbf{f}$ in the plane. The simultaneous equations become

$$\begin{aligned} \alpha_1\mathbf{e}_1^T\mathbf{e}_1+\alpha_2\mathbf{e}_1^T\mathbf{e}_2 &= \mathbf{e}_1^T\mathbf{f} \\ \alpha_1\mathbf{e}_2^T\mathbf{e}_1+\alpha_2\mathbf{e}_2^T\mathbf{e}_2 &= \mathbf{e}_2^T\mathbf{f} \end{aligned} \tag{2.4}$$

If the vectors become nearly parallel, as the angle ϕ in Fig. 1 goes to zero, as long as they are not identically parallel, they can still be used mathematically to represent $\mathbf{f}$ perfectly. An important feature is that even if the lengths of $\mathbf{e}_{1,}\mathbf{e}_2, \mathbf{f}$ are all order-one, the expansion coefficients $a_{1,2}$ can have extremely large magnitudes when the angle ϕ becomes small and $\mathbf{f}$ is nearly orthogonal to both (measured by angle η). That is to say, we find readily from (2.4),

$$\alpha_1 = \frac{\left(\mathbf{e}_1^T\mathbf{f}\right)\left(\mathbf{e}_2^T\mathbf{e}_2\right)-\left(\mathbf{e}_2^T\mathbf{f}\right)\left(\mathbf{e}_1^T\mathbf{e}_2\right)}{\left(\mathbf{e}_1^T\mathbf{e}_1\right)\left(\mathbf{e}_2^T\mathbf{e}_2\right)-\left(\mathbf{e}_1^T\mathbf{e}_2\right)^2}, \tag{2.5}$$

$$\alpha_2 = \frac{\left(\mathbf{e}_2^T\mathbf{f}\right)\left(\mathbf{e}_1^T\mathbf{e}_1\right)-\left(\mathbf{e}_1^T\mathbf{f}\right)\left(\mathbf{e}_2^T\mathbf{e}_1\right)}{\left(\mathbf{e}_1^T\mathbf{e}_1\right)\left(\mathbf{e}_2^T\mathbf{e}_2\right)-\left(\mathbf{e}_1^T\mathbf{e}_2\right)^2}. \tag{2.6}$$

Suppose for simplicity, that $\mathbf{f}$ has unit length, and that the $\mathbf{e}_i$ have also been normalized to unit length as shown in Figure 1. We have then,

$$\alpha_1 = \frac{\cos\left(\eta-\phi\right)-\cos\boldsymbol{\phi}\cos\eta}{1-\cos^2\phi} = \frac{\sin\eta}{\sin\phi} \tag{2.7}$$

$$\alpha_2 = \cos\eta - \sin\eta\cot\phi \tag{2.8}$$

and whose magnitudes can become arbitrarily large as $\phi \to 0$. One can imagine a situation in which $\alpha_1\mathbf{e}_1$ and $\alpha_2\mathbf{e}_2$ were separately measured and found to be very large. One could then erroneously infer that the sum vector, $\mathbf{f}$, was equally large. This property of the expansion in non-orthogonal vectors potentially producing large coefficients becomes important later (Chapter 4) as a way of gaining insight into the behavior of so-called non-normal operators. The generalization to higher dimensions is left to the reader's intuition. One anticipates that as ϕ becomes very small, numerical problems can arise in using these "almost parallel" vectors.

Gram-Schmidt Process

One often has a set of p-independent, but non-orthonormal vectors $\mathbf{h}_i$ and it is convenient to find a new set $\mathbf{g}_i$, which are orthonormal. The "Gram-Schmidt process" operates by induction. Suppose we have orthonormalized the first k of the $\mathbf{h}_i$ to a new set, $\mathbf{g}_i$, and wish to generate the $k+1-$st. Let

$$\mathbf{g}_{k+1} = \mathbf{h}_{k+1} - \sum_j^k \gamma_j \mathbf{g}_j \,. \tag{2.9}$$

Because $\mathbf{g}_{k+1}$ must be orthogonal to the preceding $\mathbf{g}_i$, $i = 1,\ k$, we take the dot products of (2.9) with each of these vectors, producing a set of simultaneous equations for the unknown γ_j. The resulting $\mathbf{e}_{k+1}$ is easily given unit norm by division by its length.

If one has the first k of N necessary vectors, one needs an additional $N-k$ independent vectors h_i. There are several possibilities. One might simply generate the necessary extra vectors by filling their elements with random numbers. Or one might take a very simple trial set like $\mathbf{h}_{k+1} = [1, 0, 0, ..., 0]^T$, $\mathbf{h}_{k+2} = [0, 1, 0, ...0], \ldots$. If one is unlucky, the set might prove not to be independent of the existing $\mathbf{g}_i$. But a simple numerical perturbation usually suffices to render them so. In practice, the algorithm is changed to what is usually called the "modified Gram-Schmidt process" for purposes of numerical stability.[2]

2.1. Matrix Multiplication and Identities. It has been found convenient and fruitful to usually define multiplication of two matrices $\mathbf{A}, \mathbf{B}$ by the operation $\mathbf{C} = \mathbf{AB}$, such that

$$C_{ij} = \sum_{p=1}^{P} A_{ip} B_{pj} \,. \tag{2.10}$$

For the definition (2.10) to make sense, $\mathbf{A}$ must be a $M \times P$ matrix and $\mathbf{B}$ must be $P \times N$ (including the special case of $P \times 1$, a column vector). That is, the two matrices must be "conformable." If two matrices are multiplied, or a matrix and a vector are multiplied, conformability is implied—otherwise one can be assured that an error has been made. Note

[2] See Lawson & Hanson (1974)

that $\mathbf{AB} \neq \mathbf{BA}$ even where both products exist, except under special circumstances, and that $\mathbf{A}^2 = \mathbf{AA}$, etc. Other definitions of matrix multiplication exist, but they are not needed here.

The mathematical operation in (2.10) may appear arbitrary, but a physical interpretation is available: Matrix multiplication is the dot product of all of the rows of $\mathbf{A}$ with all of the columns of $\mathbf{B}$. Thus multiplication of a vector by a matrix represents the projections of the rows of the matrix onto the vector.

If we define a matrix, $\mathbf{E}$, each of whose columns is the corresponding vector $\mathbf{e}_i$, and a vector, $\boldsymbol{\alpha} = \{\alpha_i\}$, in the same order, the expansion (2.1) can be written in the compact form

$$\mathbf{f} = \mathbf{E}\boldsymbol{\alpha}\,. \tag{2.11}$$

The transpose of a matrix $\mathbf{A}$ is written $\mathbf{A}^T$ and is defined as $\{A_{ij}\}^T = A_{ji}$, an interchange of the rows and columns of $\mathbf{A}$. Thus $\left(\mathbf{A}^T\right)^T = \mathbf{A}$. A "symmetric matrix" is one for which $\mathbf{A}^T = \mathbf{A}$. The product $\mathbf{A}^T\mathbf{A}$ represents the array of all the dot products of the columns of $\mathbf{A}$ with themselves, and similarly, $\mathbf{AA}^T$ represents the set of all dot products of all the rows of $\mathbf{A}$ with themselves. It follows that $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$. Because we have $(\mathbf{AA}^T)^T = \mathbf{AA}^T$, $(\mathbf{A}^T\mathbf{A})^T = \mathbf{A}^T\mathbf{A}$, both these matrices are symmetric.

The "trace" of a square $M \times M$ matrix $\mathbf{A}$ is defined as $\mathrm{trace}(\mathbf{A}) = \sum_i^M A_{ii}$. A "diagonal matrix" is square and zero except for the terms along the main diagonal, although we will later generalize this definition. The operator $\mathrm{diag}(\mathbf{q})$ forms a square diagonal matrix with $\mathbf{q}$ along the main diagonal.

The special $L \times L$ diagonal matrix $\mathbf{I}_L$, with $I_{ii} = 1$, is the "identity." Usually, when the dimension of $\mathbf{I}_L$ is clear from the context, the subscript is omitted. $\mathbf{IA} = \mathbf{A}$, $\mathbf{AI} = \mathbf{I}$, for any $\mathbf{A}$ for which the products make sense. If there is a matrix $\mathbf{B}$, such that $\mathbf{BE} = \mathbf{I}$, then $\mathbf{B}$ is the "left-inverse" of $\mathbf{E}$. If $\mathbf{B}$ is the left inverse of $\mathbf{E}$ and $\mathbf{E}$ is square, a standard result is that it must also be a right inverse: $\mathbf{EB} = \mathbf{I}$, $\mathbf{B}$ is then called "the inverse of $\mathbf{E}$" and is usually written $\mathbf{E}^{-1}$. Analytical expressions exist for inverses, and numerical linear algebra books explain how to find them, when they exist. If $\mathbf{E}$ is not square, one must distinguish left and right inverses, sometimes written $\mathbf{E}^+$ and referred to as "generalized inverses." Some of them will be encountered later. A useful result is that $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$, if the inverses exist. Square matrices with inverses are "non-singular."

A definition of the "length," or norm of a vector has already been introduced. But several choices are possible; for present purposes, the conventional l_2 norm already defined,

$$\|\mathbf{f}\|_2 \equiv (\mathbf{f}^T\mathbf{f})^{1/2} = \left(\sum_{i=1}^N f_i^2\right)^{1/2}, \tag{2.12}$$

is most useful; often the subscript will be omitted. Eq. (2.12) leads in turn to the measure of distance between two vectors, **a**, **b** as

$$\|\mathbf{a}-\mathbf{b}\|_2 = \sqrt{(\mathbf{a}-\mathbf{b})^T\,(\mathbf{a}-\mathbf{b})}\,, \tag{2.13}$$

the familiar Cartesian distance. Distances can also be measured in such a way that deviations of certain elements of $\mathbf{c}=\mathbf{a}-\mathbf{b}$ count for more than others—that is, a metric, or set of weights can be introduced with a definition

$$\|\mathbf{c}\|_W = \sqrt{\sum\nolimits_i c_i W_{ii} c_i}\,, \tag{2.14}$$

depending upon the importance to be attached to magnitudes of different elements, stretching and shrinking various coordinates. Finally, in the most general form, distance can be measured in a coordinate system both stretched and rotated relative to the original one

$$\|\mathbf{c}\|_W = \sqrt{\mathbf{c}^T\mathbf{W}\mathbf{c}} \tag{2.15}$$

where $\mathbf{W}$ is an arbitrary matrix (but usually, for physical reasons, symmetric and positive definite[3]).

2.2. Linear Simultaneous Equation*s*. Consider a set of M-linear equations in N-unknowns,

$$\mathbf{E}\mathbf{x} = \mathbf{y}. \tag{2.16}$$

Because of the appearance of simultaneous equations in situations in which the y_i are observed, and where $\mathbf{x}$ are parameters we wish to determine, it is often convenient to refer to (2.16) as a set of measurements of $\mathbf{x}$ which produced the observations or data, $\mathbf{y}$. If $M > N$, the system is said to be "formally overdetermined." If $M < N$, it is "underdetermined," and if $M = N$, it is "formally just-determined." The use of the word "formally" has a purpose we will come to. Knowledge of the matrix inverse to $\mathbf{E}$ would make it easy to solve a set of L equations in L unknowns, by left-multiplying (2.16) by $\mathbf{E}^{-1}$:

$$\mathbf{E}^{-1}\mathbf{E}\mathbf{x} = \mathbf{I}\mathbf{x} = \mathbf{x} = \mathbf{E}^{-1}\mathbf{y}$$

The reader is cautioned that although matrix inverses are a very powerful tool, one is usually ill-advised to solve large sets of simultaneous equations by employing $\mathbf{E}^{-1}$; other numerical methods are available for the purpose[4].

[3] "Positive definite" will be defined below. Here it suffices to mean that $\mathbf{c}^T\mathbf{W}\mathbf{c}$ should never be negative, for any $\mathbf{c}$.

[4] Golub & Van Loan (1989)

There are several ways to view the meaning of any set of linear simultaneous equations. If the columns of $\mathbf{E}$ continue to be denoted $\mathbf{e}_i$, then (2.16) is,

$$x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_n\mathbf{e}_N = \mathbf{y}\,. \tag{2.17}$$

The ability to so describe an arbitrary $\mathbf{y}$, or to solve the equations, would thus depend upon whether the $M \times 1$, vector $\mathbf{y}$ can be specified by a sum of N-column vectors, $\mathbf{e}_i$. That is, it would depend upon their being a spanning set. In this view, the elements of $\mathbf{x}$ are simply the corresponding expansion coefficients. Depending upon the ratio of M to N, that is, the number of equations compared to the number of unknown elements, one faces the possibility that there are fewer expansion vectors $\mathbf{e}_i$ than elements of $\mathbf{y}$ $(M > N)$, or that there are more expansion vectors available than elements of $\mathbf{y}$ $(M < N)$. Thus the overdetermined case corresponds to having *fewer* expansion vectors, and the underdetermined case corresponds to having *more* expansion vectors, than the dimension of $\mathbf{y}$. It is possible that in the overdetermined case, the too-few expansion vectors are not actually independent, so that there are even fewer vectors available than is first apparent. Similarly, in the underdetermined case, there is the possibility that although it appears we have more expansion vectors than required, fewer may be independent than the number of elements of $\mathbf{y}$, and the consequences of that case need to be understood as well.

An alternative interpretation denotes the rows of $\mathbf{E}$ as $\mathbf{r}_i^T$, $1 \le i \le M$. Then Eq.(2.16) is a set of M-inner products,

$$\mathbf{r}_i^T\mathbf{x} = y_i, \quad 1 \le i \le M\,. \tag{2.18}$$

That is, the set of simultaneous equations is also equivalent to being provided with the value of M–dot products of the N–dimensional unknown vector, $\mathbf{x}$, with M known vectors, $\mathbf{r}_i$. Whether that is sufficient information to determine $\mathbf{x}$ depends upon whether the $\mathbf{r}_i$ are a spanning set. In this view, in the overdetermined case one has *more* dot products available than unknown elements x_i, and in the underdetermined case, there are *fewer* such values than unknowns.

A special set of simultaneous equations for square matrices $\mathbf{A}$ is labelled the "eigenvalue/-eigenvector problem,"

$$\mathbf{Ae} = \lambda\mathbf{e}. \tag{2.19}$$

In this set of linear simultaneous equations one seeks a special vector, $\mathbf{e}$, such that for some as yet unknown scalar eigenvalue, λ, there is a solution. An $N \times N$ matrix will have up to N solutions $(\lambda_i, \mathbf{e}_i)$, but the nature of these elements and their relations require considerable effort to deduce. We will look at this problem more later; for the moment, it again suffices to say that numerical methods for solving Eq. (2.19) are well-known.

2.3. Matrix Norms. A number of useful definitions of a matrix size, or norm, exist. For present purposes the so-called "spectral norm" or "2–norm" defined as

$$\|\mathbf{A}\|_2 = \sqrt{\text{maximum eigenvalue of } (\mathbf{A}^T\mathbf{A})} \tag{2.20}$$

is usually adequate. Without difficulty, it may be seen that this definition is equivalent to

$$\|\mathbf{A}\|_2 = \max \frac{\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \max \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \tag{2.21}$$

where the maximum is defined over all vectors $\mathbf{x}$[5]. Another useful measure is the "Frobenius norm,"

$$\|\mathbf{A}\|_F = \sqrt{\sum\nolimits_{i=1}^{M}\sum\nolimits_{j=1}^{N} A_{ij}^2} = \sqrt{\text{trace}(\mathbf{A}^T\mathbf{A})}\,. \tag{2.22}$$

Neither norm requires $\mathbf{A}$ to be square.

These norms permit one to derive various useful results. Consider one illustration. Suppose $\mathbf{Q}$ is square, and $\|\mathbf{Q}\| < 1$, then

$$(\mathbf{I}+\mathbf{Q})^{-1} = \mathbf{I} - \mathbf{Q} + \mathbf{Q}^2 - \cdots\,, \tag{2.23}$$

which may be verified by multiplying both sides by $\mathbf{I}+\mathbf{Q}$, doing term-by-term multiplication and measuring the remainders with either norm.

2.4. Identities. Differentiation. There are some identities and matrix/vector definitions which prove useful.

A square "positive definite" matrix $\mathbf{A}$, is one for which the scalar "quadratic form,"

$$J = \mathbf{x}^T\mathbf{A}\mathbf{x} \tag{2.24}$$

is positive for all possible vectors $\mathbf{x}$. (It suffices to consider only symmetric $\mathbf{A}$ because for a general matrix, $\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T[(\mathbf{A}+\mathbf{A}^T)/2]\mathbf{x}$, which follows from the scalar property of the quadratic form.) If $J \geq 0$ for all $\mathbf{x}$, $\mathbf{A}$ is "positive semi-definite," or "non-negative definite." Linear algebra books show that a necessary and sufficient requirement for positive definiteness is that $\mathbf{A}$ have only positive eigenvalues (Eq. 2.19) and a semi-definite one must have all non-negative eigenvalues.

Nothing has been said about actually finding the numerical values of either the matrix inverse or the eigenvectors and eigenvalues. Computational algorithms for obtaining them have been developed by experts, and are discussed in many good textbooks.[6] Software systems like MATLAB, Maple, IDL and Mathematica implement them in easy-to-use form. For purposes of this book, we assume the reader has at least a rudimentary knowledge of these techniques and access to a good software implementation.

[5] Haykin (1986, p. 61).

[6] Press *et al.*, 1992; Lawson & Hanson, 1974; Golub & van Loan, 1989; etc.

We end up doing a certain amount of differentiation and other operations with respect to matrices and vectors. A number of formulas are very helpful, and save a lot of writing. They are all demonstrated by doing the derivatives term-by-term. Let $\mathbf{q}$, $\mathbf{r}$ be $N \times 1$ column vectors, and $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ be matrices. Then if s is any scalar,

$$\frac{\partial s}{\partial \mathbf{q}} = \left\{ \begin{matrix} \frac{\partial s}{\partial q_1} \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial s}{\partial q_N} \end{matrix} \right\} \tag{2.25}$$

is a vector (the gradient). The second derivative of a scalar,

$$\frac{\partial^2 s}{\partial \mathbf{q}^2} = \left\{ \frac{\partial}{\partial \mathbf{q}_i} \frac{\partial s}{\partial \mathbf{q}_j} \right\} = \left\{ \begin{matrix} \frac{\partial^2 s}{\partial \mathbf{q}_1^2} & \frac{\partial^2 s}{\partial \mathbf{q}_1 \mathbf{q}_2} & \cdot & \cdot & \frac{\partial^2 s}{\partial \mathbf{q}_1 \mathbf{q}_N} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \frac{\partial^2 s}{\partial \mathbf{q}_N \mathbf{q}_1} & \cdot & \cdot & \cdot & \frac{\partial^2 s}{\partial \mathbf{q}_N^2} \\ & & & & \end{matrix} \right\} \tag{2.26}$$

is the "Hessian" of s.

The derivative of one vector by another is a matrix:

$$\frac{\partial \mathbf{r}}{\partial \mathbf{q}} = \left\{ \frac{\partial \mathbf{r}_i}{\partial \mathbf{q}_j} \right\} = \left\{ \begin{matrix} \frac{\partial \mathbf{r}_1}{\partial \mathbf{q}_1} & \frac{\partial \mathbf{r}_2}{\partial \mathbf{q}_1} & \cdot & \frac{\partial \mathbf{r}_M}{\partial \mathbf{q}_1} \\ \frac{\partial \mathbf{r}_1}{\partial \mathbf{q}_2} & \cdot & \cdot & \frac{\partial \mathbf{r}_M}{\partial \mathbf{q}_2} \\ \cdot & \cdot & \cdot & \cdot \\ \frac{\partial \mathbf{r}_1}{\partial \mathbf{q}_N} & \cdot & \cdot & \frac{\partial \mathbf{r}_M}{\partial \mathbf{q}_N} \\ & & & \end{matrix} \right\} \equiv \mathbf{B}. \tag{2.27}$$

If $\mathbf{r}$, $\mathbf{q}$ are of the same dimension, the determinant of $\mathbf{B} = \det(\mathbf{B})$ is the "Jacobian" of $\mathbf{r}$.[7]

Assuming conformability, the inner product, $J = \mathbf{r}^T\mathbf{q} = \mathbf{q}^T\mathbf{r}$, is a scalar. The differential of J is,

$$dJ = d\mathbf{r}^T\mathbf{q} + \mathbf{r}^T d\mathbf{q} = d\mathbf{q}^T\mathbf{r} + \mathbf{q}^T d\mathbf{r}, \tag{2.28}$$

and hence the partial derivatives are,

$$\frac{\partial(\mathbf{q}^T\mathbf{r})}{\partial \mathbf{q}} = \frac{\partial\left(\mathbf{r}^T\mathbf{q}\right)}{\partial \mathbf{q}} = \mathbf{r}\,, \tag{2.29}$$

$$\frac{\partial\left(\mathbf{q}^T\mathbf{q}\right)}{\partial \mathbf{q}} = 2\mathbf{q}\,. \tag{2.30}$$

[7] Determinants are used only rarely in this book. Their definition and properties are left to the references, as they are usually encountered in high school mathematics.

It follows immediately that for matrix/vector products,

$$\frac{\partial}{\partial \mathbf{q}}(\mathbf{Bq}) = \mathbf{B}^T, \; \frac{\partial}{\partial \mathbf{q}}\left(\mathbf{q}^T\mathbf{B}\right) = \mathbf{B}. \tag{2.31}$$

The first of these is used repeatedly, and attention is called to the apparently trivial fact that differentiation of $\mathbf{Bq}$ with respect to $\mathbf{q}$ produces the transpose of $\mathbf{B}$—the origin, as seen later, of so-called adjoint models. For a quadratic form,

$$\begin{aligned} J &= \mathbf{q}^T\mathbf{Aq} \\ \frac{\partial J}{\partial \mathbf{q}} &= 2\mathbf{Aq}\,, \end{aligned} \tag{2.32}$$

and its Hessian is $2\mathbf{A}$.[8]

There are a few, unfortunately unintuitive, matrix inversion identities which are essential later. They are derived by considering the square, partitioned matrix,

$$\begin{Bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{Bmatrix} \tag{2.33}$$

where $\mathbf{A}^T = \mathbf{A}$, $\mathbf{C}^T = \mathbf{C}$, but $\mathbf{B}$ can be rectangular of conformable dimensions in (2.33)[9]. The most important of the identities, sometimes called the "matrix inversion lemma" is, in one form,

$$\begin{aligned} \{\mathbf{C} - \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\}^{-1} &= \{\mathbf{I} - \mathbf{C}^{-1}\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\}^{-1}\mathbf{C}^{-1} \\ &= \mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{B}^T(\mathbf{B}\mathbf{C}^{-1}\mathbf{B}^T - \mathbf{A})^{-1}\mathbf{B}\mathbf{C}^{-1} \end{aligned} \tag{2.34}$$

where it is assumed that the inverses exist.[10] A variant is,

$$\mathbf{AB}^T(\mathbf{C} + \mathbf{BAB}^T)^{-1} = (\mathbf{A}^{-1} + \mathbf{B}^T\mathbf{C}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{C}^{-1}. \tag{2.35}$$

Eq. (2.35) is readily confirmed by left-multiplying both sides by $(\mathbf{A}^{-1} + \mathbf{B}^T\mathbf{C}^{-1}\mathbf{B})$, and right-multiplying by $(\mathbf{C} + \mathbf{BAB}^T)$ and showing that the two sides of the resulting equation are equal. Another identity, found by "completing the square," is demonstrated by directly multiplying it out, and requires $\mathbf{C} = \mathbf{C}^T$ ($\mathbf{A}$ is unrestricted, but the matrices must be conformable as shown):

$$\mathbf{ACA}^T - \mathbf{BA}^T - \mathbf{AB}^T = (\mathbf{A} - \mathbf{BC}^{-1})\mathbf{C}(\mathbf{A} - \mathbf{BC}^{-1})^T - \mathbf{BC}^{-1}\mathbf{B}^T\,. \tag{2.36}$$

[8] Rogers (1980) is an entire volume of matrix derivative identities, and many other useful properties are discussed by Magnus and Neudecker (1988).

[9] Liebelt (1967, Section 1–19)

[10] The history of this not-very-obvious identity is discussed by Haykin (1986, p. 385).

3. Simple Statistics. Regression

3.1. Probability Densities, Moments. Some statistical ideas are required, but the discussion is confined to stating some basic notions and to developing a notation.[11] We require the idea of a probability density for a random variable x. This subject is a very deep one[12], but our approach will be heuristic. Suppose that an arbitrarily large number of experiments can be conducted for the determination of the values of x, denoted X_i, $1 \leq i \leq M$, and a histogram of the experimental values found. The frequency function, or probability density, will be defined as the limit, supposing it exists, of the histogram of an arbitrarily large number of experiments, $M \rightarrow \infty$, divided into bins of arbitrarily small value ranges, and normalized by M, to produce the fraction of the total appearing in the ranges. Let the corresponding limiting frequency function be denoted $p_x(X)dX$, interpreted as the fraction (probability) of values of x lying in the range, $X \leq x \leq X + dX$. As a consequence of the definition, $p_x\left(X\right) \geq 0$ and,

$$\int_{\text{all }X} p_x\left(X\right)dX = \int_{-\infty}^{\infty} p_x\left(X\right)dX = 1. \tag{3.1}$$

(The infinite integral is a convenient way of representing an integral over "all X", as p_x simply vanishes for impossible values of X.)

The "average," or "mean," or "expected value" is denoted $\langle x\rangle$ and defined as,

$$\langle x\rangle \equiv \int_{\text{all }X} X p_x(X)dX \; = m_1. \tag{3.2}$$

The mean is the center of mass of the probability density. Knowledge of the true mean value of a random variable is commonly all that we are willing to assume known. If forced to "forecast" the numerical value of x under such circumstances, often the best we can do is to employ $\langle x\rangle$. If the deviation from the true mean is denoted x' so that $x = \langle x\rangle + x'$, such a forecast has the virtue that we are assured the average forecast error, $\langle x'\rangle$, would be zero if many such forecasts are made. The bracket operation is very important throughout this book; it has the property that if a is a non-random quantity, $\langle ax\rangle = a\langle x\rangle$ and $\langle ax + y\rangle = a\left\langle x\right\rangle + \left\langle y\right\rangle$.

Quantity $\langle x\rangle$ is the "first-moment" of the probability density. Higher order moments are defined as,

$$m_n = \langle x^n\rangle = \int_{\text{-}\infty}^{\infty} X^n p_x(X)dX,$$

where n are the non-negative integers. A useful theoretical result is that a knowledge of all the moments is usually enough to completely define the probability density itself. (There are troublesome situations with, e.g. non-existent moments, as with the so-called Cauchy distribution, $p_x\left(X\right) = \left(1/\pi\right)\left(1/\left(1+X^2\right)\right)$, whose mean is infinite.) For many important probability

[11] A good statistics text such as Cramér (1946), or one on regression such as Seber (1977), should be consulted.

[12] For example, Feller (1957) or Jeffreys (1961).

densities, including the Gaussian, a knowledge of the first two moments $n = 1, 2$ is sufficient to define all the others, and hence the full probability density. It is common to define the moments for $n > 1$ about the mean, so that one has

$$\mu_n = \langle (x - \langle x \rangle)^n \rangle = \int_{-\infty}^{\infty} (X - \langle X \rangle)^n \, p_x(X) dX.$$

For $n = 2$, it is called the variance, often written $\mu_2 = \sigma^2$, where σ is the "standard deviation."

3.2. Sample Estimates. Bias. In observational sciences, one normally must estimate the values defining the probability density from the data itself. Thus the first moment, the mean, is often computed as the "sample average,"

$$\tilde{m}_1 = \langle x \rangle_M \equiv \frac{1}{M} \sum_{i=1}^{M} X_i. \tag{3.3}$$

The notation $\tilde{m}_1$ is used to distinguish the sample estimate from the true value, m_1. On the other hand, if the experiment of computing $\tilde{m}_1$ from M samples could be repeated many times, the mean of the sample estimates would be the true mean. This conclusion is readily seen by considering the expected value of the difference from the true mean:

$$\begin{aligned} \langle \langle x \rangle_M - \langle x \rangle \rangle &= \left\langle \frac{1}{M} \sum_{i=1}^{M} X_i - \langle x \rangle \right\rangle \\ &= \frac{1}{M} \sum_{i=1}^{M} \langle X_i \rangle - \langle x \rangle = \frac{M}{M} \langle x \rangle - \langle x \rangle = 0. \end{aligned}$$

Such an estimate, is said to be "unbiassed": its expected value is the quantity one seeks.

The interpretation is that for finite M, we do not expect that the sample mean will equal the true mean, but that if we could produce sample averages from distinct groups of observations, the sample averages would themselves have an average which will fluctuate about the true mean, with equal probability of being higher or lower. There are many sample estimates however, some of which we encounter, where the expected value of the sample estimate is not equal to the true estimate. Such an estimator is said to be "biassed." The simplest example of a biassed estimator is the "sample variance," defined as

$$s^2 \equiv \frac{1}{M} \sum_{i}^{M} (X_i - \langle x \rangle_M)^2 \tag{3.4}$$

For reasons explained a bit later, one finds that

$$\langle s^2 \rangle = \frac{M-1}{M} \sigma^2$$

and thus the expected value is not the true variance. (This particular estimate is "asymptotically unbiassed" as the bias vanishes as $M \to \infty$.)

If one forms, e.g., the sample mean, we are assured that it is unbiassed. But the probability that $\langle x\rangle_M = \langle x\rangle$, that is that we obtain exactly the true value, is very small. It helps to have a measure of the extent to which $\langle x\rangle_M$ is likely to be very far from $\langle x\rangle$. We need the idea of dispersion—the expected or average squared value of some quantity about some interesting value, like its mean. The most familiar measure of dispersion is the variance, already used above, the expected fluctuation of a random variable about its mean:

$$\sigma^2 = \langle (x - \langle x\rangle)^2\rangle \,.$$

More generally, define the dispersion of any random variable z as,

$$D^2(z) = \langle z^2\rangle.$$

Thus, the variance of x is $D^2(x - \langle x\rangle)$.

We can thus ask for the variance of $\langle x\rangle_M$ about the correct value. A little algebra using the bracket notation produces,

$$D^2\left((\langle x\rangle_M - x)^2\right) = \frac{\sigma^2}{M}. \tag{3.5}$$

This result shows the well-known result that as M becomes large, any tendency of the sample mean to lie far from the true value will diminish. It does not prove that some particular value will not, by accident, be far away, merely that it becomes increasingly unlikely as M grows. (In statistics textbooks, the Chebyschev inequality is used to formalize this statement.)

An estimate which is unbiassed and whose expected dispersion about the true value goes to zero with M is evidently desirable. In more interesting estimators, a bias is often present. Then for a fixed number of samples, M, there would be two distinct sources of deviation (error) from the true value: (1) the bias—how far, on average, it is expected to be from the true value, and (2) the tendency—from purely random events—for the value to differ from the true value (the random error). In numerous cases, one discovers that tolerating a small bias error can greatly reduce the random error—and thus the bias may well be worth accepting for that reason. In some cases therefore, a bias is deliberately introduced.

3.3. Multivariable Probability Densities. Correlation. The idea of a frequency function generalizes easily to two or more random variables, x, y. We can, in concept, do an arbitrarily large number of experiments in which we count the occurrences of differing values, $(X_i,\, Y_i)$, of x, y and make a histogram normalized by the total number of samples, taking the limit to produce a joint probability density $p_{xy}(X,\, Y)$, so that $p_{xy}\left(X, Y\right) dXdY$ is the fraction of occurrences such that $X \leq x \leq X + dX, Y \leq y \leq Y + dY$. A simple example would be the probability density for the simultaneous measurement of the two components of horizontal velocity in a current meter or anemometer. Again, from the definition, $p_{xy}(X,Y) \geq 0$, and

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_{xy}(X,Y)\, dX dY = 1. \tag{3.6}$$

An important use of joint probability densities is in what is known as "conditional probability." Suppose that the joint probability density for x, y is known and furthermore, $y = Y$, that is, information is available concerning the actual value of y. What then, is the probability density for x given that a particular value for y is known to have occurred? This new frequency function is usually written as $p_{x|y}(X|Y)$ and read as "the probability of x, given that y has occurred," or, "the probability of x conditioned on y." It follows immediately from the definition of the probability density that

$$p_{x|y}(X|Y) = \frac{p_{xy}(X,\, Y)}{p_y(Y)} \tag{3.7}$$

(this equation is readily understood by going back to the original experimental concept, and understanding the restriction on x, given that y is known to lie within a strip paralleling the X axis).

Using the joint frequency function, define the average product as,

$$\langle xy \rangle = \int\int_{\text{all } X,Y} XY p_{xy}(X,\, Y) dX\, dY\,. \tag{3.8}$$

Suppose that upon examining the joint frequency function, one finds that $p_{xy}(X,\, Y) = p_x(X) p_y(Y)$, that is it factors into two distinct functions. In that case, x, y are said to be "independent." Many important results follow including,

$$\langle xy \rangle = \langle x \rangle \langle y \rangle\,.$$

Non-zero mean values are often primarily a nuisance. One can always define modified variables, e.g. $x' = x- < x >$ such that the new variables have zero mean. Alternatively, one computes statistics centered on the mean. Should the centered product $\langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle$ be non-zero, x, y are said to "co-vary" or to be "correlated." Alternatively, if $\langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle = 0$, then the two variables are uncorrelated. If x, y are independent, then $\langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle = 0$. Independence thus implies lack of correlation, but the reverse is not necessarily true. (These are theoretical relationships, and if $\langle x \rangle, \langle y \rangle$ are determined from observation, as described below, one must carefully distinguish actual behavior from that expected theoretically.)

If the two variables are independent, then (3.7) is

$$p_{x|y}(X|Y) = p_x(X)\,, \tag{3.9}$$

that is, knowledge of the value of y does not change the probability density for x—a sensible result—and *there is then no predictive power for one variable given knowledge of the other.*

Suppose there are two random variables x, y between which there is anticipated to be some linear relationship,

$$x = ay + n, \tag{3.10}$$

where n represents any contributions to x that remain unknown despite knowledge of y, and a is a constant. Then,

$$\langle x\rangle = a\langle y\rangle + \langle n\rangle \,, \tag{3.11}$$

and (3.10) can be re-written as,

$$x - \langle x\rangle = a(y - \langle y\rangle) + (n - \langle n\rangle),$$

or

$$x' = ay' + n', \quad x' = x - \langle x\rangle, \quad \text{etc.} \tag{3.12}$$

From this last equation,

$$a = \frac{\langle x'y'\rangle}{\langle y'^2\rangle} = \frac{\langle x'y'\rangle}{(\langle y'^2\rangle\langle x'^2\rangle)^{1/2}}\,\frac{\langle x'^2\rangle^{1/2}}{\langle y'^2\rangle^{1/2}} = \rho\,\frac{\langle x'^2\rangle^{1/2}}{\langle y'^2\rangle^{1/2}}\,, \tag{3.13}$$

where it was supposed that $\langle y'n'\rangle = 0$, thus defining n'. The quantity

$$\rho \equiv \frac{\langle x'y'\rangle}{\langle y'^2\rangle^{1/2}\langle x'^2\rangle^{1/2}} \tag{3.14}$$

is the "correlation coefficient" and is easily shown to have the property $|\rho| \leq 1$. If ρ should vanish, then so does a. If a vanishes, then knowledge of y' carries no information about the value of x'. If $\rho = \pm 1$, then it follows from the definitions that $n = 0$ and knowledge of a permits perfect prediction of x' from knowledge of y' (because probabilities are being used, rigorous usage would state "perfect prediction almost always," but this distinction will be ignored).

A measure of how well the prediction of x' from y' will work can be obtained in terms of the variance of x'. We have

$$\langle x'^2\rangle = a^2\langle y'^2\rangle + \left\langle n'^2\right\rangle = \rho^2\langle x'^2\rangle + \langle n'^2\rangle$$

or

$$(1-\rho^2)\langle {x'}^2\rangle = \langle n'^2\rangle. \tag{3.15}$$

That is, $(1-\rho^2)\langle {x'}^2\rangle$ is the fraction of the variance in x' that is *unpredictable* from knowledge of y' and is the "unpredictable power." Conversely, $\rho^2\langle x'^2\rangle$ is the "predictable" power in x' given knowledge of y'. The limits as $\rho \to 0$, 1 are readily apparent.

Thus we interpret the statement that two variables x', y' "are correlated" or "co-vary" to mean that knowledge of one permits at least a partial prediction of the other, the expected success of the prediction depending upon the size of ρ. If ρ is not zero, the variables cannot be independent, and the conditional probability $p_{x|y}\left(X|Y\right) \neq p_x\left(X\right)$. This result represents an implementation of the statement that if two variables are not independent, then knowledge of one permits some skill in the prediction of the other. If two variables do not co-vary, but are known not to be independent, a linear model like (3.10) would not be useful—a non-linear model would be required. Such non-linear methods are possible, and are touched on briefly later. The idea that correlation or covariance between various physical quantities carries useful predictive skill between them is an essential ingredient of many of the methods taken up in this book.

In most cases, quantities like ρ, $\left\langle x'^2 \right\rangle$, are determined from the available measurements, e.g. of the form,

$$ay_i + n_i = x_i\,, \tag{3.16}$$

and are not known exactly. They are thus sample values, are not equal to the true values, and must be interpreted with caution in terms of their inevitable biases and variances. This large subject of regression analysis is left to the references.[13]

The Gaussian, or normal, probability density is one that is mathematically handy (but is potentially dangerous as a general model of the behavior of natural processes—many geophysical processes are demonstrably non-Gaussian). For a single random variable x, it is defined as,

$$p_x(X) = \frac{\exp - \frac{(X-m_x)^2}{\sigma_x^2}}{\sqrt{2\pi}\sigma_x},$$

(sometimes abbreviated as $G(m_x,\,\sigma_x)$). It is readily confirmed that $\langle x\rangle = m_x$, $\langle (x-\langle x\rangle)^2\rangle = \sigma_x^2$. Suppose that x, y are *independent* Gaussian variables $G(m_x,\,\sigma_x)$, $G(m_y,\,\sigma_y)$. Then their joint probability density is just the product of the two individual densities,

$$p_{x,y}(X,Y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left(-\frac{(X-m_x)^2}{\sigma_x^2} - \frac{(Y-m_y)^2}{\sigma_y^2}\right)\,. \tag{3.17}$$

Let two new random variables, ξ_1, ξ_2, be defined as a linear combination of x, y,

$$\begin{aligned} \xi_1 &= a_{11}(x-m_x) + a_{12}(y-m_y) + m_{\xi_1} \\ \xi_2 &= a_{21}(x-m_x) + a_{22}(y-m_y) + m_{\xi_2}, \end{aligned} \tag{3.18}$$

or in vector form,

$$\boldsymbol{\xi} = \mathbf{A}(\mathbf{x}-\mathbf{m}_x) + \mathbf{m}_\xi,$$

where $\mathbf{x} = \{x,y\}^T$, $\mathbf{m}_x = [m_x, m_y]^T$, $\mathbf{m}_y = [m_{\xi_1}, m_{\xi_2}]^T$. What is the probability density for these new variables? The general rule for changes of variable in probability densities follows from area conservation in mapping from the x, y space to the ξ_1, ξ_2 space, that is,

$$p_{\xi_1\xi_2}(\Xi_1,\Xi_2) = p_{xy}(X(\Xi_1,\Xi_2),\,Y(\Xi_1,\Xi_2))\,\frac{\partial(X,Y)}{\partial(\Xi_1,\Xi_2)} \tag{3.19}$$

[13] See, for example, Draper & Smith (1998); Seber (1979).

where $\partial(X,Y)/\partial(\Xi_1,\Xi_2)$ is the Jacobian of the transformation between the two variable sets, (see Eq. 2.27) and the numerical values satisfy the functional relations,

$$\Xi_1 = a_{11}(X - m_x) + a_{12}(Y - m_y) = m_{\xi_1}\,,$$

etc. Suppose that the relationship (3.18) is invertible, that is, we can solve for,

$$\begin{aligned} x &= b_{11}(\xi_1 - m_{\xi_1}) + b_{12}(\xi_2 - m_{\xi_2}) + m_x \\ y &= b_{21}(\xi_1 - m_{\xi_1}) + b_{22}(\xi_2 - m_{\xi_2}) + m_y, \end{aligned}$$

or,

$$\mathbf{x} = \mathbf{B}(\boldsymbol{\xi} - \mathbf{m}_\xi) + \mathbf{m}_x. \tag{3.20}$$

Then the Jacobian of the transformation is,

$$\frac{\partial(X,Y)}{\partial(\Xi_1,\Xi_2)} = b_{11}b_{22} - b_{12}b_{21} = \det(\mathbf{B}) \tag{3.21}$$

($\det(\mathbf{B})$ is the determinant). Eq. (3.18) produces

$$\begin{aligned} \langle \xi_1 \rangle &= m_{\xi_1} \\ \langle \xi_2 \rangle &= m_{\xi_2} \\ \langle (\xi_1 - \langle \xi_1 \rangle)^2 \rangle &= a_{11}^2\sigma_x^2 + a_{12}\sigma_y^2 \\ \langle (\xi_1 - \langle \xi_1 \rangle)(\xi_2 - \langle \xi_2 \rangle) \rangle &= a_{11}a_{21}\sigma_x^2 + a_{12}a_{22}\sigma_y^2 \neq 0. \end{aligned} \tag{3.22}$$

In the special case,

$$\mathbf{A} = \begin{Bmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{Bmatrix}, \qquad \mathbf{B} = \begin{Bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{Bmatrix}, \tag{3.23}$$

the transformation (3.23) is a simple coordinate rotation through angle ϕ, and the Jacobian is 1. The new second-order moments are,

$$\langle (\xi_1 - \langle \xi_1 \rangle)^2 \rangle = \sigma_{\xi_1}^2 = \cos^2\phi\,\sigma_x^2 + \sin^2\phi\,\sigma_y^2\,, \tag{3.24}$$

$$\langle (\xi_2 - \langle \xi_2 \rangle)^2 \rangle = \sigma_{\xi_2}^2 = \sin^2\phi\,\sigma_x^2 + \cos^2\phi\,\sigma_y^2\,, \tag{3.25}$$

$$\langle (\xi_1 - \langle \xi_1 \rangle)(\xi_2 - \langle \xi_2 \rangle) \rangle \equiv \mu_{\xi_1\xi_2} = (\sigma_y^2 - \sigma_x^2)\cos\phi\,\sin\phi\,. \tag{3.26}$$

The new probability density is

$$p_{\xi_1\xi_2}(\Xi_1,\Xi_2)=\frac{1}{2\pi\sigma_{\xi_1}\sigma_{\xi_2}\sqrt{1-\rho_\xi^2}}$$
$$\exp\left\{-\frac{1}{2\sqrt{1-\rho_\xi^2}}\left[\frac{(\Xi_1-m_{\xi_1})^2}{\sigma_{\xi_1}^2}\right.\right. \tag{3.27}$$
$$\left.\left.-\frac{2\rho_\xi(\Xi_1-m_{\xi_1})(\Xi_2-m_{\xi_2})}{\sigma_{\xi_1}\sigma_{\xi_2}}+\frac{(\Xi_2-m_{\xi_2})^2}{\sigma_{\xi_2}^2}\right]\right\},$$

where $\rho_\xi=(\sigma_y^2-\sigma_x^2)\sin\phi\cos\phi/\left(\sigma_{\xi_1}^2+\sigma_{\xi_2}^2\right)^{1/2}=\mu_{\xi_1\xi_2}/\sigma_{\xi_1}\sigma_{\xi_2}$ is the correlation coefficient of the new variables. A probability density derived through a linear transformation from two independent variables which are Gaussian will be said to be "jointly Gaussian" and (3.27) is a canonical form. Because a coordinate rotation is invertible, it is important to note that if we had two random variables ξ_1,ξ_2 which were jointly Gaussian with $\rho\neq 1$, then we could find a pure rotation (3.23), which produces two other variables x, y which are uncorrelated, and therefore *independent*. Notice that (3.26) shows that two such uncorrelated variables x, y will necessarily have different variances, otherwise ξ_1, ξ_2 would have zero correlation, too, by Eq. (3.26) .

As an important by-product, it is concluded that two jointly Gaussian random variables that are uncorrelated, are also independent. This property is one of the reasons Gaussians are so nice to work with; but it is not generally true of uncorrelated variables.

Vector Random Processes

Simultaneous discussion of two random processes, x, y can regarded as discussion of a vector random process $[x,\ y]^T$, and suggests a generalization to N dimensions. Let us label N random processes as x_i and define them as the elements of a vector $\mathbf{x}=[x_1,x_2,\ldots,x_N]^T$. Then the mean is a vector: $\langle\mathbf{x}\rangle=\mathbf{m}_x$, and the covariance is a matrix:

$$\mathbf{C}_{xx}=D^2(\mathbf{x}-\langle\mathbf{x}\rangle)=\langle(\mathbf{x}-\langle\mathbf{x}\rangle)(\mathbf{x}-\langle\mathbf{x}\rangle)^T\rangle\,, \tag{3.28}$$

which is necessarily symmetric and positive semi-definite. The cross-covariance of two vector processes $\mathbf{x}$, $\mathbf{y}$ is,

$$\mathbf{C}_{xy}=\langle(\mathbf{x}-\langle\mathbf{x}\rangle)(\mathbf{y}-\langle\mathbf{y}\rangle)^T\rangle, \tag{3.29}$$

and $\mathbf{C}_{xy}=\mathbf{C}_{yx}^T$.

It proves convenient to introduce two further moment matrices in addition to the covariance matrices. The "second moment" matrices will be defined as,

$$\mathbf{R}_{xx}\equiv D^2(\mathbf{x})=\langle\mathbf{x}\mathbf{x}^T\rangle,\qquad \mathbf{R}_{xy}=\langle\mathbf{x}\mathbf{y}^T\rangle,$$

that is, not taken about the means. Note $\mathbf{R}_{xy} = \mathbf{R}_{yx}^T$, etc. Let $\tilde{\mathbf{x}}$ be an "estimate" of the true value, $\mathbf{x}$. Then the dispersion of $\tilde{\mathbf{x}}$ about the true value will be called the "uncertainty" (sometimes it is called the "error covariance") and is

$$\mathbf{P} \equiv D^2(\tilde{\mathbf{x}} - \mathbf{x}) = \langle (\tilde{\mathbf{x}} - \mathbf{x})(\tilde{\mathbf{x}} - \mathbf{x})^T \rangle \,.$$

$\mathbf{P}$ is similar to $\mathbf{C}$, but differs in being taken about the true value, rather than about the mean value; the distinction can be very important.

If there are N variables, ξ_i, $1 \le i \le N$, they will be said to have an "N-dimensional jointly normal probability density" if it is of the form

$$p_{\xi_1,\dots,\xi_N}(\Xi_1,\dots,\Xi_N) = \frac{\exp -\frac{1}{2}(\boldsymbol{\Xi} - \mathbf{m})^T \mathbf{C}_{\xi\xi}^{-1} (\boldsymbol{\Xi} - \mathbf{m})}{(2\pi)^{N/2}\sqrt{\det(\mathbf{C}_{\xi\xi})}} \,. \tag{3.30}$$

One finds $\langle \boldsymbol{\xi} \rangle = \mathbf{m}$, $\langle (\boldsymbol{\xi} - \mathbf{m})(\boldsymbol{\xi} - \mathbf{m})^T \rangle = \mathbf{C}_{\xi\xi}$. Eq. (3.27) is a special case for $N = 2$, and so the earlier forms are consistent with this general definition.

Positive definite symmetric matrices can be factored as

$$\mathbf{C}_{\xi\xi} = \mathbf{C}_{\xi\xi}^{T/2} \mathbf{C}_{\xi\xi}^{1/2} \,, \tag{3.31}$$

called the "Cholesky decomposition," where $\mathbf{C}_{\xi\xi}^{1/2}$ is upper triangular (all zeros below the main diagonal) and non-singular.[14]. It follows that the transformation (a rotation and stretching),

$$\mathbf{x} = \mathbf{C}_{\xi\xi}^{-T/2}(\boldsymbol{\xi} - \mathbf{m}) \,, \tag{3.32}$$

produces new variables $\mathbf{x}$ of zero mean, and diagonal covariance, that is, a probability density

$$p_{x_1,\dots,x_N} = \frac{\exp -\frac{1}{2}(X_1^2 + \cdots X_N^2)}{(2\pi)^{N/2}} = \frac{\exp\left(-\frac{1}{2}X_1^2\right)}{(2\pi)^{1/2}} \cdots \frac{\exp\left(-\frac{1}{2}X_N^2\right)}{(2\pi)^{1/2}} \,, \tag{3.33}$$

which factors into N-independent, normal variates of zero mean and unit variance ($\mathbf{C}_{xx} = \mathbf{R}_{xx} = \mathbf{I}$). Such a process is often called Gaussian"white noise," and has the property $\langle x_i x_j \rangle = 0,\ i \neq j$.[15]

3.4. Functions and Sums of Random Variables. If the probability density of x is $p_x(x)$, then the mean of a function of x, $g(x)$ is just,

$$\langle g(x) \rangle = \int_{-\infty}^{\infty} g(X) p_x(X) dX \,, \tag{3.34}$$

which follows from the definition of the probability density as the limit of the outcome of a number of trials. The probability density for g regarded as a new random variable is given by (3.19) as,

$$p_g(G) = p_x(X(G)) \frac{dx}{dg} dG, \tag{3.35}$$

[14] Numerical schemes for finding $\mathbf{C}_{\xi\xi}^{1/2}$ are described by Lawson and Hanson (1976) and Golub and Van Loan (1989)

[15] Cramér (1946) discusses what happens when the determinant of $\mathbf{C}_{\xi\xi}$ vanishes, that is, if $\mathbf{C}_{\xi\xi}$ is singular.

where the Jacobian is just dx/dg in one-dimension.

An important special case is $g = x^2$ where x is Gaussian of zero mean and unit variance (any Gaussian variable z, of mean m and variance σ^2 can be transformed to one of zero mean and unit variance by the transformation,

$$x = \frac{z-m}{\sigma}\,.$$

Then the probability density of g is,

$$p_g(G) = \frac{1}{G^{1/2}\sqrt{2\pi}}\, \exp(-G/2)\,, \tag{3.36}$$

a special probability density usually denoted as χ_1^2 ("chi-square-sub-1").

Sums of Random Variables

It is often helpful to be able to compute the probability density of sums of independent random variables. The procedure for doing so is based upon (3.34). Let x be a random variable and consider the expected value of the function e^{ixt}:

$$\langle e^{ixt}\rangle = \int_{-\infty}^{\infty} p_x\left(X\right) e^{iXt} dX \equiv \phi_x(t)\,, \tag{3.37}$$

which is the Fourier transform of $p_x(X)$; $\phi_x(t)$ is usually termed the "characteristic function" of x. Now consider the sum of two independent random variables x, y with probability densities p_x, p_y, respectively, and define a new random variable $z = x+y$. What is the probability density of z? One starts by first determining the characteristic function, $\phi_z(t)$ for z and then using the Fourier inversion theorem to obtain $p_x(Z)$. To obtain ϕ_z,

$$\phi_z(t) = \langle e^{izt}\rangle = \langle e^{i(x+y)t}\rangle = \langle e^{ixt}\rangle\langle e^{iyt}\rangle$$

where the last step depends upon the independence assumption. This last equation shows

$$\phi_z(t) = \phi_x(t)\phi_y(t)\,. \tag{3.38}$$

That is, the characteristic function for a sum of two independent variables is the product of the characteristic functions. The "convolution theorem"[16] asserts that the Fourier transform (forward or inverse) of a product of two functions is the convolution of the Fourier transforms of the two functions. That is,

$$p_z\left(Z\right) = \int_{-\infty}^{\infty} p_x\left(r\right) p_y\left(Z-r\right) dr. \tag{3.39}$$

We will not explore this relation in any detail, leaving the reader to pursue the subject in the references.[17]. But it follows immediately that the multiplication of the characteristic functions of

[16] Bracewell (1978).

[17] Cramér (1946).

a sum of independent Gaussian variables produces a new variable, which is also Gaussian, with a mean equal to the sum of the means and a variance which is the sum of the variances ("sums of Gaussians are Gaussian"). It also follows immediately from Eq. (3.39) that if a variable ξ is defined as,

$$\xi = x_1^2 + x_2^2 + \cdots + x_v^2, \tag{3.40}$$

where each x_i is Gaussian of zero mean and unit variance, that the probability density for ξ is,

$$p_\xi(\Xi) = \frac{\Xi^{\nu/2-1}}{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)} \exp(-\Xi/2)\,, \tag{3.41}$$

known as χ_v^2 or "chi-square with ν degrees of freedom." The chi-square probability density is central to the discussion of the expected sizes of vectors, such as $\tilde{\mathbf{n}}$, measured as $\tilde{\mathbf{n}}^T\tilde{\mathbf{n}} = \|\tilde{\mathbf{n}}\|^2 = \sum_i \tilde{n}_i^2$ if the elements of $\tilde{\mathbf{n}}$ can be assumed to be independent and Gaussian. Eq. (3.36) is the special case $\nu = 1$.

Degrees-of-Freedom

The number of independent variables described by a probability density is usually called the "number of degrees-of-freedom." Thus the densities in (3.30) and (3.33) have N-degrees of freedom and (3.41) has ν of them. If a sample average (3.3) is formed, it is said to have N-degrees of freedom if each of the x_j is independent. But what if the x_j have a covariance $\mathbf{C}_{xx}$ which is non-diagonal? This question of how to interpret averages of correlated variables will be explicitly discussed on P. **??**.

Consider the special case of the sample variance Eq. (3.4)—which we claimed was biassed. The reason is that even if the sample values, x_i, are independent, the presence of the sample average in the sample variance means that there are only $N-1$ independent terms in the sum. That this is so is most readily seen by examining the two-term case. Two samples produce a sample mean, $\langle x\rangle_2 = (x_1+x_2)/2$. The two-term sample variance is,

$$s^2 = \tfrac{1}{2}\left[(x_1 - \langle x\rangle_2)^2 + (x_2 - \langle x\rangle_2)^2\right]\,,$$

but knowledge of x_1 and of the sample average, permits perfect prediction of x_2. Thus the second term in the sample variance as written is not independent of the first term, and thus there is just one independent piece of information in the two-term sample variance. To show it in general, assume without loss of generality that $\langle x\rangle = 0$, so that $\sigma^2 = \left\langle x^2\right\rangle$. The sample variance about the sample mean (which will not vanish) of independent samples is given by Eq.

(3.4) and so,

$$\begin{aligned}
\left\langle s^2\right\rangle &= \frac{1}{M}\sum_{i=1}^{M}\left\langle\left(x_i-\frac{1}{M}\sum_{j=1}^{M}x_j\right)\left(x_i-\frac{1}{M}\sum_{p=1}^{M}x_p\right)\right\rangle \\
&= \frac{1}{M}\sum_{i=1}^{M}\left\{\left\langle x_i^2\right\rangle-\frac{1}{M}\sum_{j=1}^{M}\left\langle x_ix_j\right\rangle-\frac{1}{M}\sum_{p=1}^{M}\left\langle x_ix_p\right\rangle+\frac{1}{M^2}\sum_{j=1}^{M}\sum_{p=1}^{M}\left\langle x_jx_p\right\rangle\right\} \\
&= \frac{1}{M}\sum_{i=1}^{M}\left\{\sigma^2-\frac{\sigma^2}{M}\sum_{j}\delta_{ij}-\frac{\sigma^2}{M}\sum_{p}\delta_{ip}+\frac{\sigma^2}{M^2}\sum_{j}\sum_{p}\delta_{jp}\right\} \\
&= \frac{\sigma^2\left(M-1\right)}{M}\neq\sigma^2
\end{aligned}$$

Stationarity

Consider a vector random variable, with element x_i where the subscript i denotes a position in time or space. Then x_i, x_j are two different random variables—for example, the temperature at two different positions in the ocean, or the temperature at two different times at the same position. If the physics governing these two different random variables are independent of the parameter i (i.e., independent of time or space), then x_i is said to be "stationary"—meaning that all the underlying statistics are independent of i. Specifically, $\langle x_i\rangle = \langle x_j\rangle \equiv \langle x\rangle$, $D^2(x_i) = D^2(x_j) = D^2\left(x\right)$, etc. Furthermore, x_i, x_j have a covariance

$$C_{xx}(i,j) = \langle(x_i-\langle x_i\rangle)(x_j-\langle x_j\rangle)\rangle = C_{xx}(|i-j|)\,, \tag{3.42}$$

that is, independent of i, j, and depending only upon the difference $|i-j|$. The distance, $|i-j|$, is often called the "lag." $C_{xx}(|i-j|)$ is called the "autocovariance" of $\mathbf{x}$ or just the covariance, because we now regard x_i, x_j as intrinsically the same process.[18] If C_{xx} does not vanish, then by the discussion above, knowledge of the numerical value of x_i implies some predictive skill for x_j and vice-versa—a result of great importance when we examine map-making and objective analysis. For stationary processes, all moments having the same $|i-j|$ are identical; it is seen that all diagonals of such a matrix $\{C_{xx}\left(i,j\right)\}$, are constant, for example, $\mathbf{C}_{\xi\xi}$ in Eq. (3.30). Matrices with constant diagonals are thus defined by the vector $C_{xx}(|i-j|)$, and are said to have a "Toeplitz form."

[18] If the means and variances are independent of i, j and the first cross-moment is dependent only upon $|i-j|$, the process x is said to be stationary in the "wide-sense." If all higher moments also depend only on $|i-j|$, the process is said to be stationary in the "strict-sense," or more simply, just stationary. A Gaussian process has the unusual property that wide-sense stationarity implies strict-sense stationarity.

4. Least-Squares

Much of what follows in this book can be described using very elegant and powerful mathematical tools. On the other hand, by restricting ourselves to discrete models and finite numbers of measurements (what goes into a computer), almost everything can also be viewed as a form of ordinary least-squares, providing a much more intuitive approach than one through functional analysis. It is thus useful to go back and review what "everyone knows" about this most-familiar of all approximation methods.

4.1. Basic Formulation. Consider the elementary problem motivated by the "data" shown in figure 2. t is supposed to be an independent variable, which could be time, or a spatial coordinate or just an index. Some physical variable, call it $\theta(t)$, perhaps temperature at a point in the ocean, has been measured at coordinates $t = t_i$, $1 \leq i \leq M$, as depicted in the figure.

We have reason to believe that there is a linear relationship between $\theta(t)$ and t in the form $\theta(t) = a + bt$, so that the measurements are,

$$y(t_i) = \theta(t_i) + n(t_i) = a + bt_i + n(t_i), \tag{4.1}$$

where $n(t)$ is the inevitable measurement noise. The straight line relationship might as well be referred to as a "model" as it represents our present conception of the data structure. We want to determine a, b.

The set of observations can be written in the general standard form,

$$\mathbf{Ex} + \mathbf{n} = \mathbf{y} \tag{4.2}$$

where in the present special case,

$$\mathbf{E} = \left\{\begin{matrix} 1 & t_1 \\ 1 & t_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & t_M \end{matrix}\right\}, \qquad \mathbf{x} = \begin{bmatrix} a \\ b \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y(t_1) \\ y(t_2) \\ \cdot \\ \cdot \\ y(t_M) \end{bmatrix}, \qquad \mathbf{n} = \begin{bmatrix} n(t_1) \\ n(t_2) \\ \cdot \\ \cdot \\ n(t_M) \end{bmatrix}. \tag{4.3}$$

Equation sets like (4.2) are seen in many practical situations, including the ones described in Chapter 2. The matrix $\mathbf{E}$ in general represents arbitrarily complicated (linear) relations between the parameters $\mathbf{x}$, and the observations $\mathbf{y}$. In some real cases, it has many thousands of rows and columns. Its construction involves specifying what those relations are, and in a very general sense, it requires a "model" of the data set. Unfortunately, the term "model" is used in a variety of other ways in this context, including statistical assumptions, and often for auxiliary relationships among the elements of $\mathbf{x}$ which are independent of those contained in $\mathbf{E}$. To separate these difference usages, we will sometimes append various adjectives to the use ("statistical model", "exact relationships" etc.).

One sometimes sees (4.2) written as

$$\mathbf{Ex} \sim \mathbf{y}$$

or even

$$\mathbf{Ex} = \mathbf{y}\,.$$

But Eq. (4.2) is preferable, because it explicitly recognizes that $\mathbf{n} = \mathbf{0}$ is exceptional. Sometimes, by happenstance or arrangement, one finds $M = N$ and that $\mathbf{E}$ has an inverse. But the obvious solution, $\mathbf{x} = \mathbf{E}^{-1}\mathbf{y}$, leads to the conclusion, $\mathbf{n} = \mathbf{0}$, which should be unacceptable if the $\mathbf{y}$ are the result of measurements. We will need to return to this case, but for now, let us consider the conventional problem where $M > N$.

Commonly, then, one sees a "best possible" solution—defined as producing the smallest possible value of $\mathbf{n}^T\mathbf{n}$, that is the minimum of

$$J = \sum_{i=1}^{M} n_i^2 = \mathbf{n}^T\mathbf{n} = (\mathbf{y} - \mathbf{Ex})^T(\mathbf{y} - \mathbf{Ex})\,. \tag{4.4}$$

(Whether the smallest noise solution really is the best one is something that must be seriously considered later.) In the special case of the straight line model,

$$J = \sum_{i=1}^{M} (y_i - a - bt_i)^2\,. \tag{4.5}$$

J is an example of what is called an "objective" or "cost" function.[19]

Differentiating 4.5 with respect to a, b or $\mathbf{x}$ (using (2.32)) and by setting,

$$\begin{aligned} dJ &= \frac{\partial J}{\partial x_i}dx_i = \left(\frac{\partial J}{\partial \mathbf{x}}\right)^T d\mathbf{x} \\ &= 2\left(\mathbf{E}^T\mathbf{y} - \mathbf{E}^T\mathbf{Ex}\right)^T d\mathbf{x} = 0. \end{aligned} \tag{4.6}$$

This equation is of the form

$$dJ = \sum a_i dx_i = 0. \tag{4.7}$$

It is an elementary result of multivariable calculus that an extreme value (here a minimum) of J is found where $dJ = 0$. Because the x_i are free to vary independently, dJ will vanish only if the coefficients of the dx_i are separately zero or,

$$\mathbf{E}^T\mathbf{y} - \mathbf{E}^T\mathbf{Ex} = \mathbf{0}.$$

That is,

$$\mathbf{E}^T\mathbf{Ex} = \mathbf{E}^T\mathbf{y}, \tag{4.8}$$

[19] The terminology "least-squares" is reserved in this book, conventionally, for the minimization of discrete sums such as Eq. (4.4). This usage contrasts with that of Bennett (2002) who applies it to continuous integrals, such as, $\int_a^b (u\,(q) - r\,(q))^2\,dq$ leading to the calculus of variations and Euler-Lagrange equations.

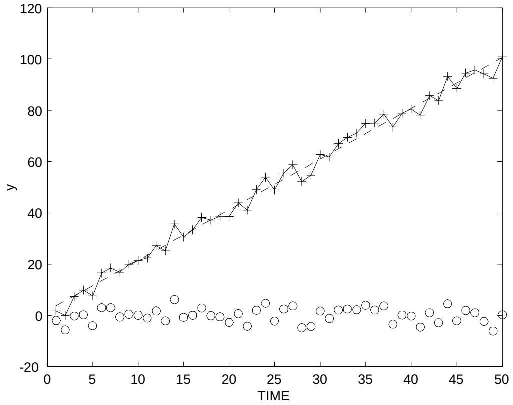


Figure 2. "Data" generated through the rule $y = 1 + 2t + n_t$, where $\langle n_t \rangle = 0$, $\langle n_i n_j \rangle = 9\delta_{ij}$ shown as + connected by the solid line. Dashed line is the simple least-squares fit, $\tilde{y} = 1.69 \pm 0.83 + (1.98 \pm 0.03)\, t$. Residuals are plotted as open circles, and at least visually, show no obvious structure. Note that the fit is correct within its estimated standard errors. The sample variance of the estimated noise was used for calculating the uncertainty, not the theoretical value.

called the "normal equations." Making the sometimes-valid-assumption that $(\mathbf{E}^T\mathbf{E})^{-1}$ exists,

$$\tilde{\mathbf{x}} = (\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\mathbf{y}\,. \tag{4.9}$$

By looking at the second derivatives of J with respect to $\mathbf{x}$, we could show what is intuitively clear—that we have a minimum and not a maximum.

The solution is written as $\tilde{\mathbf{x}}$ rather than as $\mathbf{x}$ because the relationship between (4.9) and the "correct" value is not clear. The fit is displayed in Fig. 2, as are the residuals,

$$\tilde{\mathbf{n}} = \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}} = (\mathbf{I} - \mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T)\mathbf{y}\,. \tag{4.10}$$

That is, the M equations have been used to estimate N values, $\tilde{\mathbf{x}}_i$, and M values $\tilde{\mathbf{n}}_i$, or $M + N$ altogether. The combination

$$\mathbf{H} = \mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T \tag{4.11}$$

occurs sufficiently often that it is worth a special symbol. Note the "idempotent" property $\mathbf{H}^2 = \mathbf{H}$. If the solution $\tilde{\mathbf{x}}$ is substituted into the original equations, the result is,

$$\mathbf{E}\tilde{\mathbf{x}} = \mathbf{H}\mathbf{y} = \tilde{\mathbf{y}}, \tag{4.12}$$

and $\tilde{\mathbf{n}}^T\tilde{\mathbf{y}} = \left[(\mathbf{I} - \mathbf{H})\,\mathbf{y}\right]^T \mathbf{H}\mathbf{y} = \mathbf{0}$, and the residuals are orthogonal (normal) to the inferred noise-free "data" $\tilde{\mathbf{y}}$. This result explains the label "normal equations" used to describe the solution of a least-squares problem.

All this is easy and familiar and applies to any set of simultaneous linear equations, not just the straight-line example. Before proceeding, let us apply some of the statistical machinery

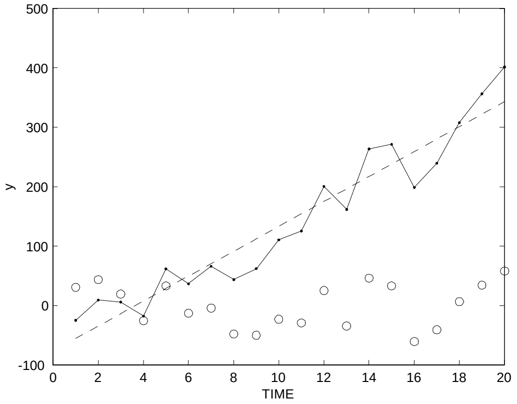


FIGURE 3. Here the "data" were generated from a quadratic rule, $y = 1 + t^2 + n(t)$, $\langle n^2 \rangle = 900$. Note that only the first $1 \le t \le 20$ data points are used. An incorrect straight line fit was used resulting in $\tilde{y} = (-76.3 \pm 17.3) + (20.98 \pm 1.4)\, t$, which is incorrect, but the residuals at least visually, do not appear unacceptable. At this point some might be inclined to claim the model has been "verified," or "validated."

to understanding (4.9). Notice that no statistics were used in obtaining (4.9), but we can nonetheless ask the extent to which this value for $\tilde{\mathbf{x}}$ is affected by the random elements: the noise in $\mathbf{y}$. Let $\mathbf{y}_0$ be the value of $\mathbf{y}$ that would be obtained in the hypothetical situation for which $\mathbf{n} = \mathbf{0}$. Assume further that $\langle \mathbf{n} \rangle = \mathbf{0}$ and that $\mathbf{R}_{nn} = \mathbf{C}_{nn} = \langle \mathbf{n}\mathbf{n}^T \rangle$ is known. Then the expected value of $\tilde{\mathbf{x}}$ is

$$\langle \tilde{\mathbf{x}} \rangle = (\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\mathbf{y}_0\,. \tag{4.13}$$

If the matrix inverse exists, then in many situations, including the problem of fitting a straight line to data, perfect observations would produce the correct answer, and Eq. (4.9) provides an unbiassed estimate of the true solution, $\langle \tilde{\mathbf{x}} \rangle = \langle \mathbf{x} \rangle$.

On the other hand, if the data were actually produced from physics governed for example, by a quadratic rule, $\theta(t) = a + ct^2$, then fitting the linear rule to such observations, even if they are perfect, could never produce the right answer and the solution would be biassed. An example of such a fit is shown in figures 3, 4. Such errors are conceptually distinguishable from the noise of observation, and are properly labeled "model errors." Assume however, that the correct model is being used, and therefore that $\langle \tilde{\mathbf{x}} \rangle = \mathbf{x}$. Then the uncertainty of the solution is,

$$\begin{aligned}\mathbf{P} = \mathbf{C}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} &= \langle (\tilde{\mathbf{x}} - \mathbf{x})(\tilde{\mathbf{x}} - \mathbf{x})^T \rangle \\ &= (\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T \langle \mathbf{n}\mathbf{n}^T \rangle\, \mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1} \\ &= (\mathbf{E}^T\mathbf{E})^{-1}\, \mathbf{E}^T\, \mathbf{R}_{nn}\, \mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\,.\end{aligned} \tag{4.14}$$

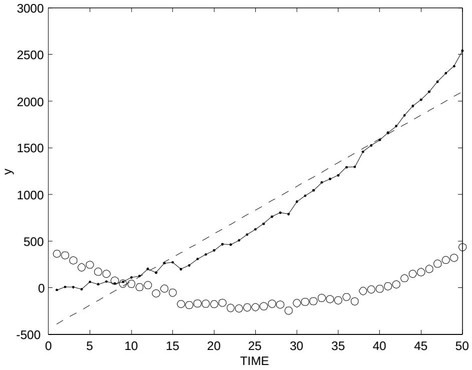


FIGURE 4. The situation as in Fig. 3, except the series was run out to 50 points. Now the residuals ('o') are visually structured, and one would have a very powerful suggestion that some hypothesis (something about the model) is not correct. This straightline fit should be rejected as being inconsistent with the assumption that the residuals are unstructured. Now the model has been "invalidated." (See caption to Fig. 3.)

In the special case, $\mathbf{R} = \sigma_n^2 \mathbf{I}$, that is, no correlation between the noise in different equations (white noise), then (4.14) simplifies to,

$$\mathbf{P} = \sigma_n^2 (\mathbf{E}^T \mathbf{E})^{-1} \,. \tag{4.15}$$

If we are not confident that $\langle \tilde{\mathbf{x}} \rangle = \mathbf{x}$, perhaps because of doubts about the straight line model, Eqs. (4.14)–(4.15) are still interpretable, but as $C_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} = D^2(\tilde{\mathbf{x}} - \langle \tilde{\mathbf{x}} \rangle) -$ the covariance of $\tilde{\mathbf{x}}$. The "standard error" of $\tilde{x}_i$ is usually defined to be $\pm\sqrt{C_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}_{ii}}}$ and is used to understand the adequacy of data for distinguishing different possible estimates of $\tilde{\mathbf{x}}$. If applied to the straight line fit of fig. 2, we obtain an estimate, $\tilde{\mathbf{x}}^T = [\tilde{a}, \tilde{b}] = [1.69 \pm 0.83, 1.98 \pm 0.03]$, which are within one standard deviation of the true values, $[a, b] = [1, 2]$. If the noise in $\mathbf{y}$ is Gaussian, it follows that the probability density of $\tilde{\mathbf{x}}$ is also Gaussian, with mean $\langle \tilde{\mathbf{x}} \rangle$ and covariance $\mathbf{C}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$. Of course, if $\mathbf{n}$ is not Gaussian, then the estimate won't be either, and one must be wary of the utility of the standard errors. A Gaussian or other assumption should be regarded as part of the model definition. One can estimate the uncertainty of the residuals as,

$$\begin{aligned} \mathbf{C}_{nn} &= \langle (\tilde{\mathbf{n}} - \langle \tilde{\mathbf{n}} \rangle) \, (\tilde{\mathbf{n}} - \langle \tilde{\mathbf{n}} \rangle) \rangle = (\mathbf{I} - \mathbf{H}) \, \mathbf{R}_{nn} \, (\mathbf{I} - \mathbf{H})^T \\ &= \sigma_n^2 \, (\mathbf{I} - \mathbf{H})^2 = \sigma_n^2 \, (\mathbf{I} - \mathbf{H}) \end{aligned} \tag{4.16}$$

where zero-mean white noise was assumed, and $\mathbf{H}$ was defined in Eq. (4.11). Notice that the true noise, $\mathbf{n}$, was assumed to be white, but that the estimated noise, $\tilde{\mathbf{n}}$, has a non-diagonal covariance and so in a formal sense does not have the expected covariance. We return to this point below.

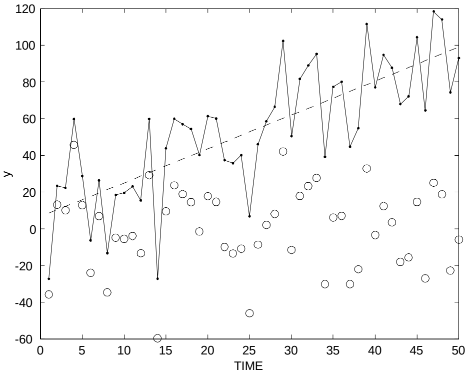


FIGURE 5. The same situation as in Fig. 2, $y = 1 + 2t$, except $\left\langle n^2 \right\rangle = 900$ to give very noisy data. Now the best fitting straight line is $y = (6.62 \pm 6.50) + (1.85 \pm 0.22)\, t$ which includes the correct answer within one standard error. Note however, that the intercept is indistinguishable from zero.

The fit of a straight line to observations demonstrates many of the issues involved in making inferences from real, noisy data that appear in more complex situations. In figure 5, the correct model used to generate the data was the same as in fig. 2, but the noise level is very high. The parameters $[\tilde{a}, \tilde{b}]$ are numerically inexact, but consistent within one standard error with the correct values, which is all one can hope for.

In figure 3, a quadratic model $y = 1 + t^2 + n\,(t)$ was used to generate the numbers, with $\left\langle n^2 \right\rangle = 900$. Using only the first 20 points, and fitting an incorrect model produces a reasonable straight line fit to the data as shown. Modeling a quadratic field with a linear model produces a systematic or "model" error, which is not easy to detect here. One sometimes hears it said that "least-squares failed" in situations such as this one. But this conclusion shows a fundamental misunderstanding: least-squares did exactly what it was asked to do—to produce the best-fitting straight-line to the data, under the assumption that the noise in the data had no structure. Here, one might conclude that "the straight-line fit *is* consistent with the data." Such a conclusion is completely different from asserting that one has proven a straight-line fit correctly "explains" the data or, in modeler's jargon, that the model has been "verified" or "validated." If the outcome of the fit were sufficiently important, one might try more powerful tests on the $\tilde{n}_i$ than a mere visual test. Such tests might lead to rejection of the straight-line hypothesis; but even if the tests are passed, the model has *never* been verified: it has only been shown to be consistent with the available data.

If the situation remains unsatisfactory (perhaps one suspects the model is inadequate, but there are not enough data to produce sufficiently powerful tests), it can be very frustrating. But sometimes the only remedy is to obtain more data. So in Fig. 4, the data length was extended

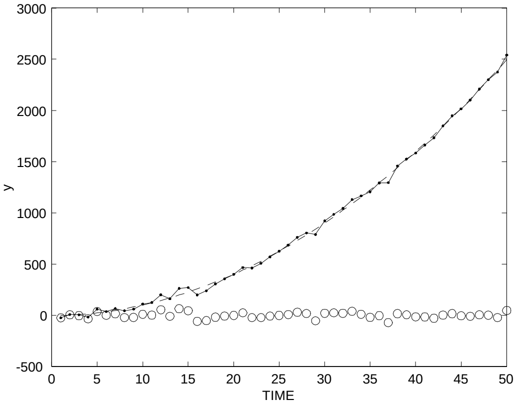


FIGURE 6. Same as Fig. 4, except a more complete model, $y = a + bt + ct^2$ was used, and which gives acceptable residuals.

to 50 points. Now, even visually, the $\tilde{n}_i$ are obviously structured, and one would almost surely reject any hypothesis that a straight-line was an adequate representation of the data. *The model has been invalidated.* If one fits a quadratic rule, $y = a + bt + ct^2$, a perfectly acceptable solution is found; see Fig. 6.

One must always confirm, after the fact, that J, which is a direct function of the residuals, behaves as expected when the solution is substituted. In particular, its expected value,

$$\langle J \rangle = \sum_i^M \left\langle n_i^2 \right\rangle = M - N, \tag{4.17}$$

assuming that the n_i have been scaled so that each has an expected value $\left\langle n_i^2 \right\rangle = 1$. That there are only $M - N$ independent terms in (4.17) follows from the N supposed-independent constraints linking the variables. For any particular solution, $\tilde{\mathbf{x}}, \tilde{\mathbf{n}}$, J will be a random variable, whose expectation is (4.17). Assuming the n_i are at least approximately Gaussian, J itself is the sum of $M - N$ independent χ_1^2 variables, and is therefore distributed in χ_{M-N}^2. One can and should make histograms of the individual n_i^2 to check them against the expected χ_1^2 probability density. This type of argument leads to the large literature on hypothesis testing.

As an illustration of the random behavior of residuals, 30 equationsin 15 unknowns, $\mathbf{Ex} + \mathbf{n} = \mathbf{y}$ were constructed, such that $\mathbf{E}^T\mathbf{E}$ was non-singular. Fifty different values of $\mathbf{y}$ were then constructed by generating 50 separate $\mathbf{n}$ using a pseudo-random number generator. An ensemble of 50 different solutions were then calculated using (4.9), producing 1500 separate values of $\tilde{n}_i^2$. These are plotted in Fig. 7 and compared to χ_1^2. The corresponding value, $\tilde{J}^{(p)} = \sum \tilde{n}_i^2$, was found for each set of equations, and also plotted. A corresponding frequency function for $\tilde{J}^{(p)}$ is compared in Fig. 7 to χ_{15}^2, with reasonably good results. The empirical mean value of all $\tilde{J}_i$ is 14.3. (The main inference here is that any particular solution may, completely correctly, produce

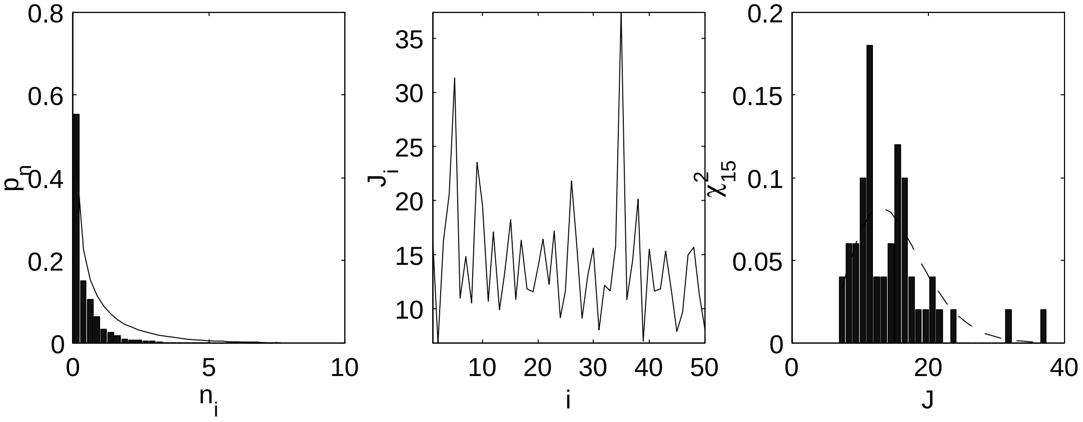


FIGURE 7. χ_1^2 probability density (left panel), and the empirical frequency function of *all* residuals, $\tilde{n}_i^2$ from 50 separate experiments for simple least-squares solution of $\mathbf{Ex}+\mathbf{n}=\mathbf{y}$. There is at least rough agreement between the theoretical and calculated frequency functions. Middle panel displays the 50 values of J_i computed from the same experiments in the left panel. Right panel displays the empirical frequency function for the J_i as compared to the theoretical value of χ_{15}^2, (dashed line). Tests exist, not discussed here, of the hypothesis that the calculated J_i are consistent with the theoretical distribution.

individual residuals $\tilde{n}_i^2$ differing considerably from the mean of $\left\langle \chi_1^2 \right\rangle = 1$, and similarly, their sums, the $J^{(p)}$ may differ greatly from $\left\langle \chi_{15}^2 \right\rangle = 15$. But one can readily calculate the probability of finding a much larger or smaller value, and employ it to help evaluate the possibility that one has used an incorrect model.

Visual tests for randomness of residuals have obvious limitations, and elaborate statistical tests in addition to the comparison with χ_1^2 exist to help determine objectively whether one should accept or reject the hypothesis that no significant structure remains in a sequence of numbers. Books on regression analysis[20] should be consulted for general methodologies. As an indication of what can be done, figure 8 shows the "sample autocorrelation,"

$$\tilde{\phi}_{nn}(\tau) = \frac{\frac{1}{M}\sum_{i=1}^{M-|\tau|} \tilde{n}_i \tilde{n}_{i+\tau}}{\frac{1}{M}\sum_{i=1}^{M} \tilde{n}_i^2}\,, \tag{4.18}$$

for the residuals of the fits shown in figs. 4, 6 is displayed. For white noise, it is reasonably obvious that,

$$\left\langle \tilde{\phi}\left(\tau\right) \right\rangle = \delta_{0\tau}, \tag{4.19}$$

[20] Seber (1977) or Box et al. (1994) or Draper and Smith (1981) are all good starting points.

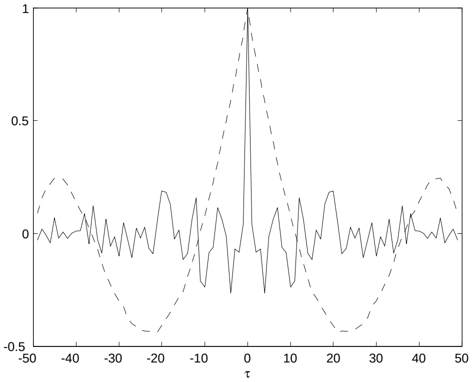


FIGURE 8. Autocorrelations of the estimated residuals in Figs. 4 (dashed line), and 6 (solid). The latter is indistinguishable, by statistical test, from a delta function at the origin, and so the residuals are not distinguishable from white noise.

and deviations of the estimated $\tilde{\phi}(t)$ from Eq. (4.19) can be used in simple tests. The adequate fit (Fig. 6) produces an autocorrelation of the residuals indistinguishable from a delta function at the origin, while the inadequate fit, shows a great deal of structure which would lead to the conclusion that the residuals are too different from white noise to be acceptable. (Not all cases are this obvious.).

As already pointed out, the residuals of the least-squares fit cannot be expected to be precisely white noise. Because there are M relationships among the parameters of the problem (M-equations), and the number of $\tilde{\mathbf{x}}$ elements determined is N, there are $M-N$-degrees of freedom in the determination of $\tilde{\mathbf{n}}$ and structures are imposed upon them. The failure, for this reason, of $\tilde{\mathbf{n}}$ strictly to be white noise, is generally only an issue in practice when $M-N$ becomes small compared to M.[21]

4.2. Weighted and Tapered Least-Squares. The least-squares solution (4.9)–(4.10) was derived by minimizing the objective function (4.4), in which each residual element is given equal weight. An important feature of least-squares is that we can give whatever emphasis we please to minimizing individual equation residuals, for example, by introducing an objective function,

$$J = \sum_i W_{ii}^{-1} n_i^2, \tag{4.20}$$

where W_{ii} are any numbers desired. The choice $W_{ii} = 1$, as used above, might be reasonable, but it is clearly an arbitrary one which without further justification does not produce a solution with any special claim to significance. In the least-squares context, we are free to make any

[21] Draper and Smith (1981, Chapter 3) and the references given there.

other reasonable choice, including demanding that some residuals should be much smaller than others—perhaps just to see if it is possible.

A general formalism is obtained by defining a diagonal weight matrix, $W = \mathrm{diag}(W_{ii})$. Divide each equation by $\sqrt{W_{ii}}$,

$$W_{ii}^{-T/2} \sum_i E_{ij} x_j + W_{ii}^{-T/2} n_i = W_{ii}^{-T/2} y_i \tag{4.21}$$

or

$$\begin{gathered} \mathbf{E}'\mathbf{x} + \mathbf{n}' = \mathbf{y}' \\ \mathbf{E}' = \mathbf{W}^{-T/2}\mathbf{E}, \quad \mathbf{n}' = \mathbf{W}^{-T/2}\mathbf{n}, \quad \mathbf{y}' = \mathbf{W}^{-T/2}\mathbf{y} \end{gathered} \tag{4.22}$$

where we used the fact that the square root of a diagonal matrix is the diagonal matrix of element-by-element square roots. Such a matrix is its own transpose. The operation in (4.21) or (4.22) is usually called "row-scaling" because it operates on the rows of $\mathbf{E}$ (as well as on $\mathbf{n}$, $\mathbf{y}$).

For the new equations (4.22), the objective function,

$$\begin{aligned} J &= \mathbf{n}'^T\mathbf{n}' = (\mathbf{y}' - \mathbf{E}'\mathbf{x})^T(\mathbf{y}' - \mathbf{E}'\mathbf{x}) \\ &= \mathbf{n}^T\mathbf{W}^{-1}\mathbf{n} = (\mathbf{y} - \mathbf{E}\mathbf{x})^T\mathbf{W}^{-1}(\mathbf{y} - \mathbf{E}\mathbf{x}) \end{aligned} \tag{4.23}$$

weights the residuals as desired. If, for some reason, $\mathbf{W}$ is non-diagonal, but symmetric and positive-definite, then it has a Cholesky decomposition, (see P. 35) and,

$$\mathbf{W} = \mathbf{W}^{T/2}\mathbf{W}^{1/2},$$

and (4.22) remains valid more generally.

The values $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$, minimizing (4.23) are,

$$\begin{aligned} \tilde{\mathbf{x}} &= (\mathbf{E}'^T\mathbf{E}')^{-1}\mathbf{E}'^T\mathbf{y}' = (\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E})^{-1}\mathbf{E}^T\mathbf{W}^{-1}\mathbf{y} \\ \tilde{\mathbf{n}} &= \mathbf{W}^{T/2}\mathbf{n}' = \left\{\mathbf{I} - \mathbf{E}(\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E})^{-1}\mathbf{E}^T\mathbf{W}^{-1}\right\}\mathbf{y} \end{aligned} \tag{4.24}$$

$$\mathbf{C}_{xx} = (\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E})^{-1}\mathbf{E}^T\mathbf{W}^{-1}\mathbf{R}_{nn}\mathbf{W}^{-1}\mathbf{E}(\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E})^{-1}. \tag{4.25}$$

Uniform diagonal weights are clearly a special case. The rationale for choosing differing diagonal weights, or a non-diagonal $\mathbf{W}$, is probably not very obvious to the reader. Often one chooses $\mathbf{W} = \mathbf{R}_{nn} = \{\langle n_i n_j \rangle\}$, that is, the weight matrix is chosen to be the expected second moment matrix of the residuals. Then

$$\left\langle \mathbf{n}'\mathbf{n}'^T \right\rangle = \mathbf{I},$$

and Eq. (4.25) simplifies to

$$\mathbf{C}_{xx} = (\mathbf{E}^T\mathbf{R}_{nn}^{-1}\mathbf{E})^{-1}. \tag{4.26}$$

In this special case, the weighting (4.22) has a ready interpretation: The equations (and hence the residuals) are rotated and stretched so that in the new coordinate system of n_i, the covariances are all diagonal and the variances are all unity. Under these circumstances, an objective function

$$J = \sum_i n_i'^2$$

as used in the original form of least-squares (Eq. (4.4)) is eminently reasonable.

But we emphasize that this choice of $\mathbf{W}$ is a very special one and has confused many users of inverse methods. To emphasize again: Least-squares is an approximation procedure in which $\mathbf{W}$ is a set of weights wholly at the disposal of the investigator; setting $\mathbf{W} = \mathbf{R}_{nn}$ is a special case whose significance is best understood after we examine a different, statistical, estimation procedure.

Whether the equations are scaled or not, the previous limitations of the simple least-squares solutions remain. In particular, we still have the problem that the solution may produce elements in $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$, whose relative values are not in accord with expected or reasonable behavior and the solution uncertainty or variances could be unusably large, as the solution is determined, mechanically, and automatically, from combinations such as $(\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E})^{-1}$. Operators like these are neither controllable nor very easy to understand; if any of the the matrices is singular, they won't even exist.

It was long ago recognized that some control over the magnitudes of $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$, $\mathbf{C}_{xx}$ could be obtained in the simple least-squares context by modifying the objective function (4.4) to have an additional term:

$$\begin{aligned} J' &= \mathbf{n}^T\mathbf{W}^{-1}\mathbf{n} + \alpha^2\mathbf{x}^T\mathbf{x} \qquad (4.27)\\ &= (\mathbf{y} - \mathbf{E}\mathbf{x})^T\mathbf{W}^{-1}(\mathbf{y} - \mathbf{E}\mathbf{x}) + \alpha^2\mathbf{x}^T\mathbf{x}, \end{aligned}$$

in which α^2 is a positive constant.

If the minimum of (4.27) is sought by setting the derivatives with respect to $\mathbf{x}$ to zero, we obtain,

$$\tilde{\mathbf{x}} = \left(\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E}+\alpha^2\mathbf{I}\right)^{-1}\mathbf{E}^T\mathbf{W}^{-1}\mathbf{y} \qquad (4.28)$$

$$\tilde{\mathbf{n}} = \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}} \qquad (4.29)$$

$$\mathbf{C}_{xx} = \qquad (4.30)$$

$$\left(\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E}+\alpha^2\mathbf{I}\right)^{-1}\mathbf{E}^T\mathbf{W}^{-1}\mathbf{R}_{nn}\mathbf{W}^{-1}\mathbf{E}\left(\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E}+\alpha^2\mathbf{I}\right)^{-1}.$$

(Eq. (4.30) simplifies a bit if $\mathbf{W}=\mathbf{R}_{nn}$.) By letting $\alpha^2\to 0$, the solution 4.24, 4.25 is recovered, and if $\alpha^2\to\infty$, $\|\tilde{\mathbf{x}}\|_2\to 0$, $\tilde{\mathbf{n}}\to\mathbf{y}$; α^2 is called a "trade-off parameter," because it trades the magnitude of $\tilde{\mathbf{x}}$ against that of $\tilde{\mathbf{n}}$. By varying the size of α^2 we gain some influence over the norm of the residuals relative to that of $\tilde{\mathbf{x}}$. The expected value of $\tilde{\mathbf{x}}$ is now,

$$\langle\tilde{\mathbf{x}}\rangle=\left[\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E}+\alpha^2\mathbf{I}\right]^{-1}\mathbf{E}^T\mathbf{W}^{-1}\mathbf{y}_0\,. \tag{4.31}$$

If the true solution is believed to be Eq. (4.13), then this new solution is biassed . But the variance of $\tilde{\mathbf{x}}$ has been reduced, (4.30), by introduction of $\alpha^2>0$—that is, the acceptance of a bias reduces the variance, possibly very greatly. Eqs. (4.28-4.29) are sometimes known as the "tapered least-squares" solution, a label whose implication becomes clear later. $\mathbf{C}_{nn}$, which is not displayed, is readily found by direct computation as in Eq. (4.16).

The most basic and commonly seen form of this solution assumes $\mathbf{W}=\mathbf{R}_{nn}=\sigma_n^2\mathbf{I}$, and then,

$$\tilde{\mathbf{x}}=\left(\mathbf{E}^T\mathbf{E}+\alpha^2\mathbf{I}\right)^{-1}\mathbf{E}^T\mathbf{y}, \tag{4.32}$$

$$\mathbf{C}_{xx}=\sigma_n^2\left(\mathbf{E}^T\mathbf{E}+\alpha^2\mathbf{I}\right)^{-1}\mathbf{E}^T\mathbf{E}\left(\mathbf{E}^T\mathbf{E}+\alpha^2\mathbf{I}\right)^{-1}, \tag{4.33}$$

obviously a special case.

A physical motivation for the modified objective function (4.27) is obtained by noticing that a preference for a bounded $\|\mathbf{x}\|$ is easily produced by adding an equation set, $\mathbf{x}+\mathbf{n}_1=\mathbf{0}$, so that the combined set is,

$$\mathbf{E}\mathbf{x}+\mathbf{n}=\mathbf{y} \tag{4.34}$$

$$\alpha^2\left(\mathbf{x}+\mathbf{n}_1\right)=\mathbf{0} \tag{4.35}$$

or

$$\mathbf{E}_1\mathbf{x}+\mathbf{n}_2=\mathbf{y}_2$$

$$\mathbf{E}_1=\left\{\begin{matrix}\mathbf{E}\\ \alpha^2\mathbf{I}\end{matrix}\right\},\quad \mathbf{n}_2^T=\left[\mathbf{n}^T\quad\alpha^2\mathbf{n}_1^T\right],\quad \mathbf{y}_2^T=\left[\mathbf{y}^T\quad\mathbf{0}^T\right], \tag{4.36}$$

and in which α^2 expresses a preference for fitting the first or second sets more closely. Then J in Eq. (4.27) becomes the natural objective function to use. A preference that $\mathbf{x}\approx\mathbf{x}_0$ is readily imposed instead, with an obvious change in (4.27) or (4.35).

Note the important points, to be shown later, that the matrix inverses in Eqs. (4.28-4.29) will *always* exist, as long as $\alpha^2>0$, and that the expressions remain valid even if $M<N$. Tapered least-squares produces some control over the sum of squares of the relative norms of $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$, but still does not produce control over the individual elements $\tilde{x}_i$.

To gain control of the elements of $\tilde{x}_i$, we can further generalize the objective function by introducing another non-singular $N \times N$ weight matrix, $\mathbf{S}$ (which is usually symmetric) and,

$$\begin{aligned} J &= \mathbf{n}^T\mathbf{W}^{-1}\mathbf{n} + \mathbf{x}^T\mathbf{S}^{-1}\mathbf{x} && (4.37) \\ &= (\mathbf{y} - \mathbf{E}\mathbf{x})^T\mathbf{W}^{-1}(\mathbf{y} - \mathbf{E}\mathbf{x}) + \mathbf{x}^T\mathbf{S}^{-1}\mathbf{x}\,, && (4.38) \end{aligned}$$

for which Eq. (4.27) is a special case. Setting the derivatives with respect to $\mathbf{x}$ to zero results in,

$$\tilde{\mathbf{x}}= \left(\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E} + \mathbf{S}^{-1}\right)\mathbf{E}^T\mathbf{W}^{-1}\mathbf{y} \qquad (4.39)$$

$$\tilde{\mathbf{n}}=\mathbf{y} - \mathbf{E}\tilde{\mathbf{x}} \qquad (4.40)$$

$$\mathbf{C}_{xx} = \qquad (4.41)$$

$$\left(\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E} + \mathbf{S}^{-1}\right)\mathbf{E}^T\mathbf{W}^{-1}\mathbf{R}_{nn}\mathbf{W}^{-1}\mathbf{E}\left(\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E} + \mathbf{S}^{-1}\right),$$

and Eqs. (4.28-4.30) are a special case, with $\mathbf{S}^{-1} = \alpha^2\mathbf{I}$.

To interpret this result, suppose $\mathbf{S}$, $\mathbf{W}$ are positive definite and symmetric and thus have Cholesky decompositions. Then we can employ both matrices directly on the equations, $\mathbf{E}\mathbf{x} + \mathbf{n} = \mathbf{y}$,

$$\mathbf{W}^{-T/2}\mathbf{E}\mathbf{S}^{T/2}\mathbf{S}^{-T/2}\mathbf{x} + \mathbf{W}^{-T/2}\mathbf{n}=\mathbf{W}^{-T/2}\mathbf{y} \qquad (4.42)$$

$$\mathbf{E}'\mathbf{x}'+\mathbf{n}'=\mathbf{y}' \qquad (4.43)$$

$$\mathbf{E}'=\mathbf{W}^{-T/2}\mathbf{E}\mathbf{S}^{T/2},\ \mathbf{x}' = \mathbf{S}^{-T/2}\mathbf{x},\ \mathbf{n}' = \mathbf{W}^{-T/2}\mathbf{n},\ \mathbf{y}'= \mathbf{W}^{-T/2}\mathbf{y} \qquad (4.44)$$

The use of $\mathbf{S}$ in this way is “column scaling” because it weights the columns of $\mathbf{E}$. With Eqs. (4.43) the obvious objective function is,

$$J = \mathbf{n}'^T\mathbf{n}' + \mathbf{x}'^T\mathbf{x}', \qquad (4.45)$$

which is identical to Eq. (4.37) but in the new variables. In these rotated and stretched variables, all squared elements are on an equal footing and this simplest objective function is a sensible one. A clearer interpretation is obtained later by specific choice of $\mathbf{S}, \mathbf{W}$.

Like $\mathbf{W}$, one is completely free to choose $\mathbf{S}$ as one pleases. A common choice is to write, where $\mathbf{F}$ is $N \times N$,

$$\begin{aligned} \mathbf{S} &= \mathbf{F}^T\mathbf{F} \\ \mathbf{F} &= \alpha^2\left\{\begin{matrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 & 1 \end{matrix}\right\}, \end{aligned} \qquad (4.46)$$

whose effect is to minimize a term $\alpha^2 \sum_i (x_i - x_{i+1})^2$, which can be regarded as a "smoothest" solution, and using α^2 to trade smoothness against the size of $\|\tilde{\mathbf{n}}\|_2$, $\alpha\mathbf{F}$ is obtained from the Cholesky decomposition of $\mathbf{S}$.

By invoking the matrix inversion lemma, an alternative form for Eqs. (4.42 – 4.44) is found,

$$\tilde{\mathbf{x}} = \mathbf{S}\mathbf{E}^T \left(\mathbf{E}\mathbf{S}\mathbf{E}^T + \mathbf{W}\right)^{-1} \mathbf{y} \tag{4.47}$$

$$\tilde{\mathbf{n}} = \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}} \tag{4.48}$$

$$\mathbf{C}_{xx} = \mathbf{S}\mathbf{E}^T \left(\mathbf{E}\mathbf{S}\mathbf{E}^T + \mathbf{W}\right)^{-1} \mathbf{R}_{nn} \left(\mathbf{E}\mathbf{S}\mathbf{E}^T + \mathbf{W}\right)^{-1} \mathbf{E}\mathbf{S} \tag{4.49}$$

So far, all of this is conventional. But we have made a special point of displaying explicitly not only the elements $\tilde{\mathbf{x}}$, but also those of the residuals, $\tilde{\mathbf{n}}$. Notice that although we have considered only the formally over-determined system, $M > N$, we *always* determine not only the $N-$elements of $\tilde{\mathbf{x}}$, but also the M-elements of $\tilde{\mathbf{n}}$, for a total of $M + N$ values—extracted from the M-equations. It is apparent that any change in any element $\tilde{n}_i$ forces changes in $\tilde{\mathbf{x}}$. In this view, to which we adhere, systems of equations involving observations *always* contain more unknowns than equations. Another way to make the point is to re-write Eqs. (4.2) without distinction between $\mathbf{x}, \mathbf{n}$ as,

$$\mathbf{E}_1 \boldsymbol{\xi} = \mathbf{y} \tag{4.50}$$

$$\mathbf{E}_1 = \{\mathbf{E}, \mathbf{I}_M\}, \ \boldsymbol{\xi}^T = [\mathbf{x}, \mathbf{n}]^T \tag{4.51}$$

A combined weight matrix,

$$\mathbf{S}_1 = \left\{ \begin{array}{cc} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{W} \end{array} \right\}, \tag{4.52}$$

would be used, and any distinction between the $\mathbf{x}, \mathbf{n}$ solution elements is suppressed. Eqs. (4.50) are a formally underdetermined system, derived from the formally over-determined observed one. This identity leads us to the problem of formal underdetermination in the next Section.

In general with least-squares problems, the solution we seek can be regarded as any of the following equivalents:

(1) The $\tilde{\mathbf{x}}, \tilde{\mathbf{n}}$ satisfying

$$\mathbf{E}\mathbf{x} + \mathbf{n} = \mathbf{y}. \tag{4.53}$$

(2) $\tilde{\mathbf{x}}, \tilde{\mathbf{n}}$ satisfying the normal equations arising from J (Eq. 4.37).

(3) $\tilde{\mathbf{x}}, \tilde{\mathbf{n}}$ producing the minimum of J in (Eq. 4.37)

The point of this list lies with item 3: algorithms exist to find minima of functions by deterministic methods ("go downhill" from an initial guess)[22], or stochastic search methods (Monte Carlo) or even, conceivably, through a shrewd guess by the investigator. If an acceptable minimum of J is found, by whatever means, it is an acceptable solution (subject to further testing, and the possibility that there is more than one such solution). Search methods become essential for the nonlinear problems taken up later.

[22] Gill, Murray and Wright (1981).

4.3. Underdetermined Systems and Lagrange Multipliers. What does one do when the number, M, of equations is less than the number, N, of unknowns and no more observations are possible? We have seen that the claim that a problem involving observations is ever overdetermined is misleading—because each equation or observation always has a noise unknown, but to motivate some of what follows, it is helpful to first pursue a conventional approach.

One often attempts when $M < N$ to reduce the number of unknowns so that the formal overdeterminism is restored. Such a parameter reduction procedure may be sensible; but there are pitfalls. Let $p_i(t)$, $0 \leq i$ be a set of polynomials, e.g. Chebyschev or Laguerre, etc. Consider data produced from the formula,

$$y(t) = 1 + a_M p_M(t) + n(t), \tag{4.54}$$

which might be deduced by fitting a parameter set $[a_0, \ldots, a_M]^T$. The above fit of a quadratic term to data is of this form. If there are fewer than M observations, an attempt to fit with fewer parameters,

$$y = \sum_{j=0}^{Q} a_j p_j(t), \quad Q < M \tag{4.55}$$

may give a good, even perfect fit; but it would be incorrect. The reduction in model parameters in such a case biasses the result, perhaps hopelessly so. One is better off retaining the underdetermined system and making inferences concerning the possible values of a_i rather than using the form (4.54), in which any possibility of learning something about a_M has been eliminated.

Example 1. *Consider a tracer problem, not unlike those encountered in medicine, hydrology, oceanography, etc., described in Chapter 1. A box (see Fig. 2 of Chapter 1) is observed to contain a steady tracer concentration C_0, and is believed fed from two reservoirs each with tracer concentration of C_1, C_2 respectively. One seeks to determine the rates J_1, J_2 that fluid is fed from these two reservoirs into the box. Tracer balance is*

$$J_1 C_1 + J_2 C_2 - J_0 C_0 = 0,$$

where J_0 is rate at which fluid is removed from the box. Mass balance then requires

$$J_1 + J_2 = J_0.$$

Evidently, there are but two equations in three unknowns (and a perfectly good solution would be $J_1 = J_2 = J_3 = 0$); but as many have noticed, we can nonetheless, determine the relative fraction of the fluid coming from each reservoir. Divide both equations through by J_0,

$$\begin{aligned} \frac{J_1}{J_0} C_1 + \frac{J_2}{J_0} C_2 &= C_0 \\ \frac{J_1}{J_0} + \frac{J_2}{J_0} &= 1 \end{aligned}$$

producing two equations in two unknowns„ J_1/J_0, J_2/J_0, which has a unique stable solution (we are obviously ignoring the noise reality). Many examples can be given of such calculations in the literature, calculating the flux ratios—apparently definitively. But suppose the investigator is suspicious that there might be a third reservoir with tracer concentration C_3. Then the equations become

$$\begin{aligned}\frac{J_1}{J_0}C_1+\frac{J_2}{J_0}C_2+\frac{J_3}{J_0}C_3 &= C_0\\ \frac{J_1}{J_0}+\frac{J_2}{J_0}+\frac{J_3}{J_0} &= 1\end{aligned}$$

now underdetermined with two equations in three unknowns. If it is obvious that no such third reservoir exists, then the reduction to two equations in two unknowns is the right thing to do. But if there is even a suspicion of a third (or more) reservoir, one should solve these equations with one of the methods we will develop permitting construction and understanding of all possible solutions.

In general terms, parameter reduction can lead to model errors, that is, bias errors, which can produce wholly illusory results.[23] A common situation particularly in problems involving tracer movements in groundwater, ocean, or atmosphere, fitting a one or two-dimensional model to data which represent a fully three-dimensional field. The result may be pleasing, but possibly completely erroneous.

A general approach to solving underdetermined problems is to render the answer apparently unique by minimizing an objective function, subject to satisfaction of the linear constraints. To see how this can work, suppose that (4.2) are indeed formally underdetermined, that is, $M < N$, and seek the solution which exactly satisfies the equations and simultaneously renders an objective function, $J = \mathbf{x}^T\mathbf{x}$, as small as possible. Direct minimization of J leads to,

$$dJ = \frac{\partial J}{\partial \mathbf{x}}^T d\mathbf{x} = 2\mathbf{x}^T d\mathbf{x} = 0, \tag{4.56}$$

but unlike the case in Eq. (4.6), the coefficients of the individual dx_i can no longer be separately set to zero (i.e., $\mathbf{x} = 0$ is an incorrect solution) because the dx_i no longer vary independently, but are restricted to values satisfying $\mathbf{Ex} = \mathbf{y}$. One approach is to use the known dependencies to reduce the problem to a new one in which the differentials are independent. For example, suppose that there are general functional relationships

$$\begin{bmatrix} x_1 \\ \vdots \\ x_L \end{bmatrix} = \begin{bmatrix} \xi_1(x_{L+1},\ldots,x_N) \\ \vdots \\ \xi_L(x_{L+1},\ldots,x_N) \end{bmatrix}.$$

[23] Wunsch & Minster (1982).

Then the first $L = M - N$ elements of x_i may be eliminated, and the objective function becomes,

$$J = \left[\xi_1(x_{L+1}, \dots, x_N)^2 + \cdots + \xi_L(x_{L+1}, \dots, x_N)^2\right] + \left[x_{L+1}^2 + \cdots + x_N^2\right],$$

in which the remaining x_i, $L+1 \le i \le N$ are independently varying. In the present case, one can choose (arbitrarily) the first $N - M$ unknowns, $\mathbf{q} = [x_i]$, and define the last M unknowns $\mathbf{r} = [x_i]$, $N - M + 1 \le i \le N$, and rewrite the equations as

$$\left\{ \mathbf{E}_1 \quad \mathbf{E}_2 \right\} \begin{bmatrix} \mathbf{q} \\ \mathbf{r} \end{bmatrix} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \tag{4.57}$$

or,

$$\mathbf{q} = \mathbf{E}_1^{-1}\left(\mathbf{y}_1 - \mathbf{E}_2\mathbf{r}\right). \tag{4.58}$$

If $\mathbf{E}_1^{-1}$ does not exist, one can try any other subset of the x_i to eliminate until a suitable group is found. This approach is completely correct, but finding an explicit solution for L elements of $\mathbf{x}$ in terms of the remaining ones may be difficult or inconvenient.

Lagrange Multipliers and Adjoints

When it is inconvenient to find such an explicit representation eliminating some variables in favor of others, a standard procedure for finding the constrained minimum is to introduce a new vector "Lagrange multiplier," $\boldsymbol{\mu}$, of M-unknown elements, to make a new objective function,

$$\begin{aligned} J' &= J - 2\boldsymbol{\mu}^T(\mathbf{Ex} - \mathbf{y}) \\ &= \mathbf{x}^T\mathbf{x} - 2\boldsymbol{\mu}^T(\mathbf{Ex} - \mathbf{y}) \end{aligned} \tag{4.59}$$

and ask for its stationary point—treating both $\boldsymbol{\mu}$ and $\mathbf{x}$ as independently varying unknowns. The numerical 2 is introduced solely for notational tidiness.

The rationale for this procedure is straightforward.[24] Consider first, a very simple example, of one equation in one two unknowns,

$$x_1 - x_2 = 1 \tag{4.60}$$

and we seek the minimum norm solution,

$$\text{minimize: } J = x_1^2 + x_2^2, \tag{4.61}$$

subject to Eq. (4.60). The differential,

$$dJ = 2x_1dx_1 + 2x_2dx_2 = 0, \tag{4.62}$$

[24] Morse & Feshbach (1953, p. 238); Strang (1986).

leads to the unacceptable solution $x_1 = x_2 = 0$, if we should incorrectly set the coefficients of dx_1, dx_2 to zero. (The step is incorrect because x_1, x_2 cannot vary independently, being connected through Eq. (4.60).) Consider instead a modified objective function

$$J' = J - 2\mu\left(x_1 - x_2 - 1\right), \tag{4.63}$$

where μ is unknown. The differential of J' is

$$dJ' = 2x_1dx_1 + 2x_2dx_2 - 2\mu\left(dx_1 - dx_2\right) - 2\left(x_1 - x_2 - 1\right)d\mu = 0, \tag{4.64}$$

or

$$dJ'/2 = dx_1\left(x_1 - \mu\right) + dx_2\left(x_2 + \mu\right) + d\mu\left(x_1 - x_2 - 1\right) = 0. \tag{4.65}$$

We are free to choose, $x_1 = \mu$ which kills off the differential involving dx_1. But now only the differentials $dx_2, d\mu$ remain; as they can vary independently, their coefficients must vanish separately and we have,

$$x_2 = -\mu \tag{4.66}$$

$$x_1 - x_2 = 1. \tag{4.67}$$

Note that the second of these recovers the original equation. Substituting $x_1 = \mu$, we have $2\mu = 1$, or $\mu = 1/2$, and $x_1 = 1/2, x_2 = -1/2$, $J = 0.5$, and one can confirm that this is indeed the "constrained" minimum. (A "stationary" value of J' was found, not an absolute minimum value, because J' is no longer necessarily positive: it has a "saddle point", which we have found.)

Before writing out the general case, note the following question: Suppose the constraint equation was changed to,

$$x_1 - x_2 = \Delta. \tag{4.68}$$

How much would J change as Δ is varied? With variable Δ, (4.64) becomes,

$$dJ' = 2dx_1\left(x_1 - \mu\right) + 2dx_2\left(x_2 + \mu\right) + 2d\mu\left(x_1 - x_2 - \Delta\right) + 2\mu d\Delta. \tag{4.69}$$

But the first three terms on the right vanish, and hence

$$\frac{\partial J'}{\partial \Delta} = 2\mu = \frac{\partial J}{\partial \Delta}, \tag{4.70}$$

because $J = J'$ at the stationary point (from (4.68). *Thus μ is the sensitivity of the objective function J to perturbations in the right-hand side of the constraint equation.* If Δ is changed from 1, to 1.2, it can be confirmed that the approximate change in the value of J is 0.2 as one deduces immediately from Eq. (4.70).

Now we develop this method generally. Reverting to Eq. (4.59),

$$\begin{aligned} &dJ' = dJ - 2\boldsymbol{\mu}^T\mathbf{E}d\mathbf{x}- \left(\mathbf{Ex}-\mathbf{y}\right)^T d\boldsymbol{\mu} \\ = \; &\left(\frac{\partial J}{\partial x_1} - 2\boldsymbol{\mu}^T\mathbf{e}_1\right)dx_1 + \left(\frac{\partial J}{\partial x_2} - 2\boldsymbol{\mu}^T\mathbf{e}_2\right)dx_2 + \cdots + \left(\frac{\partial J}{\partial x_N} - 2\boldsymbol{\mu}^T\mathbf{e}_N\right)dx_N - \left(\mathbf{Ex}-\mathbf{y}\right)^T d\boldsymbol{\mu} \\ = \; &\left(2x_1 - 2\boldsymbol{\mu}^T\mathbf{e}_1\right)dx_1 + \left(2x_2 - 2\boldsymbol{\mu}^T\mathbf{e}_2\right)dx_2 + ... + \left(2x_N - 2\boldsymbol{\mu}^T\mathbf{e}_N\right)dx_N - \left(\mathbf{Ex}-\mathbf{y}\right)^T d\boldsymbol{\mu} = 0 \end{aligned}$$

Here the $\mathbf{e}_i$ are the corresponding columns of $\mathbf{E}$. The coefficients of the first $M-$differentials dx_ican be set to zero by assigning, $x_i = \boldsymbol{\mu}^T\mathbf{e}_i$, leaving $N-M$ differentials dx_i whose coefficients must separately vanish (hence they *all* vanish, but for two separate reasons), plus the coefficient of the $M-d\mu_i$which must also vanish separately. This recipe produces, from Eq. (4.59),

$$\tfrac{1}{2}\frac{\partial J'}{\partial \mathbf{x}} = \mathbf{x} - \mathbf{E}^T\boldsymbol{\mu} = 0 \tag{4.71}$$

$$\tfrac{1}{2}\frac{\partial J'}{\partial \boldsymbol{\mu}} = \mathbf{Ex} - \mathbf{y} = \mathbf{0}\,, \tag{4.72}$$

where the first equation set is the result of the vanishing of the coefficients of dx_i and the second, which is the original set of equations, arises from the vanishing of the coefficients of the $d\mu_i$. The convenience of being able to treat all the x_i as independently varying is offset by the increase in problem dimensions by the introduction of the $M-$unknown μ_i. The first set is $N-$equations for $\boldsymbol{\mu}$ in terms of $\mathbf{x}$, and the second set is $M-$equations in $\mathbf{x}$ in terms of $\mathbf{y}$. Taken together, these are $M+N$ equations in $M+N$ unknowns, and hence just-determined no matter what the ratio of M to N, assuming $\mathbf{E}$ is full-rank.

Eq. (4.72) is,

$$\mathbf{E}^T\boldsymbol{\mu} = \mathbf{x} \tag{4.73}$$

and substituting for $\mathbf{x}$ into (4.71),

$$\mathbf{E}\mathbf{E}^T\boldsymbol{\mu} = \mathbf{y}\,,$$

$$\tilde{\boldsymbol{\mu}} = (\mathbf{E}\mathbf{E}^T)^{-1}\mathbf{y}\,, \tag{4.74}$$

assuming the inverse exists, and

$$\tilde{\mathbf{x}} = \mathbf{E}^T(\mathbf{E}\mathbf{E}^T)^{-1}\mathbf{y} \tag{4.75}$$

$$\tilde{\mathbf{n}} = \mathbf{0} \tag{4.76}$$

$$\mathbf{C}_{xx} = 0. \tag{4.77}$$

($\mathbf{C}_{xx} = 0$ because formally we estimate $\tilde{\mathbf{n}} = \mathbf{0}$).

Eqs.(4.75-4.77) are the classical solution of minimum norm of $\mathbf{x}$, satisfying the constraints exactly while minimizing the solution length. That a minimum is achieved can be verified by evaluating the second derivatives of J' at the solution point. The minimum occurs at a saddle point in $\mathbf{x}$, $\boldsymbol{\mu}$ space[25] and where the term proportional to $\boldsymbol{\mu}$ necessarily vanishes. The operator $\mathbf{E}^T(\mathbf{E}\mathbf{E}^T)^{-1}$ is sometimes called a "Moore-Penrose inverse." If $\mathbf{E}\mathbf{E}^T$ does not have an inverse, this solution does not exist (but we will find ways to generalize it to deal with the singular case).

Eqs. (4.73) for $\boldsymbol{\mu}$ in terms of $\mathbf{x}$ involves the coefficient matrix $\mathbf{E}^T$. An intimate connection exists between matrix transposes and adjoints of differential equations (see the Appendix to this Chapter), and thus $\boldsymbol{\mu}$ is sometimes called the "adjoint solution," with $\mathbf{E}^T$ defining the "adjoint model"[26] in Eq.(4.73), and with $\mathbf{x}$ acting as a forcing term. The original Eqs. (4.2) were assumed formally underdetermined, and thus the adjoint model equations in (4.73) are necessarily formally overdetermined.

The physical interpretation of $\boldsymbol{\mu}$ can be obtained as above by considering the way in which J would vary with infinitesimal changes in $\mathbf{y}$. As in the special case done above, $J = J'$ at the stationary point. Hence,

$$dJ' = dJ - 2\boldsymbol{\mu}^T\mathbf{E}d\mathbf{x} - 2\left(\mathbf{E}\mathbf{x}-\mathbf{y}\right)^T d\boldsymbol{\mu} + 2\boldsymbol{\mu}^T d\mathbf{y} = 0, \tag{4.78}$$

or, since the first three terms on the right vanish at the stationary point,

$$\frac{\partial J'}{\partial \mathbf{y}} = \frac{\partial J}{\partial \mathbf{y}} = 2\boldsymbol{\mu}, \tag{4.79}$$

and the Lagrange multipliers are the sensitivity of J, at the stationary point, to perturbations in the parameters $\mathbf{y}$. This conclusion leads, in Chapter 4, to the scrutiny of the Lagrange multipliers as a means of understanding the sensitivity of models and the flow of information within them.

If the Eqs. (4.2) are first column scaled using $\mathbf{S}^{-T/2}$, Eqs. (4.75)–(4.77) are in the primed variables, and the solution in the original variables is

$$\tilde{\mathbf{x}} = \mathbf{S}\mathbf{E}^T(\mathbf{E}\mathbf{S}\mathbf{E}^T)^{-1}\mathbf{y} \tag{4.80}$$

$$\tilde{\mathbf{n}} = \mathbf{0} \tag{4.81}$$

$$\mathbf{C}_{xx} = \mathbf{0}, \tag{4.82}$$

and the result depends directly upon $\mathbf{S}$. If a row scaling with $\mathbf{W}^{-T/2}$ is used, it is readily shown that $\mathbf{W}$ disappears from the solution and has no effect on it.

Eqs. (4.80)–(4.82) are a solution, but there is the same fatal defect as in Eq. (4.81)—$\tilde{\mathbf{n}} = \mathbf{0}$ is usually unacceptable when $\mathbf{y}$ are observations. Furthermore, $\|\tilde{\mathbf{x}}\|$ is again uncontrolled, and $\mathbf{E}\mathbf{S}\mathbf{E}^T$ may not have an inverse.

$\mathbf{n}$ must be regarded as fully an element of the solution, as much as $\mathbf{x}$. Equations representing observations can always be written as in (4.50), and can be solved exactly. Therefore, we now

[25] See Sewell (1987) for an interesting discussion.

[26] But the matrix transpose is not what the older literature calls the "adjoint matrix," and which is quite different. In the more recent literature the latter has been termed the "adjugate" matrix to avoid confusion.

use a modified objective function, allowing for general $\mathbf{S}, \mathbf{W}$,

$$J = \mathbf{x}^T\mathbf{S}^{-1}\mathbf{x} + \mathbf{n}^T\mathbf{W}^{-1}\mathbf{n} - 2\boldsymbol{\mu}^T(\mathbf{Ex} + \mathbf{n} - \mathbf{y}), \tag{4.83}$$

with both $\mathbf{x}$, $\mathbf{n}$ appearing in the objective function. Setting the derivatives of (4.83) with respect to $\mathbf{x}$, $\mathbf{n}$, $\boldsymbol{\mu}$ to zero, and solving the resulting normal equations produces,

$$\tilde{\mathbf{x}} = \mathbf{SE}^T\left(\mathbf{ESE}^T + \mathbf{W}\right)^{-1}\mathbf{y} \tag{4.84}$$

$$\tilde{\mathbf{n}} = \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}} \tag{4.85}$$

$$\mathbf{C}_{xx} = \mathbf{SE}^T\left(\mathbf{ESE}^T + \mathbf{I}\right)^{-1}\mathbf{R}_{nn}\left(\mathbf{ESE}^T + \mathbf{I}\right)^{-1}\mathbf{ES} \tag{4.86}$$

$$\tilde{\boldsymbol{\mu}} = \tilde{\mathbf{n}} \tag{4.87}$$

which is identical to Eqs. (4.47-4.49) or to the alternate from Eq.$(4.39 - 4.41)$ derived from an objective function without Lagrange multipliers.

Eqs. (4.47-4.49) and (4.84-4.86) result from two very different appearing objective functions—one in which the equations are imposed in the mean-square, and one in which they are imposed exactly, using Lagrange multipliers. Constraints in the mean-square will be termed "soft", and those imposed exactly are "hard."[27] The distinction is however, largely illusory: although (4.2) are being imposed exactly, it is only the presence of the error term, $\mathbf{n}$, which permits the equations to be written as equalities and thus as hard constraints. The hard and soft constraints here produce an identical solution. In some (rare) circumstances, which we will discuss briefly below, one may wish to impose exact constraints upon the elements of $\tilde{x}_i$. The solution (4.75)–(4.77) was derived from the noise-free hard constraints, $\mathbf{Ex} = \mathbf{y}$, but we ended by rejecting it as generally inapplicable.

Once again, $\mathbf{n}$ is only by convention discussed separately from $\mathbf{x}$, and is fully a part of the solution. The combined form (4.50), which literally treats $\mathbf{x}$, $\mathbf{n}$ as the solution, are imposed through a hard constraint on the objective function,

$$J = \boldsymbol{\xi}^T\boldsymbol{\xi} - 2\boldsymbol{\mu}^T(\mathbf{E}_1\boldsymbol{\xi} - \mathbf{y}), \tag{4.88}$$

where $\boldsymbol{\xi} = [\mathbf{S}^{-T/2}\mathbf{x}, \mathbf{W}^{-T/2}\mathbf{n}]^T$, which is identical to Eq. (4.83). (There are numerical advantages however, in working with objects in two spaces of dimensions M and N, rather than a single space of dimension $M + N$.)

[27] In the meteorological terminology of Sasaki (1970) and others, exact relationships are called "strong" constraints, and those imposed in the mean-square are "weak" ones.

4.4. Interpretation of Discrete Adjoints. When the operators are matrices, as they are in discrete formulations, then the adjoint is just the transposed matrix. Sometimes the adjoint has a simple physical interpretation. Suppose, e.g., that scalar y was calculated from a sum,

$$y = \mathbf{A}\mathbf{x}, \ \mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & . & 1 \end{bmatrix}. \tag{4.89}$$

Then the adjoint operator applied to y is evidently,

$$\mathbf{r} = \mathbf{A}^T y = \begin{bmatrix} 1 \\ 1 \\ 1 \\ . \\ 1 \end{bmatrix} y = \mathbf{x} \tag{4.90}$$

Thus the adjoint operator "sprays" the average back out onto the originating vector, and might be thought of as a kind of inverse operator.

A more interesting case is a first-difference forward operator,

$$\mathbf{A} = \left\{ \begin{array}{cccccc} -1 & 1 & & & & \\ & -1 & 1 & & & \\ & & -1 & 1 & & \\ & & & . & . & . \\ & & & & -1 & 1 \\ & & & & & -1 \end{array} \right\}, \tag{4.91}$$

that is,

$$y_i = x_{i+1} - x_i, \tag{4.92}$$

(with the exception of the last element, $y_N = -x_N$).

Then its adjoint is,

$$\mathbf{A}^T = \left\{ \begin{array}{cccccc} -1 & & & & & \\ 1 & -1 & & & & \\ & 1 & -1 & & & \\ & & . & . & & \\ & & & 1 & -1 & \\ & & & & 1 & -1 \end{array} \right\} \tag{4.93}$$

that is a first-difference *backward* operator with $\mathbf{z} = \mathbf{A}^T\mathbf{y}$, producing $z_i = y_{i-1} - y_i$ with again, the exception of the end point, now z_1.

In general, the transpose matrix, or adjoint operator is *not* simply interpretable as an inverse operation as in the summation/spray-out case might have suggested.[28] A more general

[28] Claerbout (2001) displays more examples, and Lanczos (1960) gives a very general discussion of operators and their adjoints, Green functions, and their adjoints. See also the Appendix to this Chapter.

understanding of the relationship between adjoints and inverses will be obtained in the next Section.

5. The Singular Vector Expansion

Least-squares is a very powerful, very useful method for finding solutions of linear simultaneous equations of any dimensionality and one might wonder why it is necessary to discuss any other form of solution. But even in the simplest form of least-squares, the solution is dependent upon the inverses of $\mathbf{E}^T\mathbf{E}$, or $\mathbf{E}\mathbf{E}^T$. In practice, their existence cannot be guaranteed, and we need to understand what that means, the extent to which solutions can be found when the inverses do not exist and the effect of introducing weight matrices $\mathbf{W}$, $\mathbf{S}$. This problem is intimately related to the issue of controlling solution and residual norms. Second, the relationship between the equations and the solutions is somewhat impenetrable, in the sense that structures in the solutions are not easily relatable to particular elements of the data y_i. For many purposes, particularly physical insight, understanding the structure of the solution is essential. We will return to examine the least-squares solutions using some extra machinery.

5.1. Simple Vector Expansions. Consider again the elementary problem (2.1) of representing an L–dimensional vector $\mathbf{f}$ as a sum of a complete set of L–orthonormal vectors $\mathbf{g}_i$, $1 \le i \le L$, $\mathbf{g}_i^T\mathbf{g}_j = \delta_{ij}$. Without error,

$$\mathbf{f} = \sum_{j=1}^{L} a_j \mathbf{g}_j, \quad a_j = \mathbf{g}_j^T \mathbf{f}\,. \tag{5.1}$$

But if for some reason, only the first K coefficients a_j are known, we can only approximate $\mathbf{f}$ by its first K terms:

$$\begin{aligned}\tilde{\mathbf{f}} &= \sum_{j=1}^{K} b_j \mathbf{g}_j \\ &= \mathbf{f} + \delta\mathbf{f}_1,\end{aligned} \tag{5.2}$$

and there is an error, $\delta\mathbf{f}_1$. From the orthogonality of the $\mathbf{g}_i$, it follows that $\delta\mathbf{f}_1$ will have minimum l_2 norm only if it is orthogonal to the K vectors retained in the approximation, and then only if $b_j = a_j$ as given by (5.1). The only way the error could be reduced further is by increasing K.

Define an $L \times K$ matrix, $\mathbf{G}_K$ whose columns are the first K of the $\mathbf{g}_j$. Then $\mathbf{a} = \mathbf{G}_K^T\mathbf{f}$ is the vector of coefficients $a_j = \mathbf{g}_j^T\mathbf{f}$, $1 \le j \le K$, and the finite representation (5.2) is (one should write it out),

$$\tilde{\mathbf{f}} = \mathbf{G}_K\mathbf{a} = \mathbf{G}_K(\mathbf{G}_K^T\mathbf{f}) = (\mathbf{G}_K\mathbf{G}_K^T)\mathbf{f}\,, \quad \mathbf{a} = \{a_i\}, \tag{5.3}$$

where the third equality follows from the associative properties of matrix multiplication. This expression shows that *representation of a vector in an incomplete orthonormal set produces a resulting approximation which is a simple linear combination of the elements of the correct values*

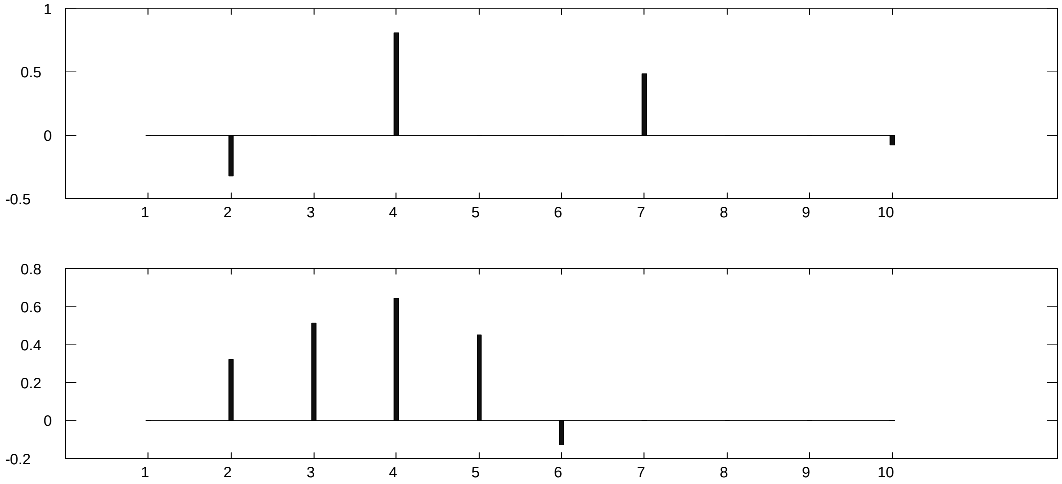


FIGURE 9. Example of a row, p, of a 10×10 resolution matrix, perhaps the fourth one, showing widely distributed averaging in forming $\mathbf{f}_p$ (upper panel). Lower panel shows so-called compact resolution, in which the solution e.g., is a readily interpreted local average of the true solution. Such situations are not common.

(i.e., a weighted average, or "filtered" version of them). Row i of $\mathbf{G}_K\mathbf{G}_K^T$ produces the weighted linear combination of the true elements of $\mathbf{f}$ which will appear as $\tilde{f}_i$.

Because the columns of $\mathbf{G}_K$ are orthonormal, $\mathbf{G}_K^T\mathbf{G}_K = \mathbf{I}_K$, that is, the $K \times K$ identity matrix; but $\mathbf{G}_K\mathbf{G}_K^T \neq \mathbf{I}_L$ unless $K = L$ (that $\mathbf{G}_L\mathbf{G}_L^T = \mathbf{I}_L$ for $K = L$ follows from the theorem for *square* matrices that shows a left inverse is also a right inverse.) If $K < L$, $\mathbf{G}_K$ is "semi-orthogonal." If $K = L$, it is "orthogonal"; in this case, $\mathbf{G}_L^{-1} = \mathbf{G}_L^T$. If it is only semi-orthogonal, $\mathbf{G}_K^T$ is a left inverse, but not a right inverse. Any orthogonal matrix has the property that its transpose is identical to its inverse.

$\mathbf{G}_K\mathbf{G}_K^T$ is known as a "resolution matrix," with a simple interpretation. Suppose the true value of $\mathbf{f}$ were $\mathbf{e}_p = [\,0\;0\;0\;\ldots\;0\;1\;0\;.\;0\;..\;0\,]^T$, that is, a Kronecker delta δ_{jp}, with unity in element p and zero otherwise. Then the incomplete expansion (5.2) or (5.3) would not reproduce the delta function, but rather

$$\tilde{\mathbf{f}}_{j0} = \mathbf{G}_K\mathbf{G}_K^T\mathbf{e}_p\,, \tag{5.4}$$

which is row p of $\mathbf{G}_K\mathbf{G}_K^T$. Each row of the resolution matrix tells one what the corresponding form of the vector would be, if its true form were a Kronecker delta.

To form a Kronecker delta requires a complete set of vectors. An analogous elementary result of Fourier analysis shows that a Dirac delta function demands contributions from all frequencies to represent a narrow, very high pulse. Removal of some of the requisite vectors (sinusoids) produces peak broadening and sidelobes. Here, depending upon the precise structure of the $\mathbf{g}_i$, the broadening and sidelobes can be complicated. If one is lucky, the effect could be a simple broadening (schematically shown in figure 9) without distant sidelobes), leading to the tidy interpretation of the result as a local average of the true values, called "compact resolution."[29] This sometimes happens, but cannot in general be expected.

A resolution matrix has the property

$$\operatorname{trace}(\mathbf{G}_K\mathbf{G}_K^T) = K\,, \tag{5.5}$$

which follows from noting that,

$$\operatorname{trace}(\mathbf{G}_K^T\mathbf{G}_K) = \operatorname{trace}\left(\mathbf{G}_K\mathbf{G}_K^T\right) = \operatorname{trace}(\mathbf{I}_K) = K.$$

5.2. Square-Symmetric Problem. Eigenvalues/Eigenvectors. Orthogonal vector expansions are particularly simple to use and interpret, but might seem irrelevant to dealing with simultaneous equations where neither the row nor column vectors of the coefficient matrix are so simply related. What we will show however, is that we can always find sets of orthonormal vectors to greatly simplify the job of solving simultaneous equations. To do so, we digress to recall the basic elements of the "eigenvector/eigenvalue problem" mentioned only in passing on P. 23.

Consider a *square*, $M \times M$ matrix $\mathbf{E}$ and the simultaneous equations

$$\mathbf{E}\mathbf{g}_i = \lambda_i \mathbf{g}_i, \quad 1 \le i \le M\,, \tag{5.6}$$

that is, the problem of finding a set of vectors $\mathbf{g}_i$ whose dot products with the rows of $\mathbf{E}$ are proportional to themselves. Such vectors are "eigenvectors," and the constants of proportionality are the "eigenvalues." Under special circumstances, the eigenvectors form an orthonormal spanning set: *textbooks show that if* $\mathbf{E}$ *is square and symmetric, such a result is guaranteed.* It is easy to see that if two λ_j, λ_k are distinct, then the corresponding eigenvectors are orthogonal:

$$\mathbf{E}\mathbf{g}_j = \lambda_j \mathbf{g}_j, \tag{5.7}$$

$$\mathbf{E}\mathbf{g}_k = \lambda_k \mathbf{g}_k \tag{5.8}$$

Left multiply the first of these by $\mathbf{g}_k^T$, and the second by $\mathbf{g}_j^T$ and subtract:

$$\mathbf{g}_k^T\mathbf{E}\mathbf{g}_j - \mathbf{g}_j^T\mathbf{E}\mathbf{g}_k = (\lambda_j - \lambda_k)\,\mathbf{g}_k^T\mathbf{g}_j. \tag{5.9}$$

But because $\mathbf{E} = \mathbf{E}^T$, the left-hand side vanishes, and hence $\mathbf{g}_k^T\mathbf{g}_j$ by the assumption $\lambda_j \neq \lambda_k$. A similar construction proves that the λ_i are all real, and an elaboration shows that for coincident λ_i, the corresponding eigenvectors can always be made orthogonal.

[29] Wiggins (1972).

Suppose for the moment that we have such a special case, and recall how eigenvectors can be used to solve (2.16). By convention, the pairs $(\lambda_i, \mathbf{g}_i)$ are ordered in the sense of decreasing λ_i. If some λ_i are repeated, an arbitrary order choice is made.

With an orthonormal, spanning set, both the known $\mathbf{y}$ and the unknown $\mathbf{x}$ can be written as,

$$\mathbf{x} = \sum_{i=1}^{M} \alpha_i \mathbf{g}_i, \quad \alpha_i = \mathbf{g}_i^T \mathbf{x}\,, \tag{5.10}$$

$$\mathbf{y} = \sum_{i=1}^{M} \beta_i \mathbf{g}_i, \quad \beta_i = \mathbf{g}_i^T \mathbf{y}\,. \tag{5.11}$$

By convention, $\mathbf{y}$ is known, and therefore β_i can be regarded as given. If we could find the α_i, $\mathbf{x}$ would be known.

Substitute (5.10) into $\mathbf{Ex} = \mathbf{y}$,

$$\mathbf{E} \sum_{i=1}^{M} \alpha_i \mathbf{g}_i = \sum_{i=1}^{M} \left(\mathbf{g}_i^T \mathbf{y}\right) \mathbf{g}_i, \tag{5.12}$$

or, using the eigenvector property,

$$\sum_{i=1}^{M} \alpha_i \lambda_i \mathbf{g}_i = \sum_{i}^{M} \left(\mathbf{g}_i^T \mathbf{y}\right) \mathbf{g}_i. \tag{5.13}$$

But the expansion vectors are orthonormal and so

$$\lambda_i \alpha_i = \mathbf{g}_i^T \mathbf{y} \tag{5.14}$$

$$\alpha_i = \frac{\mathbf{g}_i^T \mathbf{y}}{\lambda_i} \tag{5.15}$$

$$\mathbf{x} = \sum_{i=1}^{M} \frac{\mathbf{g}_i^T \mathbf{y}}{\lambda_i} \mathbf{g}_i\,. \tag{5.16}$$

Apart from an obvious difficulty if an eigenvalue vanishes, the problem is now completely solved. If we define a diagonal matrix, $\mathbf{\Lambda}$, with elements, λ_i, ordered by convention in descending numerical value, and the matrix $\mathbf{G}$, whose columns are the corresponding $\mathbf{g}_i$ in the same order, the solution to (2.16) can be written, from (5.10), (5.14)–(5.16) as

$$\boldsymbol{\alpha} = \mathbf{\Lambda}^{-1} \mathbf{G}^T \mathbf{y} \tag{5.17}$$

$$\mathbf{x} = \mathbf{G} \mathbf{\Lambda}^{-1} \mathbf{G}^T \mathbf{y} \tag{5.18}$$

where $\mathbf{\Lambda}^{-1} = (1/\lambda_i)$.

Vanishing eigenvalues, $i = i_0$, cause trouble and we must consider them. Let the corresponding eigenvectors be $\mathbf{g}_{i_0}$. Then any part of the solution which is proportional to such an eigenvector is "annihilated" by $\mathbf{E}$, that is, $\mathbf{g}_{i_0}$ is orthogonal to all the rows of $\mathbf{E}$. Such a result means that there is no possibility that anything in $\mathbf{y}$ could provide any information about the coefficient α_{i_0}. If $\mathbf{y}$ corresponds to a set of observations (data), then $\mathbf{E}$ represents the connection ("mapping") between system unknowns and observations. The existence of zero eigenvalues shows that the act of observation of $\mathbf{x}$ removes certain structures in the solution which are then indeterminate. Vectors $\mathbf{g}_{i_0}$ (and there may be many of them) are said to lie in the "nullspace" of $\mathbf{E}$. Eigenvectors corresponding to non-zero eigenvalues lie in its "range." The simplest example is given by the "observations,"

$$x_1 + x_2 = 3\,,$$
$$x_1 + x_2 = 3\,.$$

Any structure in $\mathbf{x}$ such that $x_1 = -x_2$ is destroyed by these observations, and by inspection, the nullspace vector must be $\mathbf{g}_2 = [1, -1]^T/\sqrt{2}$. (The purpose of showing the observation twice is to produce an $\mathbf{E}$ which is square.)

Suppose there are $K < M$ non-zero λ_i. Then for $i > K$, Eq. (5.14) is

$$0\alpha_i = \mathbf{g}_i^T\mathbf{y}, \quad K + 1 \le i \le M\,, \tag{5.19}$$

and two cases must be distinguished.

Case (1):

$$\mathbf{g}_i^T\mathbf{y} = 0\,, \quad K + 1 \le i \le M\,. \tag{5.20}$$

We could then put $\alpha_i = 0$, $K + 1 \le i \le M$, and the solution can be written

$$\tilde{\mathbf{x}} = \sum_{i=1}^{K} \frac{\mathbf{g}_i^T\mathbf{y}}{\lambda_i}\mathbf{g}_i, \tag{5.21}$$

and $\mathbf{E}\tilde{\mathbf{x}} = \mathbf{y}$, *exactly.* We have put a tilde over $\mathbf{x}$ because a solution of the form,

$$\tilde{\mathbf{x}} = \sum_{i=1}^{K} \frac{\mathbf{g}_i^T\mathbf{y}}{\lambda_i}\mathbf{g}_i + \sum_{i=K+1}^{M} \alpha_i\mathbf{g}_i\,, \tag{5.22}$$

with the remaining α_i taking on arbitrary values also satisfies the equations exactly. That is, the true value of $\mathbf{x}$ *could* contain structures proportional to the nullspace vectors of $\mathbf{E}$, but the equations (2.16) neither require their presence, nor provide information necessary to determine their amplitudes. We thus have a situation with a "solution nullspace." Define the matrix $\mathbf{G}_K$ to be $M \times K$, carrying only the first K of the $\mathbf{g}_i$, that is the range vectors, $\mathbf{\Lambda}_K$ to be $K \times K$ with only the first K, non-zero eigenvalues, and the columns of $\mathbf{Q}_G$ are the $M - K$ nullspace vectors (it is $M \times (M - K)$), then the solutions (5.21) and (5.22) are,

$$\tilde{\mathbf{x}} = \mathbf{G}_K\mathbf{\Lambda}_K^{-1}\mathbf{G}_K^T\mathbf{y}\,, \tag{5.23}$$

$$\tilde{\mathbf{x}} = \mathbf{G}_K\mathbf{\Lambda}_K^{-1}\mathbf{G}_K^T\mathbf{y} + \mathbf{Q}_G\boldsymbol{\alpha}_G \tag{5.24}$$

where $\boldsymbol{\alpha}_G$ is the vector of unknown nullspace coefficients, respectively.

Eq. (5.20) is often known as a "solvability condition." The solution in (5.23), with no nullspace contribution will be called the "particular" solution.

If $\mathbf{G}$ is written as a partitioned matrix,

$$\mathbf{G} = \{\mathbf{G}_K \quad \mathbf{Q}_G\}\,,$$

it follows from the column orthonormality that

$$\mathbf{G}\mathbf{G}^T = \mathbf{I} = \mathbf{G}_K\mathbf{G}_K^T + \mathbf{Q}_G\mathbf{Q}_G^T \tag{5.25}$$

or

$$\mathbf{Q}_G\mathbf{Q}_G^T = \mathbf{I} - \mathbf{G}_K\mathbf{G}_K^T\,. \tag{5.26}$$

Vectors $\mathbf{Q}_G$ span the nullspace of $\mathbf{G}$.

<u>Case (2)</u>:

$$\mathbf{g}_i^T\mathbf{y} \neq 0\,, \quad i > K+1\,, \tag{5.27}$$

for one or more of the nullspace vectors. In this case, eq. (5.14) is the contradiction,

$$0\alpha_i \neq 0,$$

and eq. (5.13) is actually

$$\sum_{i=1}^{K} \lambda_i\alpha_i\mathbf{g}_i = \sum_{i=1}^{M}(\mathbf{g}_i^T\mathbf{y})\mathbf{g}_i, \quad K < M\,, \tag{5.28}$$

that is, with differing upper limits on the sums. Owing to the orthonormality of the $\mathbf{g}_i$, there is no choice of α_i, $1 \leq i \leq K$ on the left which can match the last $M-K$ terms on the right. Evidently there is no solution in the conventional sense unless (5.20) is satisfied, hence the name "solvability condition." What is the best we might do? Define "best" to mean that the solution $\tilde{\mathbf{x}}$ should be chosen such that,

$$\mathbf{E}\tilde{\mathbf{x}} = \tilde{\mathbf{y}},$$

where the difference, $\tilde{\mathbf{n}} = \mathbf{y} - \tilde{\mathbf{y}}$, which we call the "residual," should be as small as possible (in the l_2 norm). If this choice is made, then the orthogonality of the $\mathbf{g}_i$ shows immediately that the best choice is still (5.15), $1 \leq i \leq K$. No choice of nullspace vector coefficients, nor any other value of the coefficients of the range vectors, can reduce the norm of $\tilde{\mathbf{n}}$. The best solution is then also (5.21) or (5.23).

In this situation, we are no longer solving the equations (2.16), but rather are dealing with a set that could be written,

$$\mathbf{E}\mathbf{x} \sim \mathbf{y}, \tag{5.29}$$

where the demand is for a solution that is the "best possible," in the sense just defined. Such statements of approximation are awkward, and it is more useful to always rewrite (5.29) as,

$$\mathbf{Ex} + \mathbf{n} = \mathbf{y}, \tag{5.30}$$

where $\mathbf{n}$ is the residual. If $\tilde{\mathbf{x}}$ is given by (5.22) then,

$$\tilde{\mathbf{n}} = \sum_{i=K+1}^{M} (\mathbf{g}_i^T \mathbf{y})\mathbf{g}_i, \tag{5.31}$$

by (5.28). Notice that $\tilde{\mathbf{n}}^T \tilde{\mathbf{y}} = \mathbf{0} : \tilde{\mathbf{y}}$ is orthogonal to the residuals.

This situation, where we started with M-equations in M-unknowns, but found in practice that some structures of the solution could not actually be determined, is labeled "formally just-determined," where the expression "formally" alludes to the fact that the appearance of a just-determined system did not mean that the characterization was true in practice. One or more vanishing eigenvalues mean that neither the rows nor columns of $\mathbf{E}$ are spanning sets.

Some decision has to be made about the coefficients of the nullspace vectors in (5.24). We could use the form as it stands, regarding it at as the "general solution." The analogy with the solution of differential equations should be apparent—typically, there is a particular solution and a homogeneous solution—here the nullspace vectors. When solving a differential equation, determination of the magnitude of the homogeneous solution requires additional information, often provided by boundary or initial conditions; here additional information is also necessary, but missing.

Despite the presence of indeterminate elements in the solution, we know a great deal about them: they are proportional to the known nullspace vectors. Depending upon the specific situation, we might conceivably be in a position to obtain more observations, and would seriously consider observational strategies directed at observing these missing structures. The reader is also reminded of the discussion of the Neumann problem in Chapter 1.

Another approach is to define a "simplest" solution, appealing to what is usually known as "Occam's Razor," or the "principal of parsimony," that in choosing between multiple explanations of a given phenomenon, the simplest one is usually the best. What is "simplest" can be debated, but here there is a compelling choice: The solution (5.23), that is without any nullspace contributions, is less structured than any other solution. (It is often, but not always, true that the nullspace vectors are more "wiggily" than those in the range: recall the Neumann problem.) In any case, including any vector not required by the data is arguably producing more structure than is required.) Setting all the unknown α_i to zero is thus one plausible choice. It follows from the orthogonality of the $\mathbf{g}_i$ that this particular solution is also the one of minimum solution norm. Later, we will see some other choices for the nullspace vectors.

If the nullspace vector contributions are set to zero, the true solution has been expanded in an incomplete set of orthonormal vectors. Thus, $\mathbf{G}_K\mathbf{G}_K^T$ is the resolution matrix, and the relationship between the true solution and the minimal one is just,

$$\tilde{\mathbf{x}} = \mathbf{G}_K\mathbf{G}_K^T\mathbf{x} = \mathbf{x} - \mathbf{Q}_G\boldsymbol{\alpha}_G, \quad \tilde{\mathbf{y}} = \mathbf{G}_K\mathbf{G}_K^T\mathbf{y}, \quad \tilde{\mathbf{n}} = \mathbf{Q}_G\mathbf{Q}_G^T\mathbf{y}\,. \tag{5.32}$$

The relative contributions of any structure in $\mathbf{y}$, determined by the projection, $\mathbf{g}_i^T\mathbf{y}$ will depend upon the ratio $\mathbf{g}_i^T\mathbf{y}/\lambda_i$. Comparatively weak values of $\mathbf{g}_i^T\mathbf{y}$ may well be amplified by small, but non-zero, elements of λ_i. One must keep track of both $\mathbf{g}_i^T\mathbf{y}$, and $\mathbf{g}_i^T\mathbf{y}/\lambda_i$.

Before leaving this special case, note one more useful property of the eigenvector/eigenvalues. For the moment, let $\mathbf{G}$ have all its columns, containing both the range and nullspace vectors, with the nullspace vectors being last in arbitrary order. It is thus an $M \times M$ matrix. Correspondingly, let $\boldsymbol{\Lambda}$ contain all the eigenvalues on its diagonal, including the zero ones; it too, is $M \times M$. Then the eigenvector definition (5.6) produces

$$\mathbf{E}\mathbf{G} = \mathbf{G}\boldsymbol{\Lambda}\,. \tag{5.33}$$

Multiply both sides of (5.33) by $\mathbf{G}^T$:

$$\mathbf{G}^T\mathbf{E}\mathbf{G} = \mathbf{G}^T\mathbf{G}\boldsymbol{\Lambda} = \boldsymbol{\Lambda}. \tag{5.34}$$

$\mathbf{G}$ is said to "diagonalize" $\mathbf{E}$. Now multiply both sides of (5.34) on the left by $\mathbf{G}$ and on the right by $\mathbf{G}^T$:

$$\mathbf{G}\mathbf{G}^T\mathbf{E}\mathbf{G}\mathbf{G}^T = \mathbf{G}\boldsymbol{\Lambda}\mathbf{G}^T \tag{5.35}$$

or, using the orthogonality of $\mathbf{G}$,

$$\mathbf{E} = \mathbf{G}\boldsymbol{\Lambda}\mathbf{G}^T\,, \tag{5.36}$$

a useful decomposition of $\mathbf{E}$, consistent with its symmetry.

Recall that $\boldsymbol{\Lambda}$ has zeros on the diagonal corresponding to the zero eigenvalues, and the corresponding rows and columns are entirely zero. Writing out (5.36), these zero rows and columns multiply all the nullspace vector columns of $\mathbf{G}$ by zero, and it is found that the nullspace columns of $\mathbf{G}$ can be eliminated, $\boldsymbol{\Lambda}$ reduced to its $K \times K$ form, and the decomposition (5.36) is still exact—in the form,

$$\mathbf{E} = \mathbf{G}_K\boldsymbol{\Lambda}_K\mathbf{G}_K^T\,. \tag{5.37}$$

Then the simultaneous equations (5.30) are,

$$\mathbf{G}_K\boldsymbol{\Lambda}_K\mathbf{G}_K^T\mathbf{x} + \mathbf{n} = \mathbf{y}\,. \tag{5.38}$$

Left multiply both sides by $\boldsymbol{\Lambda}_K^{-1}\mathbf{G}_K^T$ (existence of the inverse is guaranteed by the removal of the zero eigenvalues) and,

$$\mathbf{G}_K^T\mathbf{x} + \boldsymbol{\Lambda}_K^{-1}\mathbf{G}_K^T\mathbf{n} = \boldsymbol{\Lambda}_K^{-1}\mathbf{G}_K^T\mathbf{y}\,. \tag{5.39}$$

But $\mathbf{G}_K^T\mathbf{x}$ are the projection of $\mathbf{x}$ onto the range vectors of $\mathbf{E}$, and $\mathbf{G}_K^T\mathbf{n}$ is the projection of the noise. We have agreed to set the latter to zero, and obtain,

$$\mathbf{G}_K^T\mathbf{x} = \mathbf{\Lambda}_K^{-1}\mathbf{G}_K^T\mathbf{y}\,,$$

the dot products of the range of $\mathbf{E}$ with the solution. Hence, it must be true, since the range vectors are orthonormal that,

$$\tilde{\mathbf{x}} \equiv \mathbf{G}_K\mathbf{G}_K^T\mathbf{x} \equiv \mathbf{G}_K\mathbf{\Lambda}_K^{-1}\mathbf{G}_K^T\mathbf{y}\,, \tag{5.40}$$

$$\tilde{\mathbf{y}} = \mathbf{E}\tilde{\mathbf{x}} = \mathbf{G}_K\mathbf{G}_K^T\mathbf{y}\,, \tag{5.41}$$

which is identical to the particular solution (5.21). The residuals are

$$\tilde{\mathbf{n}} = \mathbf{y} - \tilde{\mathbf{y}} = \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}} = (\mathbf{I}_L - \mathbf{G}_K\mathbf{G}_K^T)\mathbf{y} = \mathbf{Q}_G\mathbf{Q}_G\mathbf{y}, \tag{5.42}$$

with $\tilde{\mathbf{n}}^T\tilde{\mathbf{y}} = 0$. Notice that matrix $\mathbf{H}$ of Eq. (4.11) is just $\mathbf{G}_K\mathbf{G}_K^T$, and hence $(\mathbf{I}-\mathbf{H})$ is the projector of $\mathbf{y}$ onto the nullspace vectors.

The expected value of the solution (5.21) or (5.40) is,

$$\langle\tilde{\mathbf{x}} - \mathbf{x}\rangle = \mathbf{G}_K\mathbf{\Lambda}_K^{-1}\mathbf{G}_K^T\langle\mathbf{y}\rangle - \sum_{i=1}^{N}\alpha_i\mathbf{g}_i = -\mathbf{Q}_G\boldsymbol{\alpha}_G \tag{5.43}$$

and so the solution is biassed unless $\boldsymbol{\alpha}_G = 0$.

The uncertainty is,

$$\begin{aligned}\mathbf{P} = D^2(\tilde{\mathbf{x}} - \mathbf{x}) &= \langle\mathbf{G}_K\mathbf{\Lambda}_K^{-1}\mathbf{G}_K^T(\mathbf{y}_0 + \mathbf{n} - \mathbf{y}_0)(\mathbf{y}_0 + \mathbf{n} - \mathbf{y}_0)^T\mathbf{G}_K\mathbf{\Lambda}_K^{-1}\mathbf{G}_K^T\rangle \\ &\quad + \langle\mathbf{Q}_G\boldsymbol{\alpha}_G\boldsymbol{\alpha}_G^T\mathbf{Q}_G^T\rangle \\ &= \mathbf{G}_K\mathbf{\Lambda}_K^{-1}\mathbf{G}_K^T\langle\mathbf{n}\mathbf{n}^T\rangle\mathbf{G}_K\mathbf{\Lambda}_K^{-1}\mathbf{G}_K^T + \mathbf{Q}_G\langle\boldsymbol{\alpha}_G\boldsymbol{\alpha}_G^T\rangle\mathbf{Q}_G^T \\ &= \mathbf{G}_K\mathbf{\Lambda}_K^{-1}\mathbf{G}_K^T\mathbf{R}_{nn}\mathbf{G}_K\mathbf{\Lambda}_K^{-1}\mathbf{G}_K^T + \mathbf{Q}_G\mathbf{R}_{\alpha\alpha}\mathbf{Q}_G^T \\ &= \mathbf{C}_{xx} + \mathbf{Q}_G\mathbf{R}_{\alpha\alpha}\mathbf{Q}_G^T\,,\end{aligned} \tag{5.44}$$

defining the second moments, $\mathbf{R}_{\alpha\alpha}$, of the coefficients of the nullspace vectors. Under the special circumstances that the residuals, $\mathbf{n}$, are white noise, with $\mathbf{R}_{nn} = \sigma_n^2\mathbf{I}$, (5.44) reduces to

$$\mathbf{P} = \sigma_n^2\mathbf{G}_K\mathbf{\Lambda}_K^{-2}\mathbf{G}_K^T + \mathbf{Q}_G\mathbf{R}_{\alpha\alpha}\mathbf{Q}_G^T\,. \tag{5.45}$$

Either case shows that the uncertainty of the minimal solution is made up of two distinct parts. The first part, the solution covariance, $\mathbf{C}_{xx}$, arises owing to the noise present in the observations, and generates uncertainty in the coefficients of the range vectors; the second contribution arises from the "missing" nullspace vector contribution. Either term can dominate. The magnitude of the noise term depends largely upon the ratio of the noise variance, σ_n^2, to the smallest non-zero singular value, λ_K^2.

Example 2. *Suppose*

$$\mathbf{Ax}=\mathbf{y},$$
$$\left\{\begin{array}{cc}1 & 1\\ 1 & 1\end{array}\right\}\left[\begin{array}{c}x_1\\ x_2\end{array}\right] = \mathbf{y}=\left[\begin{array}{c}1\\ 3\end{array}\right], \tag{5.46}$$

which is clearly inconsistent and has no solution in the conventional sense. **A** *is a square symmetric matrix. Solving,*

$$\mathbf{Ag}_i=\lambda_i\mathbf{g}_i \tag{5.47}$$

or

$$\left\{\begin{array}{cc}1-\lambda & 1\\ 1 & 1-\lambda\end{array}\right\}\left[\begin{array}{c}g_{i1}\\ g_{i2}\end{array}\right]=\left[\begin{array}{c}0\\ 0\end{array}\right]. \tag{5.48}$$

These equations require

$$g_{i1}\left[\begin{array}{c}1-\lambda\\ 1\end{array}\right]+g_{i2}\left[\begin{array}{c}1\\ 1-\lambda\end{array}\right]=\left[\begin{array}{c}0\\ 0\end{array}\right]$$

or,

$$\left[\begin{array}{c}1-\lambda\\ 1\end{array}\right]+\frac{g_{i2}}{g_{i1}}\left[\begin{array}{c}1\\ 1-\lambda\end{array}\right]=\left[\begin{array}{c}0\\ 0\end{array}\right].$$

Which is

$$\begin{aligned}(1-\lambda)+\frac{g_{i2}}{g_{i1}} &= 0\\ 1+\frac{g_{i2}}{g_{i1}}(1-\lambda) &= 0\end{aligned}$$

and solving for g_{i2}/g_{i1} produces, $\lambda=2,0$. This method, which can be generalized, in effect derives the usual statement that for Eq. (5.48) to have a solution, the determinant,

$$\left|\begin{array}{cc}1-\lambda & 1\\ 1 & 1-\lambda\end{array}\right|,$$

must vanish. The first solution is labelled λ_1, and substituting back in produces $\mathbf{g}_1=\frac{1}{\sqrt{2}}\left[\begin{array}{c}1\\ 1\end{array}\right]$, when given unit length. Also $\mathbf{g}_2=\frac{1}{\sqrt{2}}\left[\begin{array}{c}-1\\ 1\end{array}\right]$, $\lambda_2=0$. Hence,

$$\mathbf{A}=\frac{1}{\sqrt{2}}\left[\begin{array}{c}1\\ 1\end{array}\right]2\frac{1}{\sqrt{2}}\left[\begin{array}{c}1\\ 1\end{array}\right]^T=\left[\begin{array}{c}1\\ 1\end{array}\right]\left[\begin{array}{cc}1 & 1\end{array}\right] \tag{5.49}$$

The equations have no solution in the conventional sense. There is however, a sensible "best" solution:

$$\begin{aligned}\tilde{\mathbf{x}} &= \frac{\mathbf{g}_1^T\mathbf{y}}{\lambda_1}\mathbf{g}_1+\alpha_2\mathbf{g}_2, &(5.50)\\ &= \left(\frac{4}{2\sqrt{2}}\right)\frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}+\alpha_2\frac{1}{\sqrt{2}}\begin{bmatrix}-1\\1\end{bmatrix} &(5.51)\\ &= \begin{bmatrix}1\\1\end{bmatrix}+\alpha_2\frac{1}{\sqrt{2}}\begin{bmatrix}-1\\1\end{bmatrix}. &(5.52)\end{aligned}$$

Notice that

$$\mathbf{A}\tilde{\mathbf{x}}=\begin{bmatrix}2\\2\end{bmatrix}+0\neq\begin{bmatrix}1\\3\end{bmatrix}. \tag{5.53}$$

The solution has compromised the inconsistency. No choice of α_2 can reduce the residual in $\mathbf{y}$. *The equations would more sensibly have been written*

$$\mathbf{Ax}+\mathbf{n}=\mathbf{y},$$

and the difference, $\mathbf{n}=\mathbf{y}-\mathbf{A}\tilde{\mathbf{x}}$ is proportional to $\mathbf{g}_2$. There are clearly an infinite number of choices of $\tilde{\mathbf{x}}$; one choice is often the one with $\alpha_2=0$, the solution of minimum norm. In one sense, it is the simplest solution. A system like (5.46) would most likely arise from measurements (if both equations are divided by 2, they represent two measurements of the average of x_1, x_2), and $\mathbf{n}$ would be best regarded as the noise of observation.

Example 3. *Suppose the same problem as in Example 2 is solved using Lagrange multipliers, that is, minimizing,*

$$J=\mathbf{n}^T\mathbf{n}+\alpha^2\mathbf{x}^T\mathbf{x}-\mathbf{2}\boldsymbol{\mu}^T\left(\mathbf{y}-\mathbf{Ax}-\mathbf{n}\right).$$

Then, the normal equations are

$$\begin{aligned}\frac{1}{2}\frac{\partial J}{\partial\mathbf{x}} &= \alpha^2\mathbf{x}+\mathbf{A}^T\boldsymbol{\mu}=\mathbf{0}\\ \frac{1}{2}\frac{\partial J}{\partial\mathbf{n}} &= \mathbf{n}+\boldsymbol{\mu}=\mathbf{0}\\ \frac{1}{2}\frac{\partial J}{\partial\boldsymbol{\mu}} &= \mathbf{y}-\mathbf{Ax}-\mathbf{n}=\mathbf{0},\end{aligned}$$

which easily produces,

$$\begin{aligned}\tilde{\mathbf{x}} &= \mathbf{A}^T\left(\mathbf{AA}^T+\boldsymbol{\alpha}^2\mathbf{I}\right)^{-1}\mathbf{y}\\ &= \left\{\begin{matrix}2.0 & 2.0\\ 2.0 & 2.0\end{matrix}\right\}\left\{\left\{\begin{matrix}2.0 & 2.0\\ 2.0 & 2.0\end{matrix}\right\}+\alpha^2\left\{\begin{matrix}1.0 & 0\\ 0 & 1.0\end{matrix}\right\}\right\}^{-1}\begin{bmatrix}1\\3\end{bmatrix}.\end{aligned}$$

The limit $\alpha^2 \to \infty$ *is readily evaluated. Letting* $\alpha^2 \to 0$ *involves inverting a singular matrix. To understand what is going on, let us use,*

$$\mathbf{A} = \mathbf{G}\mathbf{\Lambda}\mathbf{G}^T = \mathbf{g}_1\lambda_1\mathbf{g}_1^T + 0 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix} 2\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}^T \tag{5.54}$$

Hence,

$$\mathbf{A}\mathbf{A}^T = \mathbf{G}\mathbf{\Lambda}^2\mathbf{G}^T$$

Note that the full $\mathbf{G}, \mathbf{\Lambda}$ *are being used. Note also that* $\mathbf{I} = \mathbf{G}\mathbf{G}^T$

$$\left(\mathbf{A}\mathbf{A}^T{+}\alpha^2\mathbf{I}\right) = \left(\mathbf{G}\mathbf{\Lambda}^2\mathbf{G}^T + \mathbf{G}\left(\alpha^2\right)\mathbf{G}^T\right) = \mathbf{G}\left(\mathbf{\Lambda}^2 + \alpha^2\mathbf{I}\right)\mathbf{G}^T.$$

By inspection, the inverse of this last matrix is necessarily,

$$\left(\mathbf{A}\mathbf{A}^T{+}\mathbf{I}/\alpha^2\right)^{-1} = \mathbf{G}\left(\mathbf{\Lambda}^2 + \alpha^2\mathbf{I}\right)^{-1}\mathbf{G}^T.$$

But,

$$\left(\mathbf{\Lambda}^2 + \alpha^2\mathbf{I}\right)^{-1} = \operatorname{diag}\left\{1/\left(\lambda_i^2 + \alpha^2\right)\right\}$$

Then

$$\begin{aligned}
\tilde{\mathbf{x}} &= \mathbf{A}^T\left(\mathbf{A}\mathbf{A}^T{+}\alpha^2\mathbf{I}\right)^{-1}\mathbf{y} = \mathbf{G}\mathbf{\Lambda}\mathbf{G}^T\left(\mathbf{G}\operatorname{diag}\left\{1/\left(\lambda_i^2+\alpha^2\right)\right\}\mathbf{G}^T\right)\mathbf{y} \\
&= \mathbf{G}\operatorname{diag}\left\{\lambda_i/\left(\lambda_i^2+\alpha^2\right)\right\}\mathbf{G}^T\mathbf{y} \\
&= \sum_{i=1}^{K}\mathbf{g}_i\frac{\lambda_i}{\lambda_i^2+\alpha^2}\mathbf{g}_i^T\mathbf{y} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}\frac{2}{2+\alpha^2}\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}^T\begin{bmatrix} 1 \\ 3 \end{bmatrix} + 0 \\
&= \frac{4}{2+\alpha^2}\begin{bmatrix} 1 \\ 1 \end{bmatrix}
\end{aligned}$$

The solution always exists as long as $\alpha^2 > 0$*. It is a tapered-down form of the solution with* $\alpha^2 = 0$ *if all* $\lambda_i \neq 0$*.*

$$\mathbf{n} = \begin{bmatrix} 1 \\ 3 \end{bmatrix} - \frac{4}{2+\alpha^2}\mathbf{A}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix} - \frac{4}{2+\alpha^2}\begin{bmatrix} 2.0 \\ 2.0 \end{bmatrix}$$

so that $\alpha^2 \to \infty$*, the solution* $\tilde{\mathbf{x}}$ *is minimized, becoming 0 and the residual is equal to* $\mathbf{y}$*.*

5.3. Arbitrary Systems.

5.3.1. *The Singular Vector Expansion and Singular Value Decomposition.* It may be objected that this entire development is of little use, because most problems, including those outlined in Chapter 1, produced $\mathbf{E}$ matrices which could not be guaranteed to have complete orthonormal sets of eigenvectors. Indeed, the problems considered produce matrices which are usually non-square, and for which the eigenvector problem is not even defined.

For arbitrary *square* matrices, the question of when a complete orthonormal set of eigenvectors exists is not difficult to answer, but becomes somewhat elaborate.[30] When a square matrix of dimension N is not symmetric, one must consider cases in which there are some distinct eigenvalues and where some are repeated, and the general approach requires the so-called Jordan form. But we will next find a way to avoid these intricacies, and yet deal with sets of simultaneous equations of arbitrary dimensions, not just square ones. Although the mathematics are necessarily somewhat more complicated than is employed in solving the just-determined simultaneous linear equations using a complete orthonormal eigenvector set, this latter problem provides full analogues to all of the issues in the more general case, and the reader will probably find it helpful to refer back to this situation for insight.

Consider the possibility, suggested by the eigenvector method, of expanding the solution $\mathbf{x}$ in a set of orthonormal vectors. Eq. (4.2) involves one vector, $\mathbf{x}$, of dimension N, and two vectors, $\mathbf{y}$, $\mathbf{n}$, of dimension M. We would like to use spanning orthonormal vectors, but cannot expect, with two different vector dimensions involved, to use just one set: $\mathbf{x}$ can be expanded exactly in N, N-dimensional orthonormal vectors; and similarly, $\mathbf{y}$ and $\mathbf{n}$ can be exactly represented in M, M-dimensional orthonormal vectors. There are an infinite number of ways to select two such sets. But using the structure of $\mathbf{E}$, a particularly useful pair can be identified.

The simple development leading to the discussion of the solutions in the square, symmetric case resulted from the theorem about the complete nature of the eigenvectors of such a matrix. So let us construct a new matrix,

$$\mathbf{B} = \begin{Bmatrix} \mathbf{0} & \mathbf{E}^T \\ \mathbf{E} & \mathbf{0} \end{Bmatrix}, \tag{5.55}$$

which by definition is square (dimension $M+N$ by $M+N$) and symmetric. Thus, $\mathbf{B}$ satisfies the theorem just alluded to, and the eigenvalue problem,

$$\mathbf{B}\mathbf{q}_i = \lambda_i \mathbf{q}_i \tag{5.56}$$

[30] Brogan (1985) has a succinct discussion.

will give rise to $M+N$ orthonormal eigenvectors $\mathbf{q}_i$ (an orthonormal spanning set) whether or not the λ_i are distinct or non-zero. Writing out ((5.56)),

$$\left\{\begin{matrix} \mathbf{0} & \mathbf{E}^T \\ \mathbf{E} & \mathbf{0} \end{matrix}\right\} \begin{bmatrix} q_{1i} \\ \cdot \\ q_{Ni} \\ q_{N+1,i} \\ \cdot \\ q_{N+M,i} \end{bmatrix} = \lambda_i \begin{bmatrix} q_{1i} \\ \cdot \\ q_{Ni} \\ q_{N+1,i} \\ \cdot \\ q_{N+M,i} \end{bmatrix}, \tag{5.57}$$

where q_{pi} is the pth element of $\mathbf{q}_i$. Taking note of the zero matrices, (5.57) may be rewritten,

$$\mathbf{E}^T \begin{bmatrix} q_{N+1,i} \\ \cdot \\ q_{N+M,i} \end{bmatrix} = \lambda_i \begin{bmatrix} q_{1i} \\ \cdot \\ q_{Ni} \end{bmatrix}, \tag{5.58}$$

$$\mathbf{E} \begin{bmatrix} q_{1i} \\ \cdot \\ q_{Ni} \end{bmatrix} = \lambda_i \begin{bmatrix} q_{N+1,i} \\ \cdot \\ q_{N+M,i} \end{bmatrix}. \tag{5.59}$$

Define,

$$\mathbf{u}_i = \begin{bmatrix} q_{N+1,i} \\ \cdot \\ q_{N+M,i} \end{bmatrix}, \qquad \mathbf{v}_i = \begin{bmatrix} q_{1i} \\ \cdot \\ q_{Ni} \end{bmatrix} \text{ or, } \mathbf{q}_i = \begin{bmatrix} \mathbf{v}_i \\ \mathbf{u}_i \end{bmatrix}, \tag{5.60}$$

that is, defining the first N elements of $\mathbf{q}_i$ to be called $\mathbf{v}_i$ and the last M to be called $\mathbf{u}_i$. Then (5.58)–(5.59) are

$$\mathbf{E}\mathbf{v}_i = \lambda_i \mathbf{u}_i \tag{5.61}$$

$$\mathbf{E}^T\mathbf{u}_i = \lambda_i \mathbf{v}_i\,. \tag{5.62}$$

If (5.61) is left multiplied by $\mathbf{E}^T$, and using (5.62), one has,

$$\mathbf{E}^T\mathbf{E}\mathbf{v}_i = \lambda_i^2 \mathbf{v}_i\,. \tag{5.63}$$

Similarly, left multiplying (5.62) by $\mathbf{E}$ and using (5.61) produces,

$$\mathbf{E}\mathbf{E}^T\mathbf{u}_i = \lambda_i^2 \mathbf{u}_i\,. \tag{5.64}$$

These last two equations show that the $\mathbf{u}_i$, $\mathbf{v}_i$ each separately satisfy two independent eigenvector/eigenvalue problems of the square symmetric matrices $\mathbf{E}\mathbf{E}^T$, $\mathbf{E}^T\mathbf{E}$. If one of M, N is much smaller than the other, only the smaller one must be solved for either of $\mathbf{u}_i$, $\mathbf{v}_i$; the other set is immediately calculated from (5.61) or (5.62). Evidently, in the limiting case, of either a single equation or a single unknown, the eigenvalue/eigenvector problem is completely trivial, involving a pure scalar, no matter how large is the other dimension.

The $\mathbf{u}_i$, $\mathbf{v}_i$ are called "singular vectors," and the λ_i are the "singular values." By convention, the λ_i are ordered in decreasing numerical value. Also by convention, they are all non-negative (taking the negative values of λ_i produces singular vectors differing only by a sign from those corresponding to the positive roots, and thus they are not independent vectors). Equations (5.61)–(5.62) provide a relationship between each $\mathbf{u}_i$ and each $\mathbf{v}_i$. But because in general, $M \neq N$, there will be more of one set than another. The only way equations (5.61)–(5.62) can be consistent is if $\lambda_i = 0$, $i > \min(M, N)$ (where $\min(M, N)$ is read as "the minimum of M and N"). Suppose $M < N$. Then (5.64) is solved for $\mathbf{u}_i$, $1 \leq i \leq M$, and (5.61) is used to find the corresponding $\mathbf{v}_i$. There are $N - M$ $\mathbf{v}_i$ not generated this way, but which can be found using the Gram-Schmidt method described on page 20.

Let there be K non-zero λ_i; then

$$\mathbf{E}\mathbf{v}_i \neq 0\,, \quad 1 \leq i \leq K\,. \tag{5.65}$$

These $\mathbf{v}_i$ are known as the "range vectors of $\mathbf{E}$" or the "solution range vectors." For the remaining $N - K$ vectors $\mathbf{v}_i$,

$$\mathbf{E}\mathbf{v}_i = 0\,, \quad K+1 \leq i \leq N\,, \tag{5.66}$$

known as the "nullspace vectors of $\mathbf{E}$" or the "nullspace of the solution." If $K < M$, there will be K of the $\mathbf{u}_i$ such that,

$$\mathbf{E}^T\mathbf{u}_i \neq 0\,, \quad 1 \leq i \leq K\,, \tag{5.67}$$

which are the "range vectors of $\mathbf{E}^T$" and $M - K$ of the $\mathbf{u}_i$ such that

$$\mathbf{E}^T\mathbf{u}_i = 0\,, \quad K+1 \leq i \leq M\,, \tag{5.68}$$

the "nullspace vectors of $\mathbf{E}^T$" or the "data, or observation, nullspace vectors." The "nullspace" of $\mathbf{E}$ is spanned by its nullspace vectors, the "range" of $\mathbf{E}$ is spanned by the range vectors, etc., in the sense, for example, that an arbitrary vector lying in the range is perfectly described by a sum of the range vectors. We now have two complete orthonormal sets in the two different spaces. Note that (5.66) implies that

$$\mathbf{u}_i^T\mathbf{E} = 0, \quad K+1 \leq i \leq N, \tag{5.69}$$

expressing relationships among the columns of $\mathbf{E}$ in the same way that $\mathbf{v}_i$ expresses relations among its rows.

Because the $\mathbf{u}_i$, $\mathbf{v}_i$ are complete in their corresponding spaces, we can expand $\mathbf{x}$, $\mathbf{y}$, $\mathbf{n}$ without error:

$$\mathbf{x} = \sum_{i=1}^{N} \alpha_i \mathbf{v}_i\,, \quad \mathbf{y} = \sum_{j=1}^{M} \beta_i \mathbf{u}_i\,, \quad \mathbf{n} = \sum_{i=1}^{M} \gamma_i \mathbf{u}_i\,, \tag{5.70}$$

where $\mathbf{y}$ has been measured, so that we know $\beta_j = \mathbf{u}_j^T\mathbf{y}$. To find $\mathbf{x}$, we need α_i, and to find $\mathbf{n}$, we need the γ_i. Substitute (5.70) into the equations (4.2), and using (5.61)–(5.62),

$$\begin{aligned}\sum_{i=1}^{N}\alpha_i\mathbf{E}\mathbf{v}_i + \sum_{i=1}^{M}\gamma_i\mathbf{u}_i &= \sum_{i=1}^{K}\alpha_i\lambda_i\mathbf{u}_i + \sum_{i=1}^{M}\gamma_i\mathbf{u}_i \qquad (5.71)\\ &= \sum_{i=1}^{M}(\mathbf{u}_i^T\mathbf{y})\mathbf{u}_i\,.\end{aligned}$$

Notice the differing upper limits on the summations. Because of the orthonormality of the singular vectors, (5.71) can be solved as,

$$\alpha_i\lambda_i + \gamma_i = \mathbf{u}_i^T\mathbf{y}\,, \quad i = 1 \text{ to } M\,, \tag{5.72}$$

$$\alpha_i = (\mathbf{u}_i^T\mathbf{y} - \gamma_i)/\lambda_i\,, \quad \lambda_i \neq 0\,,\ 1 \le i \le K\,. \tag{5.73}$$

In these equations, if $\lambda_i \neq 0$, nothing prevents setting $\gamma_i = 0$, that is,

$$\mathbf{u}_i^T\mathbf{n} = 0\,, \quad 1 \le i \le K \tag{5.74}$$

should we wish, and which will have the effect of making the noise norm as small as possible (there is obviously a degree of arbitrariness in this choice, and later we will choose γ_i differently). Then (5.73) produces,

$$\alpha_i = \frac{\mathbf{u}_i^T\mathbf{y}}{\lambda_i}\,, \quad 1 \le i \le K\,. \tag{5.75}$$

But, because $\lambda_i = 0$, $i > K$, the only solution to (5.72) for these values of i is $\gamma_i = \mathbf{u}_i^T\mathbf{y}$, and α_i is indeterminate. These γ_i are non-zero, except in the event (unlikely with real data) that,

$$\mathbf{u}_i^T\mathbf{y} = 0\,, \quad K+1 \le i \le N\,. \tag{5.76}$$

This last equation is a solvability condition— in direct analogy to (5.20).

The solution obtained in this manner now has the following form:

$$\tilde{\mathbf{x}} = \sum_{i=1}^{K}\frac{\mathbf{u}_i^T\mathbf{y}}{\lambda_i}\mathbf{v}_i + \sum_{i=K+1}^{N}\alpha_i\mathbf{v}_i \tag{5.77}$$

$$\tilde{\mathbf{y}} = \mathbf{E}\tilde{\mathbf{x}} = \sum_{i=1}^{K}(\mathbf{u}_i^T\mathbf{y})\mathbf{u}_i \tag{5.78}$$

$$\tilde{\mathbf{n}} = \sum_{i=K+1}^{M}(\mathbf{u}_i^T\mathbf{y})\mathbf{u}_i\,. \tag{5.79}$$

Attention is called to the differing summation limits.

The coefficients of the last $N-K$ of the $\mathbf{v}_i$ in Eq. (5.77), the solution nullspace vectors, are arbitrary, representing structures in the solution about which the equations provide no information. A nullspace is always present unless $K=N$. The solution residuals are directly proportional to the nullspace vectors of $\mathbf{E}^T$ and will vanish only if $K=M$, or the solvability conditions are met.

Just as in the simpler square symmetric case, no choice of the coefficients of the solution nullspace vectors can have any effect on the size of the size of the residuals. If we choose once again to exercise Occam's razor, and regard the simplest solution as best, then setting the nullspace coefficients to zero,

$$\tilde{\mathbf{x}} = \sum_{i=1}^{K} \frac{\mathbf{u}_i^T \mathbf{y}}{\lambda_i} \mathbf{v}_i \tag{5.80}$$

along with (5.79), this is the "particular-SVD solution." It minimizes the residuals, and simultaneously produces the corresponding $\tilde{\mathbf{x}}$ with the smallest norm. If $\langle \mathbf{n} \rangle = 0$, the bias of (5.80) is evidently,

$$\langle \tilde{\mathbf{x}} - \mathbf{x} \rangle = -\sum_{i=K+1}^{N} \alpha_i \mathbf{v}_i \,. \tag{5.81}$$

The solution uncertainty is

$$\begin{aligned} \mathbf{P} &= \sum_{i=1}^{K}\sum_{j=1}^{K} \mathbf{v}_i \frac{\mathbf{u}_i^T \mathbf{R}_{nn} \mathbf{u}_j}{\lambda_i \lambda_j} \mathbf{v}_i^T + \sum_{i=K+1}^{N}\sum_{j=K+1}^{N} \mathbf{v}_i \langle \alpha_i \alpha_j \rangle \mathbf{v}_j^T \qquad (5.82) \\ &= \mathbf{C}_{xx} + . \sum_{i=K+1}^{N}\sum_{j=K+1}^{N} \mathbf{v}_i \langle \alpha_i \alpha_j \rangle \mathbf{v}_j^T \end{aligned}$$

If the noise is white with variance σ_n^2 or, if a row-scaling matrix $\mathbf{W}^{-T/2}$ has been applied to make it so, then (5.82) becomes

$$\mathbf{P} = \sum_{i=1}^{K} \frac{\sigma_n^2}{\lambda_i^2} \mathbf{v}_i \mathbf{v}_i^T + \sum_{i=K+1}^{N} \langle \alpha_i^2 \rangle \mathbf{v}_i \mathbf{v}_i^T \tag{5.83}$$

where it was also assumed that $\langle \alpha_i \alpha_j \rangle = \langle \alpha_i^2 \rangle \delta_{ij}$ in the nullspace. The influence of very small singular values on the uncertainty is very clear: In the solution (5.77) or (5.80) there are error terms $\mathbf{u}_i^T \mathbf{n}/\lambda_i$ that are greatly magnified by small or nearly vanishing singular values, introducing large terms proportional to σ_n^2/λ_i^2 into (5.83).

The structures dominating $\tilde{\mathbf{x}}$ are clearly a competition between the magnitudes of $\mathbf{u}_i^T \mathbf{y}$ and λ_i, given by the ratio, $\mathbf{u}_i^T \mathbf{y}/\lambda_i$. Large λ_i can suppress comparatively large projections onto $\mathbf{u}_i$, and similarly, small, but non-zero λ_i may greatly amplify comparatively modest projections. In practice,[31] one is well-advised to study the behavior of both $\mathbf{u}_i^T\mathbf{y}$, $\mathbf{u}_i^T\mathbf{y}/\lambda_i$ as a function of i to understand the nature of the solution.

The decision to omit contributions to the residuals by the range vectors of $\mathbf{E}^T$, as we did in Eqs. (5.74), (5.79) needs to be examined. Should we make some other choice, the $\tilde{\mathbf{x}}$ norm would decrease, but the residual norm would increase. Determining the desirability of such a trade-off requires understanding of the noise structure—in particular, (5.74) imposes rigid structures, and hence covariances, on the residuals.

5.4. The Singular Value Decomposition. The singular vectors and values have been used to provide a convenient pair of orthonormal spanning sets to solve an arbitrary set of simultaneous equations. The vectors and values have another use however, in providing a decomposition of $\mathbf{E}$.

Define $\mathbf{\Lambda}$ as the $M \times N$ matrix whose diagonal elements are the λ_i, in order of descending values in the same order, $\mathbf{U}$ as the $M \times M$ matrix whose columns are the $\mathbf{u}_i$, $\mathbf{V}$ as the $N \times N$ matrix whose columns are the $\mathbf{v}_i$ and whose other elements are 0. As an example, suppose $M = 3$, $N = 4$; then

$$\mathbf{\Lambda} = \left\{\begin{matrix} \lambda_i & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \end{matrix}\right\}.$$

Alternatively, if $M = 4$, $N = 3$

$$\left\{\begin{matrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \\ 0 & 0 & 0 \end{matrix}\right\},$$

therefore extending the definition of a diagonal matrix to non-square ones.

Precisely as with matrix $\mathbf{G}$ considered above, column orthonormality of $\mathbf{U}$, $\mathbf{V}$ implies that these matrices are orthogonal,

$$\mathbf{U}\mathbf{U}^T = \mathbf{I}_M\,, \qquad \mathbf{U}^T\mathbf{U} = \mathbf{I}_M\,, \tag{5.84}$$

$$\mathbf{V}\mathbf{V}^T = \mathbf{I}_N\,, \qquad \mathbf{V}^T\mathbf{V} = \mathbf{I}_N\,. \tag{5.85}$$

(It follows that $\mathbf{U}^{-1} = \mathbf{U}^T$, etc.) As with $\mathbf{G}$ above, should one or more columns of $\mathbf{U}$, $\mathbf{V}$ be deleted, the matrices will become semi-orthogonal.

The relations (5.61) to (5.64) can be written compactly as:

[31] Lawson and Hanson (1974).

$$\mathbf{E}\mathbf{V} = \mathbf{U}\boldsymbol{\Lambda}\,, \qquad \mathbf{E}^T\mathbf{U} = \mathbf{V}\boldsymbol{\Lambda}^T\,, \tag{5.86}$$

$$\mathbf{E}^T\mathbf{E}\mathbf{V} = \mathbf{V}\boldsymbol{\Lambda}^T\boldsymbol{\Lambda}\,, \qquad \mathbf{E}\mathbf{E}^T\mathbf{U} = \mathbf{U}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T\,. \tag{5.87}$$

Left multiply the first of (5.86) by $\mathbf{U}^T$ and right multiply it by $\mathbf{V}^T$, and invoking Eq. (5.85),

$$\mathbf{U}^T\mathbf{E}\mathbf{V}\mathbf{V}^T = \mathbf{U}^T\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}\mathbf{V}^T = \boldsymbol{\Lambda}\,. \tag{5.88}$$

So $\mathbf{U}$, $\mathbf{V}$ diagonalize $\mathbf{E}$ (with "diagonal" having the extended meaning for a rectangular matrix as defined above.)

Right multiplying the first of (5.86) by $\mathbf{V}^T$,

$$\mathbf{E} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^T\,. \tag{5.89}$$

This last equation represents a product, called the "singular value decomposition" (SVD), of an arbitrary matrix, of two orthogonal matrices, $\mathbf{U}$, $\mathbf{V}$, and a usually non-square diagonal matrix, $\boldsymbol{\Lambda}$.

There is one further step to take. Notice that for a rectangular $\boldsymbol{\Lambda}$, as in the examples above, one or more rows or columns must be all zero, depending upon the shape of the matrix. If any of the $\lambda_i = 0$, $i < \min(M, N)$, the corresponding rows or columns will be all zeros. Let K be the number of non-vanishing singular values (the "rank" of $\mathbf{E}$). By inspection (multiplying it out), one finds that the last $N - K$ columns of $\mathbf{V}$ and the last $M - K$ columns of $\mathbf{U}$ are multiplied by zeros only. If these columns are dropped entirely from $\mathbf{U}$, $\mathbf{V}$ so that $\mathbf{U}$ becomes $M \times K$ and $\mathbf{V}$ becomes $N \times K$, and reducing $\boldsymbol{\Lambda}$ to a $K \times K$ square matrix, then the representation (5.89) remains exact, in the form,

$$\mathbf{E} = \mathbf{U}_K\boldsymbol{\Lambda}_K\mathbf{V}_K^T\,, \tag{5.90}$$

with the subscript indicating the number of columns, where $\mathbf{U}_K$, $\mathbf{V}_K$ are then only semi-orthogonal, and $\boldsymbol{\Lambda}_K$ is now square. Eq. (5.90) should be compared to (5.37)).[32]

The SVD solution can be obtained by direct matrix manipulation, rather than vector by vector. Consider once again finding the solution to the simultaneous equations ((4.2)), but first write $\mathbf{E}$ in its reduced SVD,

$$\mathbf{U}_K\boldsymbol{\Lambda}_K\mathbf{V}_K^T\mathbf{x} + \mathbf{n} = \mathbf{y}\,. \tag{5.91}$$

[32] The singular value decomposition for arbitrary non-square matrices is apparently due to the oceanographer Carl Eckart (Eckart & Young, 1939; see the discussion in Haykin, 1986; Klema & Laub, 1980; or Stewart, 1993). A particularly lucid account is given by Lanczos (1961) who however, fails to give the decomposition a name. Other references are Noble and Daniel (1977), Strang (1986) and many recent books on applied linear algebra. The crucial role it plays in inverse methods appears to have been first noticed by Wiggins (1972).

Left multiplying by $\mathbf{U}_K^T$ and invoking the semi-orthogonality of $\mathbf{U}_K$ produces

$$\mathbf{\Lambda}_K \mathbf{V}_K^T \mathbf{x} + \mathbf{U}_K^T \mathbf{n} = \mathbf{U}_K^T \mathbf{y}\,. \tag{5.92}$$

The inverse of $\mathbf{\Lambda}_K$ (square with all non-zero diagonal elements) is easily computed and,

$$\mathbf{V}_K^T \mathbf{x} + \mathbf{\Lambda}_K^{-1} \mathbf{U}_K^T \mathbf{n} = \mathbf{\Lambda}_K^{-1} \mathbf{U}_K^T \mathbf{y}\,. \tag{5.93}$$

But $\mathbf{V}_K^T \mathbf{x}$ is the dot product of the first K of the $\mathbf{v}_i$ with the unknown $\mathbf{x}$. Eq. (5.93) thus represent statements about the relationship between dot products of the unknown vector, $\mathbf{x}$, with a set of orthonormal vectors, and therefore must represent the expansion coefficients of the solution in those vectors. If we set

$$\mathbf{U}_K^T \mathbf{n} = 0\,, \tag{5.94}$$

then

$$\mathbf{V}_K^T \mathbf{x} = \mathbf{\Lambda}_K^{-1} \mathbf{U}_K^T \mathbf{y}\,, \tag{5.95}$$

and hence

$$\tilde{\mathbf{x}} = \mathbf{V}_K \mathbf{\Lambda}_K^{-1} \mathbf{U}_K^T \mathbf{y}\,, \tag{5.96}$$

identical to the solution (5.80), which the reader is urged to confirm by writing it out explicitly. As with the square symmetric case, the contribution of any structure in $\mathbf{y}$ proportional to $\mathbf{u}_i$ depends upon the ratio of the projection, $\mathbf{u}_i^T \mathbf{y}$ to λ_i. Substituting solution (5.96) into (5.91),

$$\mathbf{U}_K \mathbf{\Lambda}_K \mathbf{V}_K^T \mathbf{V}_K \mathbf{\Lambda}_K^{-1} \mathbf{U}_K^T \mathbf{y} + \mathbf{n} = \mathbf{U}_K \mathbf{U}_K^T \mathbf{y} + \mathbf{n} = \mathbf{y}$$

or

$$\tilde{\mathbf{n}} = (\mathbf{I} - \mathbf{U}_K \mathbf{U}_K^T) \mathbf{y}\,. \tag{5.97}$$

Let the full $\mathbf{U}$ and $\mathbf{V}$ matrices be rewritten as

$$\mathbf{U} = \{\mathbf{U}_K \quad \mathbf{Q}_u\}\,,\ \mathbf{V} = \{\mathbf{V}_K \quad \mathbf{Q}_v\} \tag{5.98}$$

where $\mathbf{Q}_u$, $\mathbf{Q}_v$ are the matrices whose columns are the corresponding nullspace vectors. Then,

$$\mathbf{E}\tilde{\mathbf{x}} + \tilde{\mathbf{n}} = \mathbf{y},\ \mathbf{E}\tilde{\mathbf{x}} = \tilde{\mathbf{y}} \tag{5.99}$$

$$\tilde{\mathbf{y}} = \mathbf{U}_K \mathbf{U}_K^T \mathbf{y},\ \tilde{\mathbf{n}} = \mathbf{Q}_u \mathbf{Q}_u^T \mathbf{y} = \sum_{j=K+1}^{N} \left(\mathbf{u}_j^T \mathbf{y}\right) \mathbf{u}_i \tag{5.100}$$

which is identical to (5.78). Note,

$$\mathbf{Q}_u \mathbf{Q}_u^T = (\mathbf{I} - \mathbf{U}_K \mathbf{U}_K^T),\ \mathbf{Q}_v \mathbf{Q}_v^T = (\mathbf{I} - \mathbf{V}_K \mathbf{V}_K^T) \tag{5.101}$$

and which are idempotent. ($\mathbf{V}_K \mathbf{V}_K^T$ is matrix $\mathbf{H}$ of Eq. (4.11)). The two vector sets $\mathbf{Q}_u, \mathbf{Q}_v$ span the data and solution nullspaces respectively. The general solution is,

$$\tilde{\mathbf{x}} = \mathbf{V}_K \mathbf{\Lambda}_K^{-1} \mathbf{U}_K \mathbf{y} + \mathbf{Q}_v \boldsymbol{\alpha}, \tag{5.102}$$

where $\boldsymbol{\alpha}$ is now restricted to being the vector of coefficients of the nullspace vectors.

The solution uncertainty (5.82) is,

$$\begin{aligned} \mathbf{P} = \mathbf{V}\boldsymbol{\Lambda}_K^{-1}\mathbf{U}_K^T \left\langle \mathbf{n}\mathbf{n}^T \right\rangle \mathbf{U}_K \boldsymbol{\Lambda}_K^{-1}\mathbf{V}_K^T \\ + \mathbf{Q}_v \left\langle \alpha\alpha^T \right\rangle \mathbf{Q}_G^T = \mathbf{C}_{xx} + \mathbf{Q}_v \left\langle \boldsymbol{\alpha}\boldsymbol{\alpha}^T \right\rangle \mathbf{Q}_v^T \end{aligned} \tag{5.103}$$

or,

$$\mathbf{P} = \sigma_n^2 \mathbf{V}_K \boldsymbol{\Lambda}_K^{-2} \mathbf{V}_K^T + \mathbf{Q}_v \left\langle \boldsymbol{\alpha}\boldsymbol{\alpha}^T \right\rangle \mathbf{Q}_v^T \tag{5.104}$$

for white noise.

Least-squares solution of simultaneous solutions by SVD has several important advantages. Among other features, we can write down within one algebraic formulation the solution to systems of equations which can be under-, over-, or just-determined. Unlike the eigenvalue/eigenvector solution for the square system, the singular values (eigenvalues) are always non-negative and real, and the singular vectors (eigenvectors) can always be made a complete orthonormal set. Neither of these statements is true for the conventional eigenvector problem. Furthermore, the relations (5.61), (5.62) or (5.86, 5.87) provide a specific, quantitative statement of the connection between a set of orthonormal structures in the data, and the corresponding presence of orthonormal structures in the solution. These relations provide a very powerful diagnostic method for understanding precisely why the solution takes on the form it does.

5.5. Some Simple Examples. Algebraic Equations.

Example 4. *The simplest underdetermined system is* 1×2. *Suppose* $x_1 - 2x_2 = 3$ *so that*

$$\mathbf{E} = \{1 \quad -2\}, \quad \mathbf{U} = \{1\}, \quad \mathbf{V} = \left\{ \begin{matrix} .447 & -.894 \\ -.894 & -.447 \end{matrix} \right\}, \quad \lambda_1 = 2.23$$

where the second column of $\mathbf{V}$ *is in the nullspace of* $\mathbf{E}$. *The general solution is* $\tilde{\mathbf{x}} = [.6, -1.2]^T + \alpha_2 \mathbf{v}_2$. *Because* $K = 1$ *is the only possible choice here, it is readily confirmed that this solution satisfies the equation exactly, and a data nullspace is not possible.*

Example 5. *The most elementary overdetermined problem is* 2×1. *Suppose*

$$x_1 = 1$$
$$x_1 = 3\,.$$

The appearance of two such equations is possible if there is noise in the observations, and they should be written as,

$$x_1 + n_1 = 1$$
$$x_1 + n_2 = 3\,.$$

$\mathbf{E} = \{1,1\}^T$, $\mathbf{E}^T\mathbf{E}$ *represents the eigenvalue problem of the smaller dimension, again 1*×1 *and,*

$$\mathbf{U} = \begin{Bmatrix} .707 & -.707 \\ .707 & .707 \end{Bmatrix}, \quad \mathbf{V} = \{1\}, \quad \lambda_1 = \sqrt{2}$$

where the second column of $\mathbf{U}$ *lies in the data nullspace, there being no solution nullspace. The general solution is* $\mathbf{x} = x_1 = 2$, *which if substituted back into the original equations produces,*

$$\mathbf{E}\tilde{\mathbf{x}} = \begin{bmatrix} 2 \\ 2 \end{bmatrix} = \tilde{\mathbf{y}}\,,$$

and hence there are residuals $\tilde{\mathbf{n}} = \tilde{\mathbf{y}} - \mathbf{y} = [1, -1]^T$, *and which are necessarily proportional to* $\mathbf{u}_2$ *and thus orthogonal to* $\tilde{\mathbf{y}}$. *No other solution can produce a smaller* l_2 *norm residual than this one. The SVD produced a solution which compromised the contradiction between the two original equations.*

Example 6. *The possibility of* $K < M$, $K < N$ *simultaneously is also easily seen. Consider the system:*

$$\begin{Bmatrix} 1 & -2 & 1 \\ 3 & 2 & 1 \\ 4 & 0 & 2 \end{Bmatrix} \mathbf{x} = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix},$$

which appears superficially just-determined. But the singular values are $\lambda_1 = 5.67$, $\lambda_2 = 2.80$, $\lambda_3 = 0$. *The vanishing of the third singular value means that the row and column vectors are not linearly independent sets (not spanning sets)—indeed the third row vector one is just the sum of the first two (but the third element of* $\mathbf{y}$ *is not the sum of the first two—making the equations inconsistent). Thus there are both solution and data nullspaces, which the reader might wish to find. With a vanishing singular value,* $\mathbf{E}$ *can be written exactly using only two columns of* $\mathbf{U}$, $\mathbf{V}$ *and the linear dependence of the equations is given explicitly as* $\mathbf{u}_3^T\mathbf{E} = 0$.

Example 7. *Consider now the underdetermined system,*

$$\begin{aligned} x_1 + x_2 - 2x_3 &= 1 \\ x_1 + x_2 - 2x_3 &= 2\,, \end{aligned}$$

which has no conventional solution at all, being a contradiction, and is thus simultaneously underdetermined and incompatible. If one of the coefficients is modified by a very small quantity, ϵ, *to produce,*

$$\begin{aligned} x_1 + x_2 - (2+\epsilon)x_3 &= 1\,, \\ x_1 + x_2 - 2x_3 &= 2\,, \end{aligned} \tag{5.105}$$

not only is there a solution, there are an infinite number of them, which the reader should confirm by computing the particular SVD solution and the nullspace. Thus the slightest perturbation in *the coefficients has made the system jump from having no solution to having an infinite number, an obviously disconcerting situation. The label for such a system is "ill-conditioned." How would we know the system is ill-conditioned? There are several indicators. First, the ratio of the two singular values is determined by ϵ. In (5.105), if we take $\epsilon = 10^{-10}$, the two singular values are $\lambda_1 = 3.46$, $\lambda_2 = 4.1 \times 10^{-11}$, an immediate warning that the two equations are nearly linearly dependent. (In a mathematical problem the non-vanishing of the second singular value is enough to assure a solution. As will be discussed below, it is the inevitable slight errors in $\mathbf{y}$, which suggest sufficiently small singular values should be treated as though they were actually zero.)*

EXAMPLE 8. *A similar problem exists with the system,*

$$\begin{aligned} x_1 + x_2 - 2x_3 &= 1 \\ x_1 + x_2 - 2x_3 &= 1\,, \end{aligned}$$

which has an infinite number of solutions. But the change to

$$\begin{aligned} x_1 + x_2 - 2x_3 &= 1\,, \\ x_1 + x_2 - 2x_3 &= 1 + \epsilon \end{aligned}$$

for arbitrarily small ϵ produces a system with no solutions in the conventional mathematical sense, although the SVD will handle the system in a sensible way, which the reader should confirm.

Problems like these are simple examples of the practical issues that arise once one recognizes that unlike textbook problems, observational ones always contain inaccuracies; any discussion of how to handle data in the presence of mathematical relations must account for these inaccuracies as intrinsic—not as something to be regarded as an afterthought. But the SVD itself is sufficiently powerful that it always contains the information to warn of ill-conditioning, and by determination of K to cope with it—producing useful solutions.

EXAMPLE 9. *Tomographic Problem from Chapter 1. A square box, made up of 3*×3 *unit*

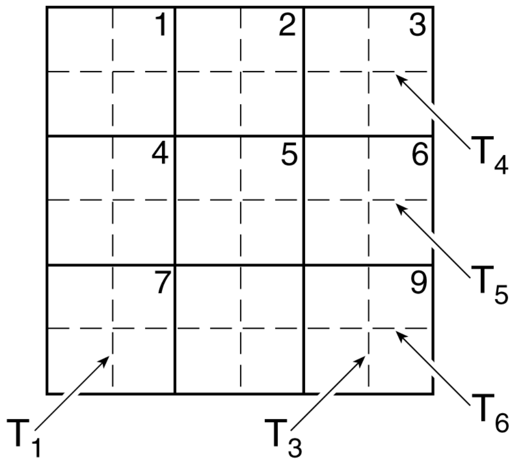


FIGURE 10. Tomographic problem with 9-unknowns and only 6-integral constraints.

dimension boxes. All rays in are in the r_x or r_y directions (Fig. 10) . So the equations are

$$\left\{\begin{array}{ccccccccc} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{array}\right\} \left[\begin{array}{c} x_1 \\ x_2 \\ . \\ . \\ . \\ . \\ . \\ . \\ x_9 \end{array}\right] = \left[\begin{array}{c} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{array}\right]$$

that is, $Ex = y$. The SVD produces (note, the columns here are the reverse of the usual convention—an anomaly of software output),

$$\mathbf{U}^{(r)} =: \left\{\begin{array}{cccccc} -0.408 & 0.0 & 0.0 & 0.816 & 0.0 & 0.408 \\ -0.408 & 0.653 & -1.93\times 10^{-2} & -0.408 & -0.271 & 0.408 \\ -0.408 & -0.653 & 1.93\times 10^{-2} & -0.408 & 0.271 & 0.408 \\ 0.408 & 0.204 & 0.653 & 1.25\times 10^{-19} & 0.446 & 0.408 \\ 0.408 & 0.104 & -0.751 & -2.37\times 10^{-20} & 0.304 & 0.408 \\ 0.408 & -0.308 & 9.83\times 10^{-2} & -6.44\times 10^{-20} & -0.750 & 0.408 \end{array}\right\}$$

,

$$\mathbf{\Lambda}^{(r)} = \left\{ \begin{array}{ccccccccc} 1.96 \times 10^{-19} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1.73 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.73 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.73 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1.73 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2.45 & 0 & 0 & 0 \end{array} \right\}$$

,

$$\mathbf{V}^{(r)} =$$

$$\left\{ \begin{array}{ccccccccc} 0.53 & 0.118 & 0.377 & 0.471 & 0.258 & 0.333 & -0.198 & -0.311 & \\ -0.613 & 0.495 & 0.366 & -0.236 & 0.101 & 0.333 & -0.156 & -0.207 & 4.2 \\ 8.26 \times 10^{-2} & -0.259 & 0.388 & -0.236 & 0.414 & 0.333 & 0.354 & 0.519 & - \\ -0.263 & 5.99 \times 10^{-2} & -0.433 & 0.471 & 0.175 & 0.333 & -0.341 & 0.507 & 4.6 \\ 0.269 & 0.437 & -0.445 & -0.236 & 1.89 \times 10^{-2} & 0.333 & 0.483 & -8.29 \times 10^{-3} & \\ -5.84 \times 10^{-3} & -0.317 & -0.422 & -0.236 & 0.332 & 0.333 & -0.142 & -0.499 & - \\ -0.267 & -0.178 & 5.67 \times 10^{-2} & 0.471 & -0.433 & 0.333 & 0.539 & -0.195 & - \\ 0.344 & 0.199 & 4.56 \times 10^{-2} & -0.236 & -0.589 & 0.333 & -0.327 & 0.216 & - \\ -7.68 \times 10^{-2} & -0.555 & 6.79 \times 10^{-2} & -0.236 & -0.277 & 0.333 & -0.212 & -2.03 \times 10^{-2} & \end{array} \right.$$

so rank $K = 5$. *Notice that there are four repeated* λ_i, *and the lack of simple symmetries in the corresponding* v_i *is a consequence of a random assignment in the eigenvectors.*

u_1 *just averages the right hand-side values, and the corresponding solution is completely uniform. The average of* y *is usually the most robust piece of information.*

The "right" answer is $x = [0, 0, 0, 0, 1, 0, 0, 0, 0]^T$. *The rank 5 answer by SVD is* $\tilde{x}$ *=[-0.1111, 0.2222, -0.1111, 0.2222, 0.5556, 0.2222, -0.1111, 0.2222, -0.1111]*^{T} *which exactly satisfies the same equations.* $\tilde{x}^T \tilde{x} = 0.556 \ < x^T x$. *Written out in two dimensions, this solution is,*

$$\begin{array}{cc} & \quad r_x \rightarrow \\ r_y \uparrow & \left[\begin{array}{ccc} -.11 & .22 & -.11 \\ .22 & .56 & .22 \\ -.11 & .22 & -.11 \end{array} \right] \end{array}. \tag{5.106}$$

This is the minimum norm solution. The sixth v_i, *which actually belongs in the null space is,*

$$\left[\begin{array}{ccc} .53 & -.61 & .01 \\ -.26 & .27 & 0 \\ -.27 & .34 & -.01 \end{array} \right]$$

and one of the remaining null-space vectors is,

$$\begin{bmatrix} -.20 & -.16 & .35 \\ -.34 & .48 & -.14 \\ .54 & -.33 & -.21 \end{bmatrix} \tag{5.107}$$

both of which, along with all the other null space vectors, are readily confirmed to produce a zero sum along any of the ray paths. u_6 *is in the data nullspace.* $u_6^T\ E = 0$ *shows that,*

$$a\left(y_1 + y_2 + y_3\right) - a\left(y_4 + y_5 + y_6\right) = 0,$$

if there is to be a solution without a residual, or alternatively, that no solution would permit this sum to be non-zero. This requirement is physically sensible, as it says that the vertical and horizontal rays cover the same territory and must therefore produce the same sum travel times. It explains why the rank is 5, and not 6.

There is no noise. The correct solution and the SVD solution differ by the null space vectors. One can easily confirm that $\tilde{x}$ *is row (or column) 5 of* $V_5V_5^T$*. Least-squares allows one to minimize (or maximize) anything one pleases. Suppose for some reason, we want the solution that minimizes the differences between the value in box 5 and its neighbors, perhaps as a way of finding a "smooth" solution. Let*

$$\mathbf{W} = \left\{ \begin{array}{ccccccccc} -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{array} \right\} \tag{5.108}$$

The last row is included to render W *a full-rank matrix. Then*

$$\mathbf{Wx} = \begin{bmatrix} x_5 - x_1 \\ x_5 - x_2 \\ . \\ x_5 - x_9 \\ x_5 \end{bmatrix} \tag{5.109}$$

and we can minimize

$$J = \mathbf{x}^T\mathbf{W}^T\mathbf{Wx} \tag{5.110}$$

subject to $\mathbf{Ex}=\mathbf{y}$ *by finding the stationary value of*

$$J' = J - 2\boldsymbol{\mu}^T\left(\mathbf{y}-\mathbf{Ex}\right) \tag{5.111}$$

The normal equations are then

$$\mathbf{W}^T\mathbf{W}\mathbf{x} = \mathbf{E}^T\boldsymbol{\mu} \tag{5.112}$$

$$\mathbf{Ex} = \mathbf{y} \tag{5.113}$$

and

$$\tilde{\mathbf{x}} = \left(\mathbf{W}^T\mathbf{W}\right)^{-1}\mathbf{E}^T\boldsymbol{\mu}$$

and then,

$$\mathbf{E}\left(\mathbf{W}^T\mathbf{W}\right)^{-1}\mathbf{E}^T\boldsymbol{\mu} = \mathbf{y}$$

The rank of $\mathbf{E}\left(\mathbf{W}^T\mathbf{W}\right)^{-1}\mathbf{E}^T$ *is* $K=5<M=6$, *and so we need a generalized inverse,*

$$\tilde{\boldsymbol{\mu}} = \left(\mathbf{E}\left(\mathbf{W}^T\mathbf{W}\right)^{-1}\mathbf{E}^T\right)^{+}\mathbf{y} = \sum_{j=1}^{5}\mathbf{g}_i\frac{\mathbf{g}_i^T\mathbf{y}}{\lambda_i}$$

The null space of $\mathbf{E}\left(\mathbf{W}^T\mathbf{W}\right)^{-1}\mathbf{E}^T$ *is readily confirmed to be the vector,*

$$\left[\begin{array}{cccccc} -0.408 & -0.408 & -0.408 & 0.408 & 0.408 & 0.408 \end{array}\right]^T, \tag{5.114}$$

which produces the solvability condition. Here, because $\mathbf{E}\left(\mathbf{W}^T\mathbf{W}\right)^{-1}\mathbf{E}^T$ *is symmetric, the SVD reduces to the symmetric decomposition.*

Finally, the mapped $\tilde{\mathbf{x}}$ *is*

$$\left[\begin{array}{ccc} -.20 & .41 & -.20 \\ .41 & .18 & .41 \\ -.20 & .41 & -.21 \end{array}\right]$$

and one cannot further decrease the sum-squared difference. One can confirm that this solution satisfies the equations. Evidently, this is a minimum, not a maximum (it suffices to show that the eigenvalues of $\mathbf{W}^T\mathbf{W}$ *are all positive). The addition of any of the nullspace vectors of* $\mathbf{E}$ *to* $\tilde{\mathbf{x}}$ *will necessarily increase the value of* J *and hence there is no bounded maximum. In real tomographic problems, the arc lengths making up matrix* $\mathbf{E}$ *are three dimensional curves and depend upon the background index of refraction in the medium, and which is usually itself determined from observations.*[33] *There are thus errors in* $\mathbf{E}$ *itself, rendering the problem one of* non-linear *estimation. Approaches to solving such problems are described in Chapter 3.*

[33] Munk et al. (1996).

EXAMPLE 10. *Consider, the flow into a four-sided box (Chapter 1, Fig. 5) with missing integration constant as described there. Total mass conservation and conservation of dye,* C_i. *Let the relative areas of each interface be 1, 2, 3, 1 units respectively. Let velocities on each side be* $1, 1/2, -2/3$,0 *respectively, with the minus sign indicating a flow out. That mass is conserved is confirmed by*

$$1\left(1\right)+2\left(\frac{1}{2}\right)+3\left(\frac{-2}{3}\right)+1\left(0\right)=0$$

Now we suppose that the total velocity is not in fact known, but an integration constant is missing on each interface, so that

$$1\left(\frac{1}{2}+b_1\right)+2\left(1+b_2\right)+3\left(\frac{1}{3}+b_3\right)+1\left(2+b_4\right)=0$$

where the $b_i = [1/2, -1/2, -1, -2]$ *respectively, but are unknown. Then the above equation becomes*

$$b_1+2b_2+3b_3+b_4=-5.5$$

or one equation in 4 *unknowns. Evidently, one linear combination of the unknown* b_i *can be determined. We would like more information. Suppose that a tracer of concentration,* $C_i = [2, 1, 3/2, 0]$ *is measured at each side, and is believed conserved. The governing equation is*

$$1\left(\frac{1}{2}+b_1\right)2+2\left(1+b_2\right)1+3\left(\frac{1}{3}+b_3\right)\frac{3}{2}+1\left(2+b_4\right)0=0$$

or

$$2b_1+2b_2+4.5b_3+0b_4=-4.5$$

giving a system of 2 equations in four unknowns

$$\left\{\begin{array}{cccc}1 & 2 & 3 & 1\\ 2 & 2 & 4.5 & 0\end{array}\right\}\left[\begin{array}{c}b_1\\ b_2\\ b_3\\ b_4\end{array}\right]=\left[\begin{array}{c}-5.5\\ -4.5\end{array}\right].$$

The SVD of the coefficient matrix, $\mathbf{E}$, *is :*

$$\mathbf{E}=\left\{\begin{array}{cc}-0.582 & -0.813\\ 0.813 & 0.582\end{array}\right\}\left\{\begin{array}{cccc}6.50 & 0 & 0 & 0\\ 0 & 1.02 & 0 & 0\end{array}\right\}\left\{\begin{array}{cccc}-.801 & 0.179 & -0.454 & 0.347\\ 0.009 & 0.832 & 0.429 & 0.340\\ -0.116 & 0.479 & -0.243 & -0.835\\ 0.581 & 0.215 & -0.742 & 0.259\end{array}\right\}$$

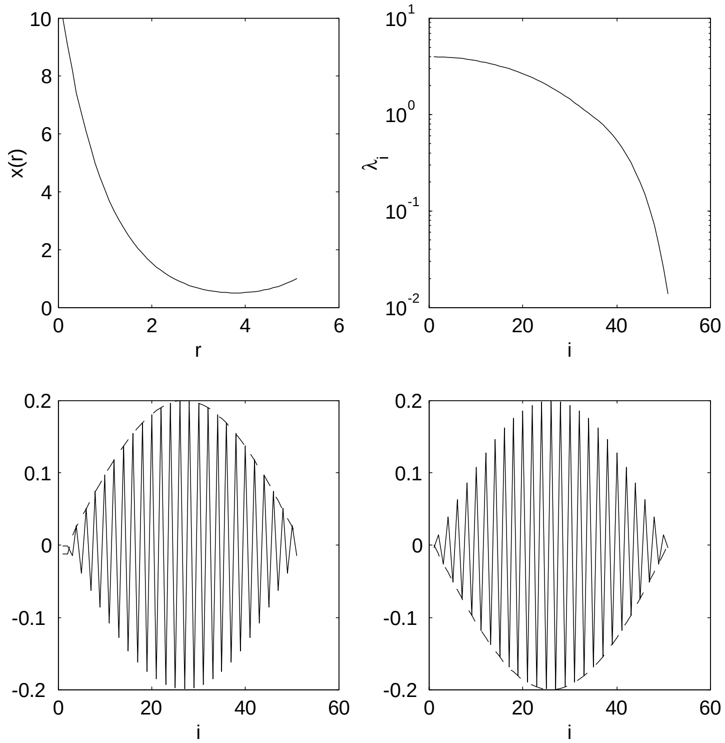


FIGURE 11. Upper left is numerical solution to Eq. (5.116) by direct solution of the simultaneous equations. Upper right panel displays the corresponding singular values. All are finite (there is no nullspace). Lower left panel idsplays $\mathbf{u}_1$ (solid curve) and $\mathbf{u}_{51}$ (dashed curve). Lower right panel shows the corresponding $\mathbf{v}_1, \mathbf{v}_{51}$. The most robust information is the *absence* of small scales in the solution.

5.6. Simple Examples. Differential and Partial Differential Equations.

EXAMPLE 11. *As an example of the use of this machinery with differential equations, consider the simple equation,*

$$\frac{d^2 x(r)}{dr^2} - k^2 x(r) = 0, \tag{5.115a}$$

subject initial and/or boundary condition. Using a simple uniform discretization, we have

$$x((m+1)\Delta r) - \left(2 + k(\Delta r)^2\right) x(m\Delta r) + x((m-1)\Delta r) = 0, \tag{5.116}$$

at all interior points. Taking the specific case, $x(\Delta r) = 10, x(51\Delta r) = 1, \Delta r = Dr = 0.1,$ *the numerical solution is depicted in Fig. 11 from the direct (conventional) solution to* $\mathbf{Ax} = \mathbf{y}$. *The first two rows of* $\mathbf{A}$ *were used to impose the boundary conditions on* $x(\Delta r), x(51\Delta r)$. *The singular values of* $\mathbf{A}$ *are also plotted in Fig.* **??**. *The range is about two orders of magnitude and there is no reason to suspect numerical difficulties. The first and last singular vectors* $\mathbf{u}_{1,51}$, $\mathbf{v}_{1,51}$, *are plotted too. One infers (by plotting additional such vectors), that the large* *singular values correspond to singular vectors showing a great deal of small-scale structure, and the smallest singular values correspond to the least structured (largest spatial scales) in both the solution and in the specific corresponding weighted averages of the equations. This result may be counterintuitive. But note that in this problem, all elements of* $\mathbf{y}$ *vanish except the first two, which are being used to set the boundary conditions. We know from the analytical solution that the true solution is large-scale; most of the information contained in the differential equation (5.115a) or its numerical counterpart, (5.116) is an assertion that all small scales are absent; this information is the most robust and corresponds to the largest singular values. The remaining information, on the exact nature of the largest scales, is contained in only two of the 51 equations, is extremely important, but less robust than that concerning the absence of small scales. (Less "robust" is being used in the sense that small changes in the boundary conditions will lead to relatively large changes in the largescale structures in the solution because of the division by relatively small* λ_i*.).*

EXERCISE 1. *Describe and discuss the above solution when* $k^2 < 0$.

EXERCISE 2. *By the same methods used in this last example, study the behavior of the solution to the modified Bessel equation*

$$r^2\frac{d^2x}{dr^2} + r\frac{dx}{dr} - r^2x = 0, a \leq r \leq b.$$

EXAMPLE 12. *Consider now the classical Neumann problem described in Chapter 1. The problem is to be solved on a* 10×10 *grid with* $\mathbf{A}_3\boldsymbol{\phi} = \mathbf{d}_3$. *The singular values of* $\mathbf{A}_3$ *are plotted in figure 12a; the largest one is* $\lambda_1 = 7.8$, *and the smallest non-zero one is* $\lambda_{99} = 0.08$. *As expected,* $\lambda_{100} = 0$. *The singular vector* $\mathbf{v}_{100}$ *corresponding to the zero singular value is shown in fig. 3–9b, and as expected is a constant;* $\mathbf{u}_{100}$ *shown in fig. 12c is not a constant, it has considerable structure—which provides the solvability condition for the Neumann problem,* $\mathbf{u}^T\mathbf{y} = 0$. *The physical origin of the solvability condition is readily understood: Neumann boundary conditions prescribe boundary flux rates, and the sum of the interior source strengths plus the boundary flux rates must sum to zero, otherwise no steady state is possible. If the boundary conditions are homogeneous, then no flow takes place through the boundary, and the interior sources must sum to zero. In particular, the value of* $\mathbf{u}_{100}$ *on the interior grid points is a constant.* The Neumann problem is thus a forward problem requiring one to deal with both a solution nullspace and a solvability condition.

EXAMPLE 13. *As another example of solution by the SVD, let there be unit positive flux into the box of the previous example on the left boundary, unit positive flux out on the right and no interior sources. The resulting particular SVD solution is shown in fig. 12. No residuals are left because the system was constructed as fully consistent, and there is an arbitrary constant which*

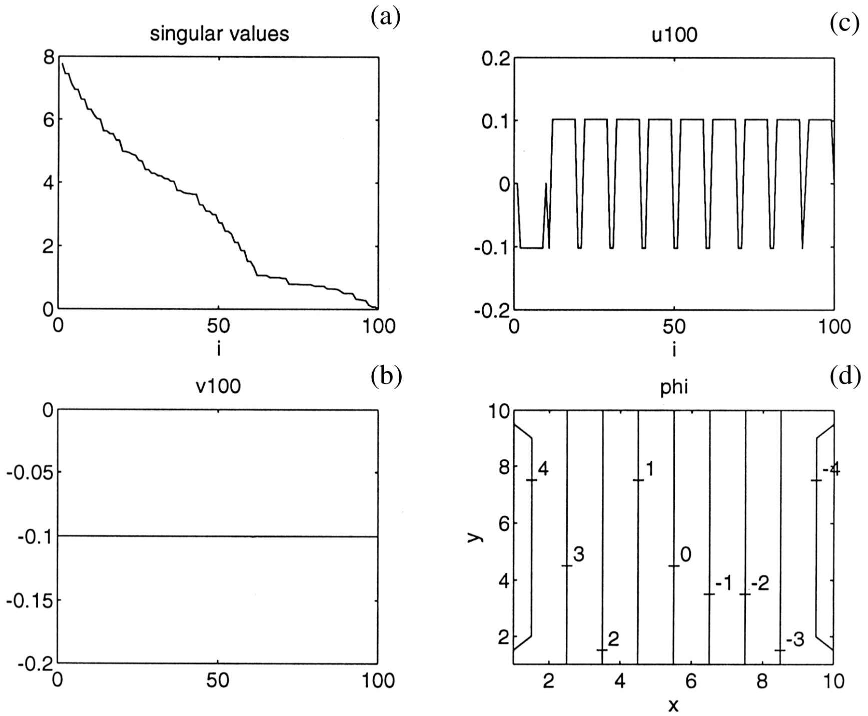


FIGURE 12. (a) Singular values of the coefficient matrix $\mathbf{A}$ of the numerical Neumann problem. All λ_i are non-zero except the last one. (b) $\mathbf{u}_{100,}$ the nullspace vector of $\mathbf{E}^T$ defining the solvability or consistency condition for a solution through $\mathbf{u}_{100}^T\mathbf{y}$ =0 meaning there is no net influx of material through the boundaries or from an interior source. (c) solution nullspace vector $\mathbf{v}_{100}$—a constant which cannot be determined. (d) Particular SVD solution with equal and opposite fluxes across the two vertical walls, zero flux across the horizontal ones, and no interior source. This solution satisfies the solvability condition.

can be added ($\mathbf{v}_{100}$). (The reader may wish to experiment with incompatible specifications for this problem.) This is an example of a forward problem solved using an inverse method (the SVD). Related inverse problems are also easily formulated in simply by interchanging conventional knowns and unknowns. One can also do a number of interesting experiments with the SVD. For example, if the equations imposing the boundary values are dropped, the resulting range vectors, $\mathbf{v}_i$, describe the particular solution of the partial differential equation, and the nullspace vectors describe the homogeneous one. In this way, one can "pick apart" the structure of the solution. A more interesting possibility is to withhold knowledge of the boundary conditions and ask for their determination, given the interior solution.

5.7. Relation of Least-Squares to the SVD. What is the relationship of the SVD solution to the least-squares solutions? To some extent, the answer is already obvious from the orthonormality of the two sets of singular vectors: they *are* the least-squares solution, where it exists. Begin by first asking when the simple least-squares solution will exist? Consider first the formally overdetermined problem, $M > N$. The solution (4.9) exists if and only if the matrix inverse exists. Substituting the SVD for $\mathbf{E}$, one finds

$$(\mathbf{E}^T\mathbf{E})^{-1} = (\mathbf{V}_N\mathbf{\Lambda}_N^T\mathbf{U}_N^T\mathbf{U}_N\mathbf{\Lambda}_N\mathbf{V}_N^T)^{-1} = (\mathbf{V}_N\mathbf{\Lambda}_N^2\mathbf{V}_N^T)^{-1}, \tag{5.117}$$

where the semi-orthogonality of $\mathbf{U}_N$ has been used. Suppose that $K = N$, its maximum possible value; then $\mathbf{\Lambda}_N^2$ is $N \times N$ with *all non-zero diagonal elements* λ_i^2. The inverse in (5.117) may be found by inspection, using $\mathbf{V}_N\mathbf{V}_N^T = \mathbf{I}_N$,

$$(\mathbf{E}^T\mathbf{E})^{-1} = \mathbf{V}_N\mathbf{\Lambda}_N^{-2}\mathbf{V}_N^T\,. \tag{5.118}$$

Then the solution (4.9) becomes

$$\tilde{\mathbf{x}} = (\mathbf{V}_N\mathbf{\Lambda}_N^{-2}\mathbf{V}_N^T)\mathbf{V}_N\mathbf{\Lambda}_N\mathbf{U}_N^T = \mathbf{V}_N\mathbf{\Lambda}_N^{-1}\mathbf{U}_N^T\mathbf{y}\,, \tag{5.119}$$

which is identical to the SVD solution (5.96). If $K < N$, $\mathbf{\Lambda}_N^2$ has at least one zero on the diagonal, no matrix inverse exists and the conventional least-squares solution is not defined. The condition for its existence is thus $K = N$, the so-called “full rank overdetermined” case. The condition $K < N$ is called “rank deficient.” The dependence of the least-squares solution magnitude upon the possible presence of very small, but non-vanishing, singular values is obvious.

That the full-rank overdetermined case is unbiassed, as previously asserted, can now be seen from

$$\langle\tilde{\mathbf{x}} - \mathbf{x}\rangle = \sum_{i=1}^{N}\frac{(\mathbf{u}_i^T\langle\mathbf{y}\rangle)}{\lambda_i}\mathbf{v}_i - \mathbf{x} = \sum_{i=1}^{N}\frac{\mathbf{u}_i^T\mathbf{y}_0}{\lambda_i}\,\mathbf{v}_i - \mathbf{x} = \mathbf{0}\,,$$

if $\langle\mathbf{n}\rangle = \mathbf{0}$, assuming that the correct $\mathbf{E}$ (model) is being used.

Now consider another least-squares problem, the conventional purely underdetermined least-squares problem, whose solution is (4.75). When does that exist? Substituting the SVD,

$$\begin{aligned}\tilde{\mathbf{x}} &= \mathbf{V}_M\mathbf{\Lambda}_M\mathbf{U}_M^T(\mathbf{U}_M\mathbf{\Lambda}_M\mathbf{V}_M^T\mathbf{V}_M\mathbf{\Lambda}_M^T\mathbf{U}_M^T)^{-1}\mathbf{y}\\ &= \mathbf{V}_M\mathbf{\Lambda}_M\mathbf{U}_M^T(\mathbf{U}_M\mathbf{\Lambda}_M^2\mathbf{U}_M^T)^{-1}\mathbf{y}\,.\end{aligned} \tag{5.120}$$

Again, the matrix inverse exists if and only if $\mathbf{\Lambda}_M^2$ has all non-zero diagonal elements, which occurs only when $K = M$. Under that specific condition, the inverse is obtained by inspection and,

$$\tilde{\mathbf{x}} = \mathbf{V}_M\mathbf{\Lambda}_M\mathbf{U}_M^T(\mathbf{U}_M\mathbf{\Lambda}_M^{-2}\mathbf{U}_M^T)\mathbf{y} = \mathbf{V}_M\mathbf{\Lambda}_M^{-1}\mathbf{U}_M^T\mathbf{y} \tag{5.121}$$

$$\tilde{\mathbf{n}} = 0\,, \tag{5.122}$$

which is once again the particular-SVD solution (5.96)—with the nullspace coefficients set to zero. This situation is usually referred to as the “full-rank underdetermined case.” Again, the possible influence of small singular values is apparent and an arbitrary sum of nullspace vectors can be added to (5.121). The bias of (5.120) is given by the nullspace elements, and its formal uncertainty is only from the nullspace contribution, because with $\tilde{\mathbf{n}} = \mathbf{0}$, the formal noise variance vanishes, and the particular-SVD solution covariance $\mathbf{C}_{xx}$ would be zero.

The particular-SVD solution thus coincides with the two simplest forms of least-squares solution, and generalizes both of them to the case where the matrix inverses do not exist. *All of the structure imposed by the SVD, in particular the restriction on the residuals in* (5.74)*, is present in the least-squares solution.* If the system is not of full rank, then the simple least-squares solutions do not exist. *The SVD generalizes these results* by determining what it can: the elements of the solution lying in the range of $\mathbf{E}$, and an explicit structure for the resulting nullspace vectors.

The SVD provides a lot of flexibility. For example, it permits one to modify the simplest underdetermined solution (4.75) to remove its greatest shortcoming, the necessity that $\tilde{\mathbf{n}} = \mathbf{0}$. One simply truncates the solution (5.80) at $K = K' < M$, thus assigning all vectors $\mathbf{v}_i$, $K'+1 \leq i \leq K$, to an "effective nullspace" (or substitutes K' for K everywhere). The resulting residual is then

$$\tilde{\mathbf{n}} = \sum_{i=K'+1}^{M} (\mathbf{u}_i^T \mathbf{y})\mathbf{u}_i \,, \tag{5.123}$$

with an uncertainty for $\tilde{\mathbf{x}}$ given by 5.103, but with the upper limit being K' rather than K. Such truncation has the effect of reducing the solution covariance contribution to the uncertainty, but increasing the contribution owing to the nullspace (and increasing the potential bias). In the presence of singular values small compared to σ_n, the resulting overall reduction in uncertainty may be very great—at the expense of a possibly very small bias. This consideration is extremely important—it says that despite the mathematical condition $\lambda_i \neq 0$, some structures in the solution cannot be estimated with sufficient reliability to be useful. The "effective rank" is then not the same as the mathematical rank.

The solution now consists of three parts,

$$\tilde{\mathbf{x}} = \sum_{i=1}^{K'} \frac{\mathbf{u}_i^T \mathbf{y}}{\lambda_i} \mathbf{v}_i + \sum_{i=K'+1}^{K} \alpha_i \, \mathbf{v}_i + \sum_{i=K+1}^{N} \alpha_i \mathbf{v}_i \,, \tag{5.124}$$

where the middle sum contains the terms appearing with singular values too small to be employed—for the given noise—and the third sum is the strict nullspace. Usually, one lumps the two nullspace sums together. The first sum, by itself, represents the particular-SVD solution in the presence of noise.

It was already noticed that the simplest form of least-squares does not provide a method to control the ratios of the solution and noise norms. Evidently, truncation of the SVD offers a simple way to do so—by modifying K'. It follows that the solution norm necessarily is reduced, and that the residuals must grow, along with the size of the solution nullspace. The issue of how to choose K', that is, "rank determination," in practice involves studying the tradeoff between the desire for the highest possible resolution (large K'), and acceptable variance (small K'). .

5.8. Row and Column Scaling. The effects on the least-squares solutions of the row and column scaling can now be understood. We discuss them in the context of noise covariances, but as always in least-squares, the weight matrices need no statistical interpretation, and can be chosen by the investigator to suit his convenience or taste.

Suppose we have two equations

$$\left\{\begin{matrix} 1 & 1 & 1 \\ 1 & 1.01 & 1 \end{matrix}\right\} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix},$$

and there is no information about the noise covariance and so no row scaling is reasonable: $\mathbf{W} = \mathbf{I}$. The SVD of $\mathbf{E}$ is

$$\mathbf{U} = \left\{\begin{matrix} 0.7059 & -0.7083 \\ 0.7083 & 0.7059 \end{matrix}\right\}, \qquad \mathbf{V} = \left\{\begin{matrix} 0.5764 & -0.4096 & 0.7071 \\ 0.5793 & 0.8151 & 0.0000 \\ 0.5764 & -0.4096 & -0.7071 \end{matrix}\right\},$$

$$\lambda_1 = 2.4536, \qquad \lambda_2 = .0058\,.$$

The SVD solutions, choosing ranks $K' = 1, 2$ in succession, are very nearly (the numbers having been rounded),

$$\begin{aligned} \tilde{\mathbf{x}} &\sim \frac{(y_1+y_2)}{2.45}\begin{bmatrix} .58 \\ .58 \\ .58 \end{bmatrix}, \\ &\sim \frac{(y_1+y_2)}{2.45}\begin{bmatrix} .58 \\ .58 \\ .58 \end{bmatrix} + \frac{(y_1-y_2)}{.0058}\begin{bmatrix} -.41 \\ .82 \\ .41 \end{bmatrix}, \end{aligned} \tag{5.125}$$

respectively, so that the first term simply averages the two measurements, y_i, and the difference between them contributes—with great uncertainty—in the second term of the rank 2 solution owing to the very small singular value. The uncertainty is

$$(\mathbf{E}\mathbf{E}^T)^{-1} = \left\{\begin{matrix} 1.51\times10^4 & -1.50\times10^4 \\ -1.50\times10^4 & 1.51\times10^4 \end{matrix}\right\}. \tag{5.126}$$

Now suppose that the covariance matrix of the noise is known to be

$$\mathbf{R}_{nn} = \left\{\begin{matrix} 1 & 0.999999 \\ 0.999999 & 1 \end{matrix}\right\}$$

(an extreme case, chosen for illustrative purposes). Then, put $\mathbf{W} = \mathbf{R}_{nn}$,

$$\mathbf{W}^{1/2} = \begin{Bmatrix} 1.0000 & 1.0000 \\ 0 & 0.0014 \end{Bmatrix}, \qquad \mathbf{W}^{-T/2} = \begin{Bmatrix} 1.0000 & 0 \\ -707.1063 & 707.1070 \end{Bmatrix}.$$

The new system to be solved is

$$\begin{Bmatrix} 1.0000 & 1.0000 & 1.0000 \\ 0.0007 & 7.0718 & 0.0007 \end{Bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} y_1 \\ 707.1(-y_1 + y_2) \end{bmatrix}.$$

The SVD is

$$\mathbf{U} = \begin{Bmatrix} 0.1456 & 0.9893 \\ 0.9893 & -0.1456 \end{Bmatrix}, \qquad \mathbf{V} = \begin{Bmatrix} 0.0205 & 0.7068 & 0.7071 \\ 0.9996 & -0.0290 & 0.0000 \\ 0.0205 & 0.7068 & -0.7071 \end{Bmatrix}$$

$$\lambda_1 = 7.1450\,, \qquad \lambda_2 = 1.3996\,.$$

The second singular value is now much larger relative to the first one, and the two solutions are,

$$\begin{aligned} \tilde{\mathbf{x}} &\sim \frac{(y_2 - y_1)}{7.1} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \\ &\sim \frac{(y_2 - y_1)}{7.1} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + \frac{y_1}{1.4} \begin{bmatrix} .71 \\ 0 \\ .71 \end{bmatrix}, \end{aligned} \tag{5.127}$$

and the rank 1 solution is given by the difference of the observations, in stark contrast to the unscaled solution. The result is quite sensible—the noise in the two equations is so nearly perfectly correlated, that it can be removed by subtraction; the difference $y_2 - y_1$ is a nearly noise-free piece of information and accurately defines the appropriate structure in $\tilde{\mathbf{x}}$. In effect, the information provided in the row scaling with $\mathbf{R}$ permits the SVD to nearly eliminate the noise at rank 1 by an effective subtraction, whereas without that information, the noise is reduced in the solution (5.125) at rank 1 only by direct averaging.

At full rank, that is, $K = 2$, it can be confirmed that the solutions (5.125) and (5.127) are identical, as they must be. But the error covariances are quite different:

$$(\mathbf{E}'\mathbf{E}'^T)^{-1} = \begin{Bmatrix} 0.5001 & -0.707 \\ -0.707 & 0.5001 \end{Bmatrix}.$$

because the imposed covariance permits a large degree of noise suppression.

It was previously asserted that in a full-rank formally underdetermined system, row scaling is irrelevant to $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$, as may be seen as follows,

$$\begin{aligned}\tilde{\mathbf{x}} &= \mathbf{E}'^T(\mathbf{E}'\mathbf{E}'^T)^{-1}\mathbf{y}' \\ &= \mathbf{E}^T\mathbf{W}^{-1/2}(\mathbf{W}^{-T/2}\mathbf{E}\mathbf{E}^T\mathbf{W}^{-1/2})^{-1}\mathbf{W}^{-T/2}\mathbf{y} \\ &= \mathbf{E}\mathbf{W}^{-1/2}\mathbf{W}^{1/2}(\mathbf{E}\mathbf{E}^T)^{-1}\mathbf{W}^{T/2}\mathbf{W}^{-T/2}\mathbf{y} \\ &= \mathbf{E}^T(\mathbf{E}\mathbf{E}^T)^{-1}\mathbf{y}\,,\end{aligned} \tag{5.129}$$

where we used the result $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ and both inverses must exist, which is possible only in the full rank situation.

There is a subtlety in row-weighting. Suppose we have two equations of form,

$$\begin{aligned}10x_1 + 5x_2 + x_3 &= 1\,, \\ 100x_1 + 50x_2 + 10x_3 &= 2\,,\end{aligned} \tag{5.130}$$

after row scaling to make the expected noise variance in each the same. A rank 1 solution to these equations by SVD is $\tilde{\mathbf{x}} = [.0165, .0083, .0017]^T$, which produces residuals $\tilde{\mathbf{y}} - \mathbf{y} = [-0.79, 0.079]^T$—much smaller in the second equation than in the first one.

Consider that the second equation is 10 times the first one—in effect we are saying that a measurement of 10 times the values of $10x_1 + 5x_2 + x_3$ has the same noise in it as a measurement of one times this same linear combination. The second equation clearly represents a much more accurate determination of this linear combination and the equation should be given much more weight in determining the unknowns—and both the SVD and ordinary least-squares does precisely that. To the extent that one finds this result undesirable (one should be careful about why it is so found), there is an easy remedy—divide the equations by their row norms $\left(\sum_j E_{ij}^2\right)^{1/2}$. But there will be a contradiction with the assertion that the noise in all equations was the same to begin with. Such row-scaling is best regarded as non-statistical in nature.

An example of this situation is readily apparent in the box balances discussed in Chapter 1. Equations such as Eq. 3.3 of Chapter 1 could have row norms much larger than those Eq. 3.2, Chapter 1 for the corresponding mass balance, simply because the tracer is measured by convention in its own units. If the tracer is e.g., oceanic salt, values are, by convention, measured on the Practical Salinity Scale, and are near 35 (but are dimensionless). Because there is nothing fundamental about the choice of units, it seems unreasonable to infer that the requirement of tracer balance has an expected error 35 times smaller than for mass. One usually proceeds in the obvious way by dividing the tracer equations by their row norms as the first step. (This approach need have no underlying statistical validity, but is often done simply on the assumption that salt equations are unlikely to be 35 times more accurate than the mass ones.) The second step is to ask whether anything further can be said about the relative errors of mass and salt balance, which would introduce a second, purely statistical row weight.

Column Scaling

In the least-squares problem, we formally introduced a "column scaling" matrix $\mathbf{S}$. Column scaling operates on the SVD solution exactly as it does in the least-squares solution, to which it reduces in the two special cases already described. That is, we should apply the SVD to sets of equations only where any knowledge of the solution element size has been removed first. If the SVD has been computed for such a column-scaled (and row-scaled) system, the solution is for the scaled unknown $\mathbf{x}'$, and the physical solution is

$$\tilde{\mathbf{x}} = \mathbf{S}^{-T/2}\tilde{\mathbf{x}}' \,. \tag{5.131}$$

But there are occasions, with underdetermined systems, where a non-statistical scaling may also be called for, the analogue to the situation considered above where a row-scaling was introduced on the basis of possible non-statistical considerations.

Example

Suppose we have one equation in two unknowns,

$$10x_1 + 1x_2 = 3\,. \tag{5.132}$$

The particular-SVD solution produces $\tilde{\mathbf{x}} = [0.2970, 0.0297]^T$ in which the magnitude of x_1 is much larger than that of x_2 and the result is readily understood. As we have seen, the SVD automatically finds the exact solution, subject to making the solution norm as small as possible. Because the coefficient of x_1 in (5.132) is 10 times that of x_2, it is obviously more efficient in minimizing the norm to give x_1 a larger value than x_2—because it contributes more efficiently in producing y.

Although we have demonstrated this dependence for a trivial example, similar behavior occurs for underdetermined systems in general. In many cases, this distribution of the elements of the solution vector $\mathbf{x}$ is desirable, the numerical value 10 appearing for good physical reasons. In other problems—the numerical values appearing in the coefficient matrix $\mathbf{E}$ are an "accident." In the box-balance example of Chapter 1, the distance defining the interfaces of the boxes are a consequence of the distance a ship steamed before stopping to make measurements. Unless one believed that velocities should be larger where the ship steamed further, or the water was deeper, then the solutions may behave unphysically. Indeed, in some situations the velocities are expected to be inverse to the water depth and such a prior statistical hypothesis is best imposed after one has removed the structural accidents from the system. (The tendency for the solutions to be proportional to the column norms is not rigid. In particular, the equations themselves may actually preclude the proportionality.)

Take a positive definite, diagonal matrix $\mathbf{S}$, and rewrite (4.2) as

$$\mathbf{E}\mathbf{S}^{-T/2}\mathbf{S}^{T/2}\mathbf{x}\ +\mathbf{n}=\mathbf{y}.$$

Then,

$$\mathbf{E}'\mathbf{x}'+\mathbf{n}=\mathbf{y}',\ \mathbf{E}'=\mathbf{E}\mathbf{S}^{T/2},\ \mathbf{x}'=\mathbf{S}^{-T/2}\mathbf{x}\,.$$

Solving

$$\tilde{\mathbf{x}}'=\mathbf{E}'^T(\mathbf{E}'\mathbf{E}'^T)^{-1}\mathbf{y}\,,\ \tilde{\mathbf{x}}=\mathbf{S}^{T/2}\tilde{\mathbf{x}}'\,. \tag{5.133}$$

How should $\mathbf{S}$ be chosen? Apply the recipe (5.133) for the simple one equation example of (5.132), with

$$\mathbf{S}=\left\{\begin{array}{cc} 1/L_{11} & 0 \\ 0 & 1/L_{22}\end{array}\right\}$$

:

$$\begin{aligned}
\mathbf{E}' &= \left\{\begin{array}{cc} 10/L_{11}^{1/2} & 1/SL_{22}^{1/2}\end{array}\right\},\ \mathbf{E}'\mathbf{E}'^T=\frac{100}{L_{11}}+\frac{1}{L_{22}} && (5.134)\\
\left(\mathbf{E}'\mathbf{E}'^T\right)^{-1} &= \frac{L_{11}L_{22}}{100L_{22}+L_{11}} && (5.135)\\
\tilde{\mathbf{x}}' &= \left\{\begin{array}{c} 10/L_{11}^{1/2} \\ 1/L_{22}^{1/2}\end{array}\right\}\frac{L_{11}}{100L_{22}+L_{11}}3, && (5.136)\\
\tilde{\mathbf{x}} &= \mathbf{S}^{-1/2}\tilde{\mathbf{x}}'=\left\{\begin{array}{c} 10/L_{11} \\ 1/L_{22}\end{array}\right\}\frac{L_{11}}{100L_{22}+L_{11}}3. && (5.137)
\end{aligned}$$

The relative magnitudes of the elements of $\tilde{\mathbf{x}}$ are proportional to $10/L_{11}$, $1/L_{22}$. To make the numerical values of the elements of $\tilde{\mathbf{x}}$ the same, we should clearly choose $L_{11}=10$, $L_{22}=1$, that is, we should divide the elements of the first column of $\mathbf{E}$ by $\sqrt{10}$ and the second column by $\sqrt{1}$. The apparent rule (which is correct and general) is to divide each column of $\mathbf{E}$ by the square root of its length. The square root of the length may be surprising, but arises because of the second multiplication by the elements of $\mathbf{S}^{T/2}$ in (5.133). This form of column scaling should be regarded as "non-statistical," in that it is based upon inferences about the numerical magnitudes of the columns of $\mathbf{E}$ and does not employ information about the statistics of the solution. Indeed, its purpose is to prevent the imposition of structure on the solution for which no statistical basis has been anticipated.

If the system is full-rank overdetermined, the column weights drop out, just as we claimed for least-squares above. To see this, consider that in the full-rank case,

$$\begin{aligned}
\tilde{\mathbf{x}}' &= (\mathbf{E}'^T\mathbf{E}')^{-1}\mathbf{E}'^T\mathbf{y}\\
\tilde{\mathbf{x}} &= \mathbf{S}^{T/2}(\mathbf{S}^{1/2}\mathbf{E}^T\mathbf{E}\mathbf{S}^{T/2})^{-1}\mathbf{S}^{1/2}\mathbf{E}^T\mathbf{y} && (5.138)\\
&= \mathbf{S}^{T/2}\mathbf{S}^{-T/2}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{S}^{-1/2}\mathbf{S}^{1/2}\mathbf{E}^T\mathbf{y}=(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\mathbf{y}\,.
\end{aligned}$$

Usually row-scaling is done prior to column scaling so that the row norms have a simple physical interpretation.

5.9. Solution and Observation Resolution. Data Ranking. Typically, either or both of the set of vectors $\mathbf{v}_i$, $\mathbf{u}_i$ used to present $\mathbf{x}$, $\mathbf{y}$ will be deficient in the sense of the expansions in (5.2). It follows immediately from eqs. (5.3) that the particular-SVD solution is,

$$\tilde{\mathbf{x}} = \mathbf{V}_K \mathbf{V}_K^T \mathbf{x} = \mathbf{T}_v \mathbf{x}, \tag{5.139}$$

and the data vector with which both it and the general solution are consistent is

$$\tilde{\mathbf{y}} = \mathbf{U}_K \mathbf{U}_K^T \mathbf{y} \ = \mathbf{T}_u \mathbf{y}. \tag{5.140}$$

It is convenient therefore, to define the solution and observation resolution matrices,

$$\mathbf{T}_v = \mathbf{V}_K \mathbf{V}_K^T, \qquad \mathbf{T}_u = \mathbf{U}_K \mathbf{U}_K^T\,. \tag{5.141}$$

The interpretation of the solution resolution matrix is identical to that in the square-symmetric case (P.69).

Interpretation of the data resolution matrix is slightly subtle. Suppose an element of $\mathbf{y}$ was fully resolved, that is, some row, j_0, of $\mathbf{U}_K \mathbf{U}_K^T$ were all zeros except for diagonal element j_0, which is one. Then a change of unity in y_{j_0} would produce a change in $\tilde{\mathbf{x}}$ which would leave unchanged all other elements of $\tilde{\mathbf{y}}$. If element j_0 is *not* fully resolved, then a change of unity in observation y_{j_0} produces a solution which leads to changes in other elements of $\tilde{\mathbf{y}}$. Stated slightly differently, if y_i is not fully resolved, the system lacks adequate information to distinguish equation i from a linear dependence on one or more other equations.

One can use these ideas to construct quantitative statements of which observations are the most important ("data ranking"). From (5.5), trace$(\mathbf{T}_u) = K$ and the relative contribution to the solution of any particular constraint is given by the corresponding diagonal element of $\mathbf{T}_u$.

Consider the example (5.130) without row weighting. At rank 1,

$$\mathbf{T}_u = \begin{Bmatrix} 0.0099 & 0.099 \\ 0.099 & 0.9901 \end{Bmatrix},$$

showing that the second equation has played a much more important role in the solution than the first one—despite the fact that we asserted the expected noise in both to be the same. The reason is that described above, the second equation in effect asserts that the measurement is 10 times more accurate than in the first equation—and the data resolution matrix informs us of that explicitly. The elements of $\mathbf{T}_u$ can be used to rank the data in order of importance to the final solution. All of the statements made above to $\mathbf{T}_u$, $\mathbf{T}_v$.

If row and column scaling have been applied to the equations prior to application of the SVD, the covariance, uncertainty, and resolution expressions apply in those new, scaled spaces.

The resolution in the original spaces is,

$$\mathbf{T}_v = \mathbf{S}^{T/2}\mathbf{T}_{v'}\mathbf{S}^{-T/2}\,, \tag{5.142}$$

$$\mathbf{T}_u = \mathbf{W}^{T/2}\mathbf{T}_{u'}\mathbf{W}^{-T/2}\,, \tag{5.143}$$

so that

$$\tilde{\mathbf{x}} = \mathbf{T}_v\mathbf{x}, \qquad \tilde{\mathbf{y}} = \mathbf{T}_u\mathbf{y} \tag{5.144}$$

where $\mathbf{T}_{v'}$, $\mathbf{T}_{u'}$ are the expressions (126), (127) in the scaled space. The uncertainty in the new space is $\mathbf{P} = \mathbf{S}^{1/2}\mathbf{P}'\mathbf{S}^{T/2}$ where $\mathbf{P}'$ is the uncertainty in the scaled space.

We have seen an interpretation of three matrices obtained from the SVD: $\mathbf{V}_K\mathbf{V}_K^T$, $\mathbf{U}_K\mathbf{U}_K^T$, $\mathbf{V}_K\Lambda_K^{-2}\mathbf{V}_K^T$. The reader may well wonder, on the basis of the symmetries between solution and data spaces, whether there is an interpretation of the remaining matrix $\mathbf{U}_K\Lambda_K^{-2}\mathbf{U}_K^T$?

To understand its use, recall, the normal equations (4.71, 4.72) that emerged from the constrained objective function (4.59). They become, using the SVD for $\mathbf{E}$,

$$\mathbf{V}\boldsymbol{\Lambda}\mathbf{U}^T\boldsymbol{\mu} = \mathbf{x}\,, \tag{5.145}$$

$$\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^T\mathbf{x} = \mathbf{q}\,. \tag{5.146}$$

These equations show that $\mathbf{U}\boldsymbol{\Lambda}^2\mathbf{U}^T\boldsymbol{\mu} = \mathbf{q}$. The particular SVD solution is,

$$\tilde{\boldsymbol{\mu}} = \mathbf{U}_K\boldsymbol{\Lambda}_K^{-2}\mathbf{U}_K^T\mathbf{q}\,, \tag{5.147}$$

involving the "missing" fourth matrix. Thus,

$$\frac{\partial J}{\partial \mathbf{q}} = 2\mathbf{U}_K\boldsymbol{\Lambda}_K^{-2}\mathbf{U}_K^T\mathbf{q}\,,$$

and taking the second derivative,

$$\frac{\partial^2 J}{\partial \mathbf{y}^2} = 2\mathbf{U}_K\boldsymbol{\Lambda}_K^{-2}\mathbf{U}_K^T \tag{5.148}$$

is the Hessian of J with respect to the data. If any of the λ_i become very small, the objective function will be extremely sensitive to small perturbations in $\mathbf{y}$—producing an effective nullspace of the problem. Eq. (5.148) supports the suggestion that perfect constraints can lead to difficulties.

5.10. Relation to Tapered and Weighted Least-Squares. In using least-squares, a shift was made from the simple objective functions (4.4) and (4.59) to the more complicated ones in (4.27) or (4.37). The change was made to permit a degree of control of the relative norms of $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$, and through the use of $\mathbf{W}$, $\mathbf{S}$ of the individual elements and the resulting uncertainties, and covariances. Application of the weight matrices $\mathbf{W}$, $\mathbf{S}$ through their Cholesky decompositions to the equations prior to the use of the SVD is equally valid—thus providing the same amount of influence over the solution elements. The SVD provides its control over the solution norms, uncertainties and covariances through choice of the effective rank K'. This approach is different from the use of the extended objective functions (4.27), but the SVD is actually useful in understanding the effect of such functions.

Assume any necessary $\mathbf{W}$, $\mathbf{S}$ have been applied. Then, the full SVD, including zero singular values and corresponding singular vectors, is substituted into 4.28,

$$\tilde{\mathbf{x}} = (\alpha^2\mathbf{I}_N + \mathbf{V}\mathbf{\Lambda}^T\mathbf{\Lambda}\mathbf{V}^T)^{-1}\mathbf{V}\mathbf{\Lambda}^T\mathbf{U}^T\mathbf{y}\,,$$

and using $\mathbf{V}\mathbf{V}^T = \mathbf{I}$, we have

$$\begin{aligned}\tilde{\mathbf{x}} &= \mathbf{V}(\mathbf{\Lambda}^T\mathbf{\Lambda} + \alpha^2\mathbf{I})^{-1}\mathbf{\Lambda}^T\mathbf{U}^T\mathbf{y} \\ &= \mathbf{V}\,\mathrm{diag}\left(\lambda_i^2 + \alpha^2\right)^{-1}\mathbf{\Lambda}^T\mathbf{U}^T\mathbf{y}.\end{aligned} \tag{5.149}$$

or,

$$\tilde{\mathbf{x}} = \sum_{i=1}^{N} \frac{\lambda_i(\mathbf{u}_i^T\mathbf{y})}{\lambda_i^2 + \alpha^2}\mathbf{v}_i\,. \tag{5.150}$$

It is now apparent what the effect of "tapering" has done in least-squares. The word refers to the tapering down of the coefficients by the presence of α^2 of the $\mathbf{v}_i$ from the values they would have in the "pure" SVD . In particular, the guarantee that matrices like $(\mathbf{E}^T\mathbf{E} + \alpha^2\mathbf{I})$ always have an inverse despite vanishing singular values, is seen to follow because the presence of $\alpha^2 > 0$, assures the inverse of the sum always exists, irrespective of the rank of $\mathbf{E}$. The simple addition of a positive constant to the diagonal of a singular matrix is a well-known ad hoc method for giving it an approximate inverse. Such methods are a form of what is usually known as "regularization," and are just procedures for suppressing nullspaces.

The residuals of the tapered least-squares solution can be written in various forms. Eqs. (4.29) are,

$$\begin{aligned}\tilde{\mathbf{n}} &= \alpha^2\mathbf{U}(\alpha^2\mathbf{I} + \mathbf{\Lambda}\mathbf{\Lambda}^T)^{-1}\mathbf{U}^T\mathbf{y} \\ &= \sum_{i=1}^{M} \frac{(\mathbf{u}_i^T\mathbf{y})\alpha^2}{\lambda_i^2 + \alpha^2}\mathbf{u}_i\,,\end{aligned} \tag{5.151}$$

that is, the projection of the noise onto the range vectors $\mathbf{u}_i$ no longer vanishes. Some of the structure of the range of $\mathbf{E}^T$ is being attributed to noise and it is no longer true that the residuals are subject to the rigid requirement (5.74) of having zero contribution from the range vectors. An increased noise norm is also deemed acceptable, as the price of keeping the solution norm small, by assuring that none of the coefficients in the sum (5.150) becomes overly large—values we can control by varying α^2. The covariance of this solution about its mean (eq. 4.30) is readily rewritten as

$$\begin{aligned}\mathbf{C}_{xx} &= \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\lambda_i\lambda_j\mathbf{u}_i\mathbf{R}_{nn}\mathbf{u}_i^T}{(\lambda_i^2+\alpha^2)(\lambda_j^2+\alpha^2)}\mathbf{v}_i\mathbf{v}_j^T \\ &= \sigma_n^2\sum_{i=1}^{N}\frac{\lambda_i^2}{(\lambda_i^2+\alpha^2)^2}\,\mathbf{v}_i\mathbf{v}_j^T \\ &= \sigma_n^2\mathbf{V}(\mathbf{\Lambda}^T\mathbf{\Lambda}+\alpha^2\mathbf{I}_N)^{-1}\mathbf{\Lambda}^T\mathbf{\Lambda}(\mathbf{\Lambda}^T\mathbf{\Lambda}+\alpha^2\mathbf{I}_N)^{-1}\mathbf{V}^T\end{aligned} \tag{5.152}$$

where the second and third lines are again the special case of white noise. The role of α^2 in controlling the solution variance, as well as the solution size, should be plain. The tapered least-squares solution is biassed —but the presence of the bias can greatly reduce the solution variance. Study of the solution as a function of α^2 is known as "ridge regression"[34]. Elaborate techniques have been developed for determining the "right" value of α^2.[35]

The uncertainty, $\mathbf{P}$, is readily found as,

$$\begin{aligned}\mathbf{P}=&\alpha^2\sum_{i=1}^{N}\frac{\mathbf{v}_i\mathbf{v}_i^T}{\left(\lambda_i^2+\alpha^2\right)^2}+\sigma_n^2\sum_{i=1}^{N}\frac{\lambda_i^2\mathbf{v}_i\mathbf{v}_i^T}{\left(\lambda_i^2+\alpha^2\right)^2} \\ = \quad &\alpha^2\mathbf{V}\left(\mathbf{\Lambda}^T\mathbf{\Lambda}+\alpha^2\mathbf{I}\right)^{-2}\mathbf{V}^T+\boldsymbol{\sigma}_n^2\mathbf{V}\left(\mathbf{\Lambda}^T\mathbf{\Lambda}+\alpha^2\mathbf{I}\right)^{-1}\mathbf{\Lambda}^T\mathbf{\Lambda}\left(\mathbf{\Lambda}^T\mathbf{\Lambda}+\alpha^2\mathbf{I}\right)^{-1}\mathbf{V}^T\end{aligned} \tag{5.153}$$

showing the variance reduction possible for finite α^2 (reduction of the second term), and the bias error incurred in compensation in the first term.

The truncated SVD and the tapered SVD-tapered least-squares solutions produce the same qualitative effect—it is possible to increase the noise norm while decreasing the solution norm. Although the solutions differ somewhat, they both achieve a purpose stated above—to extend ordinary least-squares in such a way that one can control the relative noise and solution norms. The quantitative difference between them is readily stated—the truncated form makes a clear separation between range and nullspace in both solution and residual spaces: The basic SVD solution contains only range vectors and no nullspace vectors. The residual contains only nullspace vectors and no range vectors. The tapered form permits a merger of the two different sets of vectors: Then both solution and residuals contain some contribution from both formal range and effective nullspaces.

We have already seen several times that preventing $\tilde{\mathbf{n}}$ from having any contribution from the range of $\mathbf{E}^T$ introduces covariances into the residuals, with a consequent inability to produce values which are strictly white noise in character (although it is only a real issue as the number of degrees of freedom, $M-K$, goes toward zero). In the tapered form of least-squares, or the equivalent tapered SVD, contributions from the range vectors $\mathbf{u}_i$, $i \le K$, is permitted, and a

[34] Hoerl & Kennard (1970a,b).

[35] Lawson & Hanson (1974), or Hansen (1992). Hansen's (1992) discussion is particularly interesting because he exploits the "generalized SVD," which is used to simultaneously diagonalize two matrices.

potentially more realistic residual estimate is obtained. (There is usually no good reason why $\tilde{\mathbf{n}}$ should be expected to be orthogonal to the range vectors.)

5.11. Resolution and Variance of Tapered Solutions to Simultaneous Equations. The tapered least-squares solutions have an implicit nullspace, arising from the terms corresponding to zero singular values, or values small compared to α^2. But that solution form does a good job of hiding the existence of what should still be regarded as an effective nullspace.

To obtain a measure of solution resolution when the $\mathbf{v}_i$ vectors have not been computed, consider a situation in which the true solution were $\mathbf{x}_{j_0} \equiv \delta_{j,j_0}$, that is, unity in the j_0 element and zero elsewhere. Then, in the absence of noise (the resolution analysis applies to the noise-free situation), the correct value of $\mathbf{y}$ would be

$$\mathbf{E}\mathbf{x}_{j_0} = \mathbf{y}_{j_0}\,, \tag{5.154}$$

defining $\mathbf{y}_{j_0}$. Suppose we actually knew (had measured) $\mathbf{y}_{j_0}$, what solution $\mathbf{x}_{j_0}$ would be obtained?

Suppressing all covariances matrices, tapered least-squares (Eqs. 4.32) produces

$$\tilde{\mathbf{x}}_{j_0} = \mathbf{E}^T(\mathbf{E}\mathbf{E}^T + \alpha^2\mathbf{I})^{-1}\mathbf{y}_{j_0} = \mathbf{E}^T(\mathbf{E}\mathbf{E}^T + \alpha^2\mathbf{I})^{-1}\mathbf{E}\mathbf{x}_{j_0}\,, \tag{5.155}$$

which is row (or column) j_0 of

$$\mathbf{T}_v = \mathbf{E}^T(\mathbf{E}\mathbf{E}^T + \alpha^2\mathbf{I})^{-1}\mathbf{E}\,. \tag{5.156}$$

Thus we can interpret any row or column of $\mathbf{T}_v$ as the solution resolution for a Kronecker delta, correct solution, in that element. It is an easy matter, using the SVD of $\mathbf{E}$ and letting $\alpha^2 \to 0$ to show that (5.156) reduces to $\mathbf{V}\mathbf{V}^T$. These expressions apply in the row- and column-scaled space and are suitably modified to take account of any $\mathbf{W}, \mathbf{S}$ which may have been applied, as in Eqs. (5.142), (5.143). An obvious variant of (5.156) follows from the alternative least-squares solution (4.84), with $\mathbf{W} = \alpha^2\mathbf{I}$, $\mathbf{S} = \mathbf{I}$,

$$\mathbf{T}_v = \left(\mathbf{E}^T\mathbf{E} + \alpha^2\mathbf{I}\right)^{-1}\mathbf{E}^T\mathbf{E} \tag{5.157}$$

Solution resolution matrices are obtained similarly. Let $y_j = \delta_{jj_1}$.The Eq. (4.47) produces

$$\tilde{\mathbf{x}}_{j_1} = \mathbf{E}^T\left(\mathbf{E}\mathbf{E}^T + \alpha^2\mathbf{I}\right)^{-1}\mathbf{y}_{j_1}, \tag{5.158}$$

which if substituted into the original equations is,

$$\mathbf{E}\tilde{\mathbf{x}}_{j_1} = \mathbf{E}\mathbf{E}^T\left(\mathbf{E}\mathbf{E}^T + \alpha^2\mathbf{I}\right)^{-1}\mathbf{y}_{j_1}. \tag{5.159}$$

Thus,

$$\mathbf{T}_u = \mathbf{E}\mathbf{E}^T\left(\mathbf{E}\mathbf{E}^T + \alpha^2\mathbf{I}\right)^{-1} \tag{5.160}$$

The alternate form is,

$$\mathbf{T}_u = \mathbf{E}\left(\mathbf{E}^T\mathbf{E} + \alpha^2\mathbf{I}\right)^{-1}\mathbf{E}^T. \tag{5.161}$$

All of the resolution matrices reduce properly to either $\mathbf{U}\mathbf{U}^T, \mathbf{V}\mathbf{V}^T$ as $\alpha^2 \to 0$ when the SVD for $\mathbf{E}$ is substituted.

6. Combined Least-Squares and Adjoints (Lagrange Multipliers)

6.1. Exact Constraints. Consider now a modest generalization of the constrained problem eq. (4.2) in which the unknowns $\mathbf{x}$ are also meant to satisfy some constraints exactly, or nearly so, for example

$$\mathbf{A}\mathbf{x} = \mathbf{d}\,. \tag{6.1}$$

In some contexts, (6.1) is referred to as the "model," a term also employed, confusingly, for the physics defining $\mathbf{E}$ along with the statistics assumed to describe $\mathbf{x}, \mathbf{n}$. In the end, there is no unique meaning to the term, and only the context is a guide. We will temporarily refer to Eq. (6.1) as "perfect constraints," as opposed to those involving $\mathbf{E}$, which generally always have a non-zero noise element.

An example of a model in these terms occurs in acoustic tomography (Chapter 1), where measurements exist of both density and velocity fields, and they are connected by dynamical relations; the errors in the relations are believed to be so much smaller than those in the data, that for practical purposes, the constraints (6.1) might as well be treated as though they are perfect.[36] But otherwise, the distinction between constraints (6.1) and the observations is an arbitrary one, and the introduction of an error term in the former, no matter how small, removes any particular reason to distinguish them: $\mathbf{A}$ may well be some subset of the rows of $\mathbf{E}$. What follows can in fact be obtained by imposing the zero noise limit for some of the rows of $\mathbf{E}$ in the solutions already described. Furthermore, whether the model should be satisfied exactly, or should contain a noise element too, is situation dependent. One should be wary of introducing exact equalities into estimation problems, because they carry the strong possibility of introducing small eigenvalues, or near singular relationships, into the solution, and which may dominate the results. Nonetheless, carrying one or more perfect constraints does produce some insight into how the system is behaving.

Several approaches are possible. Consider for example, the objective function,

$$J = (\mathbf{E}\mathbf{x} - \mathbf{y})^T(\mathbf{E}\mathbf{x} - \mathbf{y}) + \alpha^2(\mathbf{A}\mathbf{x} - \mathbf{d})^T(\mathbf{A}\mathbf{x} - \mathbf{d}) \tag{6.2}$$

where $\mathbf{W}$, $\mathbf{S}$ have been applied if necessary and α^2 is retained as a trade-off parameter. This objective function corresponds to the requirement of a solution of the combined equation sets,

$$\begin{Bmatrix} \mathbf{E} \\ \alpha^2\mathbf{A} \end{Bmatrix} \mathbf{x} + \begin{bmatrix} \mathbf{n} \\ \alpha^2\mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{y} \\ \alpha^2\mathbf{d} \end{bmatrix} \tag{6.3}$$

[36] Munk and Wunsch (1982).

in which $\mathbf{u}$ is the model noise. For any finite α^2, the perfect constraints are formally "soft" because they are being applied only as a minimized sum of squares. The solution follows immediately from (4.9) with

$$\mathbf{E} \longrightarrow \left\{ \begin{matrix} \mathbf{E} \\ \alpha^2 \mathbf{A} \end{matrix} \right\}, \qquad \mathbf{y} \longrightarrow \left\{ \begin{matrix} \mathbf{y} \\ \alpha^2 \mathbf{d} \end{matrix} \right\},$$

assuming the matrix inverse exists. As $\alpha^2 \to \infty$, the second set of equations is being imposed with arbitrarily great accuracy, and barring numerical issues, becomes as exactly satisfied as one wants (this is an example of a "penalty method").

Alternatively, the model can be applied as a hard constraint. All prior covariances and scalings having been applied, and Lagrange multipliers introduced, reduces the problem to one with an objective function

$$J = \mathbf{n}^T\mathbf{n} - 2\boldsymbol{\mu}^T(\mathbf{A}\mathbf{x} - \mathbf{d}) = (\mathbf{E}\mathbf{x} - \mathbf{y})^T(\mathbf{E}\mathbf{x} - \mathbf{y}) - 2\boldsymbol{\mu}^T(\mathbf{A}\mathbf{x} - \mathbf{d}), \tag{6.4}$$

which is just a variant of (4.59). But now, (6.1) is to be exactly satisfied, and the observations only approximately so.

Setting the derivatives of J with respect to $\mathbf{x}$, $\boldsymbol{\mu}$ to zero, gives the normal equations,

$$\mathbf{A}^T\boldsymbol{\mu} = \mathbf{E}^T(\mathbf{E}\mathbf{x} - \mathbf{y}) \tag{6.5}$$

$$\mathbf{A}\mathbf{x} = \mathbf{d} \tag{6.6}$$

Eq. (6.5) represents the adjoint, or "dual" model, for the adjoint or dual solution $\boldsymbol{\mu}$, and the two equation sets are to be solved simultaneously for $\mathbf{x}, \boldsymbol{\mu}$. They are again $M + N$ equations in $M + N$ unknowns (M of the μ_i, N of the x_i), but need not be full-rank. The first set, sometimes referred to as the "adjoint model," determines $\boldsymbol{\mu}$ from the *difference between* $\mathbf{Ex}$, and $\mathbf{y}$. The last set is just the exact constraints.

We can most easily solve two extreme cases in Eqs. (6.5, 6.6)— one in which $\mathbf{A}$ is square, $N \times N$, and of full-rank, and one in which $\mathbf{E}$ has this property. In the first case,

$$\tilde{\mathbf{x}} = \mathbf{A}^{-1}\mathbf{d} \tag{6.7}$$

and,

$$\tilde{\boldsymbol{\mu}} = \mathbf{A}^{-T}(\mathbf{E}^T\mathbf{E}\mathbf{A}^{-1} - \mathbf{E}^T)\mathbf{d}. \tag{6.8}$$

Here, the values of $\tilde{\mathbf{x}}$ are completely determined by the full-rank, noiseless model and the minimization of the deviation from the observations is passive. The Lagrange multipliers or adjoint solution, however, are useful in providing the sensitivity information, $\partial J/\partial \mathbf{d} = 2\boldsymbol{\mu}$, as already discussed. The uncertainty of this solution is zero because of the perfect model 6.6.

When $\mathbf{E}$ is full-rank, $K = N$,, from 6.5,

$$\tilde{\mathbf{x}} = (\mathbf{E}^T\mathbf{E})^{-1}[\mathbf{E}^T\mathbf{y} + \mathbf{A}^T\boldsymbol{\mu}] \equiv \tilde{\mathbf{x}}_u + (\mathbf{E}^T\mathbf{E})^{-1}\mathbf{A}^T\boldsymbol{\mu}$$

where $\tilde{\mathbf{x}}_u = (\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\mathbf{y}$ is the ordinary, unconstrained least-squares solution. Substituting into (6.6) produces

$$\tilde{\boldsymbol{\mu}} = [\mathbf{A}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{A}^T]^{-1}(\mathbf{d} - \mathbf{A}\tilde{\mathbf{x}}_u) \tag{6.9}$$

and

$$\tilde{\mathbf{x}} = \tilde{\mathbf{x}}_u + (\mathbf{E}^T\mathbf{E})^{-1}\mathbf{A}^T[\mathbf{A}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{A}^T]^{-1}(\mathbf{d} - \mathbf{A}\tilde{\mathbf{x}}_u)\,, \tag{6.10}$$

assuming $\mathbf{A}$ is full-rank underdetermined. The perfect model is underdetermined; its range is being fit perfectly, with its nullspace being employed to reduce the misfit to the data as far as possible. The uncertainty of this solution may be written[37],

$$\begin{aligned}\mathbf{P} \;&=\; D^2(\tilde{\mathbf{x}} - \mathbf{x}) && (6.11)\\ &= \sigma^2\left\{(\mathbf{E}^T\mathbf{E})^{-1} - (\mathbf{E}^T\mathbf{E})^{-1}\mathbf{A}^T(\mathbf{A}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{A}^T)^{-1}\mathbf{A}(\mathbf{E}^T\mathbf{E})^{-1}\right\},\end{aligned}$$

which represents a reduction in the uncertainty of the ordinary least-squares solution (first term on the right) by the information in the perfectly known constraints. The presence of $\mathbf{A}^{-1}$ in these solutions is a manifestation of the warning about the possible introduction of components dependent upon small eigenvalues of $\mathbf{A}$. If neither $\mathbf{E}^T\mathbf{E}$ nor $\mathbf{A}$ is of full-rank one can use, e.g., the SVD with the above solution; the combined $\mathbf{E}, \mathbf{A}$ may be rank deficient, or just determined.

EXAMPLE 14. *Consider the least-squares problem of solving*

$$x_1 + n_1 = 1$$

$$x_2 + n_2 = 1$$

$$x_1 + x_2 + n_3 = 3$$

with uniform, uncorrelated noise of variance 1 *in each of the equations. The least-squares solution is then*

$$\tilde{\mathbf{x}} = \begin{bmatrix}1.3333 & 1.3333\end{bmatrix}^T$$

with uncertainty

$$\mathbf{P} = \begin{Bmatrix}0.6667 & -0.3333\\ -0.333 & 0.6667\end{Bmatrix}.$$

But suppose that it is known or desired that $x_1 - x_2 = 1$*. Then* (6.10) *produces* $\tilde{\mathbf{x}} = [1.8333 \quad 0.8333]^T$, $\boldsymbol{\mu} = 0.5$, $J' = 0.8333$, *with uncertainty*

$$\mathbf{P} = \begin{Bmatrix}0.1667 & 0.1667\\ 0.1667 & 0.1667\end{Bmatrix}.$$

If the constraint is shifted to $x_1 - x_2 = 1.1$*, the new solution is* $\tilde{\mathbf{x}} = \begin{bmatrix}1.8833 & 0.7833\end{bmatrix}^T$ *and the new objective function is* $J' = 0.9383$*, consistent with the sensitivity deduced from* $\boldsymbol{\mu}$*.*

[37] Seber (1977).

A more generally useful case occurs when the errors normally expected to be present in the supposedly exact constraints are explicitly acknowledged. If the exact constraints have errors either in the "forcing," $\mathbf{d}$, or in a mis-specification of $\mathbf{A}$, then we write,

$$\mathbf{Ax} = \mathbf{d} + \mathbf{\Gamma u}\,, \tag{6.12}$$

assuming, $\langle \mathbf{u} \rangle = 0$, $\left\langle \mathbf{uu}^T \right\rangle = \mathbf{Q}$. $\mathbf{\Gamma}$ is a known coefficient matrix included for generality: If for example the errors were thought to be the same in all equations, we could write $\Gamma = [1, 1, ...1]^T$, and then $\mathbf{u}$ would be just a scalar. Let the dimension of $\mathbf{u}$ by $P \times 1$. Such representations are not unique and more will be said about them in Chapter 4. A hard constraint formulation can still be used, in which (6.12) is still to be exactly satisfied, imposed through an objective function of form,

$$J = (\mathbf{Ex} - \mathbf{y})^T \mathbf{R}_{nn}^{-1} (\mathbf{Ex} - \mathbf{y}) + \mathbf{u}^T \mathbf{Q}^{-1} \mathbf{u} - 2\boldsymbol{\mu}^T (\mathbf{Ax} - \mathbf{d} - \mathbf{\Gamma u})\,. \tag{6.13}$$

Here, we have explicitly included the noise error covariance matrix. Finding the normal equations by setting the derivatives with respect to $(\mathbf{x}, \mathbf{u}, \boldsymbol{\mu})$ to zero produces,

$$\begin{aligned} \mathbf{A}^T \boldsymbol{\mu} &= \mathbf{E}^T \mathbf{R}_{nn}^{-1} (\mathbf{Ex} - \mathbf{y}) && (6.14) \\ \mathbf{\Gamma}^T \boldsymbol{\mu} &= \mathbf{Q}^{-1} \mathbf{u} && (6.15) \\ \mathbf{Ax} + \mathbf{\Gamma u} &= \mathbf{d} && (6.16) \end{aligned}$$

This system is $(2N + P)$ equations in $(2N + P)$ unknowns, where the first equation is again the adjoint system, and dependent upon $\mathbf{Ex} - \mathbf{y}$. Because $\mathbf{u}$ is simple function of the Lagrange multipliers, the system is easily reduced to,

$$\begin{aligned} \mathbf{A}^T \boldsymbol{\mu} &= \mathbf{E}^T \mathbf{R}_{nn}^{-1} (\mathbf{Ex} - \mathbf{y}) && (6.17) \\ \mathbf{Ax} + \mathbf{\Gamma Q \Gamma}^T \boldsymbol{\mu} &= \mathbf{d} && (6.18) \end{aligned}$$

which is now $2N \times 2N$, the $\mathbf{u}$ having dropped out. If $\mathbf{A}$ is full-rank, the solution is immediate; otherwise the SVD is used.

To use a soft constraint methodology, write

$$J = (\mathbf{Ex} - \mathbf{y})^T \mathbf{R}_{nn}^{-1} (\mathbf{Ex} - \mathbf{y}) + (\mathbf{Ax} - \mathbf{q} - \Gamma \mathbf{u})^T \mathbf{Q}^{-1} (\mathbf{Ax} - \mathbf{q} - \Gamma \mathbf{u})^T, \tag{6.19}$$

and find the normal equations. It is again readily confirmed that the solutions using (6.3) or 6.19 are identical, and the hard/soft distinction is seen again to be artificial. The soft constraint method can deal with perfect constraints, by letting $\left\| \mathbf{Q}^{-1} \right\| \to 0$ but stopping when numerical instability sets in. The resulting numerical algorithms fall under the general subject of "penalty" and "barrier" methods.[38]

[38] Luenberger (1984).

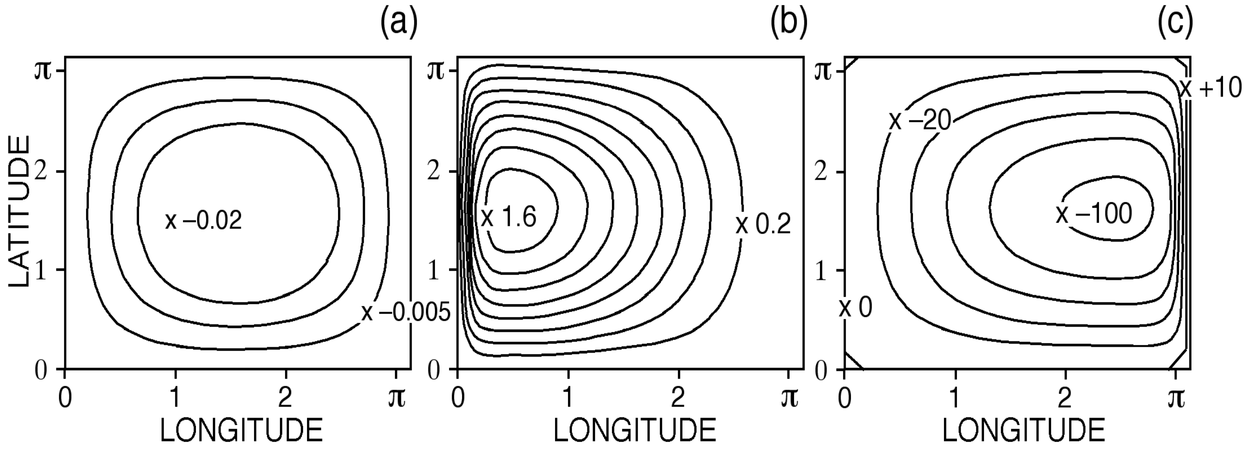


FIGURE 13. Numerical solution of the partial differential equation, Eq. 6.20. Panel (a) shows the imposed symmetric forcing $-\sin x \sin y$.(b) Displays the solution ϕ, and (c) shows the Lagrange multipliers, or adjoint solution, μ and which is a mirror image of ϕ.

EXAMPLE 15. A Partial Differential Equation[39] *Consider the partial differential equation,*

$$\epsilon\nabla^2\phi + \frac{\partial\phi}{\partial x} = -\sin x \sin y. \tag{6.20}$$

A code was written to solve it by finite differences for the case $\epsilon = 0.05$ *and* $\phi = 0$ *on the boundaries* $0 \leq x \leq \pi$, $0 \leq y \leq \pi$, *as depicted in figure 13 . The discretized form of the model is then the perfect* $N \times N$ *system*

$$\mathbf{Ax} = \mathbf{q}, \quad \mathbf{x} = \{\phi_{ij}\} \tag{6.21}$$

and $\mathbf{q}$ *is the equivalently discretized* $-\sin x \sin y$*. The theory of partial differential equations shows that this system is full-rank and generally well-behaved. But let us pretend that information is unknown to us, and seek the values* $\mathbf{x}$ *which makes the objective function*

$$J = \mathbf{x}^T\mathbf{x} - 2\boldsymbol{\mu}^T(\mathbf{Ax} - \mathbf{q}) \tag{6.22}$$

stationary with respect to $\mathbf{x}$, $\boldsymbol{\mu}$, *that is the Eqs. (6.5, 6.6) with* $\mathbf{E} = \mathbf{I}$, $\mathbf{y} = \mathbf{0}$.. *Physically,* $\mathbf{x}^T\mathbf{x}$ *is identified with the solution potential energy. The solution* $\boldsymbol{\mu}$, *corresponding to the circulation of fig. 13bis shown in fig. 13c. What is the interpretation? The Lagrange multipliers represent the sensitivity of the solution potential energy to perturbations in the forcing field. The sensitivity is greatest in the right-half of the basin, and indeed displays a boundary layer character. A physical interpretation of the Lagrange multipliers can be inferred, given the simple structure*

[39] Schröter and Wunsch (1986).

of the governing equation (6.20)*, and the Dirichlet boundary conditions. This equation is not self-adjoint; the adjoint partial differential equation is of form,*

$$\epsilon\nabla^2\boldsymbol{\mu} - \frac{\partial\boldsymbol{\mu}}{\partial x} = d, \tag{6.23}$$

where d is a forcing term, subject to mixed boundary conditions, and whose discrete form is obtained by taking the transpose of the **A** *matrix of the discretization (See the Chapter Appendix.) The forward solution exhibits a boundary layer on the left-hand wall, while the adjoint solution has a corresponding behavior in the dual space on the right-hand wall. The structure of the* $\boldsymbol{\mu}$ *would evidently change if J were changed; in the present case, J immediately describes the sensitivity of the potential energy to any perturbation in q.*[40]

The original objective function J is very closely analogous to the Lagrangian (not to be confused with the Lagrange multiplier) in classical mechanics. In mechanics, the gradients of the Lagrangian commonly are forces. The modified Lagrangian, J', is used in mechanics to impose various physical constraints, and the virtual force required to impose the constraints, for example, the demand that a particle follow a particular path, is the Lagrange multiplier.[41] In an economics/management context, the multipliers are usually called "shadow prices" as they are intimately related to the question of how much profit will change with a shift in the availability or cost of a product ingredient.

More generally, there is a close connection between the stationarity requirements imposed upon various objective functions throughout this book, and the mathematics of classical mechanics. An elegant Hamiltonian formulation of the material in Chapter 4 is possible.

6.2. Relation to Green Functions[42]**.** Consider any linear set of simultaneous equations, involving a *square* $N \times N$ matrix, $\mathbf{A}$,

$$\mathbf{Ax} = \mathbf{d}. \tag{6.24}$$

First consider the adjoint equations, for an arbitrary right-hand-side,

$$\mathbf{A}^T\mathbf{z} = \mathbf{r}. \tag{6.25}$$

Then the simple scalar identity,

$$\mathbf{z}^T\mathbf{Ax} - \mathbf{x}^T\mathbf{A}^T\mathbf{z} = 0 \tag{6.26a}$$

(the "bilinear identity") implies,

$$\mathbf{z}^T\mathbf{d} = \mathbf{x}^T\mathbf{r}. \tag{6.27}$$

[40] In oceanographic terms, the exact constraints describe the Stommel Gulf Stream solution. The eastward intensification of the adjoint solution corresponds to the change in sign of β in the adjoint model. See Schröter and Wunsch (1986) for details and an elaboration to a non-linear situation.

[41] Lanczos (1970) has a good discussion.

[42] See Lanczos (1961, Section 3.19)

In the special case, $\mathbf{r}=\mathbf{0}$, we have

$$\mathbf{z}^T\mathbf{d} =0, \tag{6.28}$$

that is, $\mathbf{d}$, the right-hand side of the original equations (6.24), must be orthogonal to any solution of the homogeneous adjoint equations. (In SVD-terms, this result is nothing but the solvability condition Eq. (5.76)). If $\mathbf{A}$ is of full rank, then there is no non-zero solution to the homogeneous adjoint equations.

Now assume that $\mathbf{A}$ is indeed full rank. We add a single equation to (6.24) of the form

$$x_p=\alpha_p \tag{6.29}$$

or

$$\mathbf{e}_p^T\mathbf{x} =\alpha_p, \tag{6.30}$$

where $\mathbf{e}_p=\delta_{ip}$ (the vector which is all zeros, except for 1 at position p). α_p is unknown, but we also demand that Eq. (6.24) should remain exactly satisfied. The combined system of (6.24) and (6.30), written as,

$$\mathbf{A}_1\mathbf{x}=\mathbf{d}_1 \tag{6.31}$$

is overdetermined. If it is to have a solution without any residual, it must be orthogonal to any solution of the homogeneous adjoint equations,

$$\mathbf{A}_1^T\mathbf{z}=\mathbf{0}. \tag{6.32}$$

There is only one such solution (because there is only one vector, $\mathbf{z}=\mathbf{u}_{N+1}$, in the null space of $\mathbf{A}_1^T$). Write $\mathbf{u}_{N+1}=[\mathbf{g}_p,\gamma]^T$, separating out the first N elements of $\mathbf{u}_{N+1}$, calling them $\mathbf{g}_p$, and calling the one remaining element γ. Thus the solvability condition is,

$$\mathbf{u}_{N+1}^T\mathbf{d}_1=\mathbf{g}_p^T\mathbf{d}+\gamma\alpha_p=0. \tag{6.33}$$

Choose $\gamma=-1$ (any other finite choice can be absorbed into $\mathbf{g}$). Then,

$$\alpha_p=\mathbf{g}_p^T\mathbf{d}. \tag{6.34}$$

If $\mathbf{g}_p$ were known, then the value of α_p in (6.34) would be the only value consistent with the solutions to (6.24), and would be the correct value of $x_{p.}$ But (6.32) is the same as,

$$\mathbf{A}^T\mathbf{g}_p= \mathbf{e}_p \tag{6.35}$$

(recalling $\gamma=-1$),and which can be solved because $\mathbf{e}_p$ is known. Because we would like to find *all* elements x_p, we would need to solve (6.35) for all $1\leq p\leq N$, that is,

$$\mathbf{A}^T\mathbf{G}=\mathbf{I}_N \tag{6.36}$$

which is N separate problems, each for the corresponding column of $\mathbf{G} = \{\mathbf{g}_1, \mathbf{g}_2, ...\mathbf{g}_N\}$. Here, $\mathbf{G}$ is the Green function. With $\mathbf{G}$ known, we have immediately,

$$\mathbf{x} = \mathbf{G}^T \mathbf{d}, \tag{6.37}$$

(from Eq. (6.34)). The Green function is an inverse to the adjoint equations (and generalizes in the continuous case to an operator inverse).

7. Gauss-Markov (Minimum Variance) Estimation and Simultaneous Equations

The fundamental objective for least-squares is minimization of the noise norm (4.4), although we complicated the discussion somewhat by introducing trade-offs against $\|\tilde{\mathbf{x}}\|$, various weights in the norms, and even the restriction that $\tilde{\mathbf{x}}$ should satisfy certain equations exactly. Least-squares methods, whether used directly as in (4.9) or indirectly through the vector representations of the SVD, are fundamentally deterministic—$\mathbf{W}$, $\mathbf{S}$, α^2 need be given no statistical interpretation—although sometimes one uses covariances for them. Statistics were used only to understand the sensitivity of the solutions to noise, and to obtain measures of the expected deviation of the solution from some supposed truth.

But there is another, radically different, approach to obtaining estimates of the solution to equation sets like (4.2), directed more clearly toward the physical goal: to find an estimate $\tilde{\mathbf{x}}$ which deviates as little as possible in the *mean-square* from the true solution. That is, we wish to minimize the statistical quantities $\langle(\tilde{\mathbf{x}}_i - \mathbf{x}_i)^2\rangle$ for all i. The next section is devoted to understanding how to find such an $\tilde{\mathbf{x}}$ (and the corresponding $\tilde{\mathbf{n}}$), through an excursion into statistical estimation theory. It is far from obvious that this $\tilde{\mathbf{x}}$ should bear any resemblance to one of the least-squares estimates; but as will be seen, under some circumstances the two are identical. Their possible identity is extremely useful, but has apparently led many investigators to seriously confuse the methodologies, and therefore the interpretation of the result.

7.1. The Fundamental Result. Suppose we are interested in making an estimate of a physical variable, $\mathbf{x}$, which might be a vector or a scalar, and might be constant with space and time, but which may vary with either or both. To be definite, let $\mathbf{x}$ be a function of an independent variable $\mathbf{r}$, written discretely as $\mathbf{r}_j$ (it might be a vector of space coordinates, or a scalar time, or an accountant's label). Let us make some suppositions about what is usually called "prior information." In particular, suppose we have an estimate of the low-order statistics describing $\mathbf{x}$, that is, we specify its mean and second moments:

$$\langle \mathbf{x} \rangle \ = \mathbf{0}\,, \qquad \langle \mathbf{x}(\mathbf{r}_i)\mathbf{x}(\mathbf{r}_i)^T \rangle = \mathbf{R}_{xx}(\mathbf{r}_i,\, \mathbf{r}_j)\,. \tag{7.1}$$

To make this problem concrete, one might think of $\mathbf{x}$ as being the temperature anomaly (about the mean) at a fixed depth in the ocean (a scalar) and $\mathbf{r}_j$ a vector of horizontal positions; or conductivity in a well, where $\mathbf{r}_j$ would be the depth coordinate, and $\mathbf{x}$ is the vector of scalars at any location, $\mathbf{r}_p$, $x_p = x\left(\mathbf{r}_p\right)$. Alternatively, $\mathbf{x}$ might be the temperature at a fixed point, with r_j being the scalar of time. But if the field of interest is the velocity vector, then each element of $\mathbf{x}$ is itself a vector, and one can extend the notation in a straightforward fashion. To keep the notation a little cleaner, however, we will treat the elements of $\mathbf{x}$ as scalars.

Now suppose that we have some observations, y_i, as a function of the same coordinate $\mathbf{r}_i$, with a known, zero mean, and second moments

$$\mathbf{R}_{yy}\left(\mathbf{r}_i, \mathbf{r}_j\right) = \left\langle \mathbf{y}\left(\mathbf{r}_i\right)\mathbf{y}\left(\mathbf{r}_j\right)^T\right\rangle, \qquad \mathbf{R}_{xy}(\mathbf{r}_i,\, \mathbf{r}_j) = \left\langle \mathbf{x}(\mathbf{r}_i)\mathbf{y}(\mathbf{r}_j)^T\right\rangle, \quad 1 \le i, j \le M \tag{7.2}$$

(the individual observation elements can also be vectors—for example, two or three components of velocity and a temperature at a point—but as with $\mathbf{x}$, the modifications required to treat this case are straightforward, and we here assume scalar observations). Could the measurements be used to make an estimate of $\mathbf{x}$ at a point $\tilde{\mathbf{r}}_\alpha$ where no measurement is available? Or to use many measurements to obtain a better estimate even at points where there is a measurement? The idea is to exploit the concept that finite covariances carry predictive capabilities from known variables to unknown ones. A specific example would be to suppose the measurements are of temperature $y(\mathbf{r}_j) = y_0(\mathbf{r}_j) + n(\mathbf{r}_j)$, where n is the noise and we wish to estimate the temperature at different locations, perhaps on a regular grid $\tilde{\mathbf{r}}_\alpha$, $1 \le \alpha \le N$. This special problem is one of gridding. or mapmaking (the tilde is placed on $\mathbf{r}_\alpha$ as a device to emphasize that this is a location where an estimate is sought; the numerical values of these places or labels are assumed known). Alternatively, and somewhat more interesting, perhaps the measurements are more indirect, with $y(r_i)$ representing a velocity field component at depth in the ocean and believed connected through the thermal wind equation to the temperature field. We might want to estimate the temperature from measurements of the velocity.

It is reasonable to ask for an estimate $\tilde{x}(\tilde{\mathbf{r}}_\alpha)$, whose dispersion about its true value, $x(\tilde{\mathbf{r}}_\alpha)$ is as small as possible, that is,

$$P\left(\tilde{\mathbf{r}}_\alpha, \tilde{\mathbf{r}}_\alpha\right) = \left\langle (\tilde{x}(\tilde{\mathbf{r}}_\alpha) - x(\tilde{\mathbf{r}}_\alpha))(\tilde{x}(\tilde{\mathbf{r}}_\beta) - x(\tilde{\mathbf{r}}_\beta))\right\rangle\big|_{\tilde{\mathbf{r}}_\alpha = \tilde{\mathbf{r}}_\beta}$$

is to be minimized. If we would like to answer the question for more than one point, and if we would like to understand the covariance of the errors of our estimates at various points $\tilde{\mathbf{r}}_\alpha$, then we can form a vector of values to be estimated, $\{\tilde{x}(\mathbf{r}_\alpha)\} \equiv \tilde{\mathbf{x}}$ and the uncertainty among them,

$$\begin{aligned}\mathbf{P}(\tilde{\mathbf{r}}_\alpha,\, \tilde{\mathbf{r}}_\beta) &= \left\langle (\tilde{x}(\tilde{\mathbf{r}}_\alpha) - x(\tilde{\mathbf{r}}_\alpha))(\tilde{x}(\tilde{\mathbf{r}}_\beta) - x(\tilde{\mathbf{r}}_\beta))\right\rangle \\ &= \left\langle (\tilde{\mathbf{x}} - \mathbf{x})(\tilde{\mathbf{x}} - \mathbf{x})^T\right\rangle, \quad 1 \le \alpha \le N,\ 1 \le \beta \le N,\end{aligned} \tag{7.3}$$

where the *diagonal* elements, $\mathbf{P}(\tilde{\mathbf{r}}_\alpha,\, \tilde{\mathbf{r}}_\alpha)$, are to be *individually* minimized (not in the sum of squares). Thus we seek the solution with *minimum variance about the correct value.*

What should the relationship be between data and estimate? At least initially, one might try a linear combination of data,

$$\tilde{x}(\tilde{\mathbf{r}}_\alpha) = \sum_{j=1}^{M} B(\tilde{\mathbf{r}}_\alpha,\, \mathbf{r}_j) y(\mathbf{r}_j)\,, \tag{7.4}$$

for all α, which makes the diagonal elements of $\mathbf{P}$ in (7.3) as small as possible. We can treat all the points $\tilde{\mathbf{r}}_\alpha$ simultaneously by letting $\mathbf{B}$ be an $M \times N$ matrix, and

$$\tilde{\mathbf{x}}\left(\mathbf{r}_\alpha\right) = \mathbf{B}(\tilde{\mathbf{r}}_\alpha,\, \mathbf{r}_j)\mathbf{y}\left(\mathbf{r}_j\right)\;. \tag{7.5}$$

(This notation is mixed. Eq. (7.5) is a shorthand for (7.4), in which the argument has been put into $\mathbf{B}$ explicitly as a reminder that there is a summation over all the data locations $\mathbf{r}_j$ for all mapping locations $\tilde{\mathbf{r}}_\alpha$, but it is automatically accounted for by the usual matrix multiplication convention.)

An important result, usually called the "Gauss-Markov theorem," produces the values of $\mathbf{B}$ that will minimize the diagonal elements of $\mathbf{P}$.[43] Substituting (7.5) into (7.3) and expanding,

$$\begin{aligned}
&\mathbf{P}(\tilde{\mathbf{r}}_\alpha,\, \tilde{\mathbf{r}}_\beta) = \langle(\mathbf{B}(\tilde{\mathbf{r}}_\alpha,\, \mathbf{r}_j)\mathbf{y}\left(\mathbf{r}_j\right) - x(\tilde{\mathbf{r}}_\alpha))(\mathbf{B}(\tilde{\mathbf{r}}_\beta,\, \mathbf{r}_l)\mathbf{y}\left(\mathbf{r}_l\right) - x(\tilde{\mathbf{r}}_\beta))^T\rangle \\
&\equiv \langle(\mathbf{B}\mathbf{y} - \mathbf{x})(\mathbf{B}\mathbf{y} - \mathbf{x})^T\rangle \\
&= \mathbf{B}\left(\tilde{\mathbf{r}}_\alpha, \mathbf{r}_j\right)\langle\mathbf{y}\left(\mathbf{r}_j\right)\mathbf{y}\left(\mathbf{r}_l\right)^T\rangle\mathbf{B}\left(\mathbf{r}_\beta, \mathbf{r}_l\right)^T - \\
&\langle\mathbf{x}\,\tilde{(}\mathbf{r}_\alpha)\mathbf{y}\left(\mathbf{r}_l\right)^T\rangle\mathbf{B}\left(\mathbf{r}_\alpha, \mathbf{r}_l\right)^T - \mathbf{B}\left(\tilde{\mathbf{r}}_\alpha, \mathbf{r}_j\right)\langle\mathbf{y}\left(\mathbf{r}_j\right)\mathbf{x}\left(\tilde{\mathbf{r}}_\beta\right)^T\rangle + \langle\mathbf{x}\left(\tilde{\mathbf{r}}_\alpha\right)\mathbf{x}\left(\tilde{\mathbf{r}}_\beta\right)^T\rangle
\end{aligned} \tag{7.6}$$

Using $\mathbf{R}_{xy} = \mathbf{R}_{yx}^T$, eq. (7.6) is,

$$\mathbf{P} = \mathbf{B}\mathbf{R}_{yy}\mathbf{B}^T - \mathbf{R}_{xy}\mathbf{B}^T - \mathbf{B}\mathbf{R}_{xy} + \mathbf{R}_{xx}\,. \tag{7.7}$$

Notice that because $\mathbf{R}_{xx}$ represents the moments of $\mathbf{x}$ evaluated at the estimation positions, it is a function of $\tilde{\mathbf{r}}_\alpha$, $\tilde{\mathbf{r}}_\beta$, whereas $\mathbf{R}_{xy}$ involves covariances of $\mathbf{y}$ at the data positions with $\mathbf{x}$ at the estimation positions, and is consequently a function $\mathbf{R}_{xy}(\tilde{\mathbf{r}}_\alpha,\, \mathbf{r}_j)$.

Now, using the matrix identity (2.34)—that is, completing the square, (7.7) becomes

$$\mathbf{P} = (\mathbf{B} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1})\mathbf{R}_{yy}(\mathbf{B} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1})^T - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{xy}^T + \mathbf{R}_{xx}\,. \tag{7.8}$$

Setting $\tilde{\mathbf{r}}_\alpha = \tilde{\mathbf{r}}_\beta$ so that (7.8) is the variance of the estimate at point $\tilde{\mathbf{r}}_\alpha$ about its true value, and noting that all three terms in Eq. (7.8) are positive definite, minimization of any diagonal element of $\mathbf{P}$ is obtained by choosing $\mathbf{B}$ so that the first term vanishes or,

$$\mathbf{B}(\tilde{\mathbf{r}}_\alpha,\, \mathbf{r}_j) = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\,. \tag{7.9}$$

[43] The derivation follows Liebelt (1967).

(The diagonal elements of $(\mathbf{B}-\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1})\mathbf{R}_{yy}(\mathbf{B}-\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1})^T$ need to be written out explicitly to see that Eq. (7.9) is necessary.) Thus the minimum variance estimate is

$$\tilde{\mathbf{x}}(\tilde{\mathbf{r}}_\alpha)=\mathbf{R}_{xy}\left(\tilde{\mathbf{r}}_\alpha,\mathbf{r}_i\right)\mathbf{R}_{yy}^{-1}\left(\mathbf{r}_i,\mathbf{r}_j\right)\mathbf{y}\left(\mathbf{r}_j\right)\ , \tag{7.10}$$

and the actual minimum value of the diagonal elements of $\mathbf{P}$ is found by substituting back into (7.7), producing

$$\mathbf{P}(\tilde{\mathbf{r}}_\alpha,\,\tilde{\mathbf{r}}_\beta)=\mathbf{R}_{xx}(\tilde{\mathbf{r}}_\alpha,\,\tilde{\mathbf{r}}_\beta)-\mathbf{R}_{xy}(\tilde{\mathbf{r}}_\alpha,\,\mathbf{r}_j)\mathbf{R}_{yy}^{-1}(\mathbf{r}_j,\,\mathbf{r}_k)\mathbf{R}_{xy}^T(\tilde{\mathbf{r}}_\beta,\,\mathbf{r}_k)\,. \tag{7.11}$$

The bias of (7.11) is

$$\langle\tilde{\mathbf{x}}-\ \mathbf{x}\rangle=\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\langle\mathbf{y}\rangle-\mathbf{x}\,. \tag{7.12}$$

If $\langle\mathbf{y}\rangle=\mathbf{x}=0$, the estimator is unbiassed , and called a "best linear unbiassed estimator," or "BLUE"; otherwise it is biassed. The whole development here began with the assumption that $\langle\mathbf{x}\rangle=\langle\mathbf{y}\rangle=0$; what is usually done is to remove the *sample* mean from the observations $\mathbf{y}$, and to ignore the difference between the true and sample means. Under some circumstances, this approximation is unacceptable, and one must account for the mapping error introduced by the use of the sample mean. A general approach falls under the label of "kriging", and which is briefly discussed in Chapter 3.

7.2. Linear Algebraic Equations. The result (7.9)–(7.11) is the abstract general case and is deceptively simple. Invocation of the physical problem of interpolating temperatures etc., is not necessary: the only information actually used is that there are finite covariances between $\mathbf{x},\mathbf{y},\mathbf{n}$. Although we will explicitly explore its use for mapping in Chapter 5.3, suppose that the observations are related to the unknown vector $\mathbf{x}$ as in our canonical problem, that is, through a set of linear equations: $\mathbf{E}\mathbf{x}+\mathbf{n}=\mathbf{y}$. The measurement covariance, $\mathbf{R}_{yy}$, can then be computed directly as:

$$\mathbf{R}_{yy}=\langle(\mathbf{E}\mathbf{x}+\mathbf{n})(\mathbf{E}\mathbf{x}+\mathbf{n})^T\rangle=\mathbf{E}\mathbf{R}_{xx}\mathbf{E}^T+\mathbf{R}_{nn}. \tag{7.13}$$

The unnecessary, but simplifying and often excellent, assumption was made that the cross-terms of form,

$$\mathbf{R}_{xn}=\mathbf{R}_{nx}^T=\mathbf{0}, \tag{7.14}$$

so that

$$\mathbf{R}_{xy}=\langle\mathbf{x}(\mathbf{E}\mathbf{x}+\mathbf{n})^T\rangle=\mathbf{R}_{xx}\mathbf{E}^T\ , \tag{7.15}$$

that is, there is no correlation between the measurement noise and the actual state vector (e.g., that the noise in a temperature measurement does not depend upon whether the true value is 10° or 25°). (The arguments $\tilde{\mathbf{r}}_\alpha,\mathbf{r}_j$, etc. are replaced by the corresponding matrix indices of $\mathbf{x},\mathbf{y}$ etc.)

Under these circumstances, eqs. (7.10), (7.11) take on the form:

$$
\begin{aligned}
\tilde{\mathbf{x}} &= \mathbf{R}_{xx}\mathbf{E}^T\left(\mathbf{E}\mathbf{R}_{xx}\mathbf{E}^T{+}\mathbf{R}_{nn}\right)^{-1}\mathbf{y} && (7.16)\\
\tilde{\mathbf{n}} &= \mathbf{y}-\mathbf{E}\tilde{\mathbf{x}} && (7.17)\\
\mathbf{P} &= \mathbf{R}_{xx}{-}\mathbf{R}_{xx}\mathbf{E}^T\left(\mathbf{E}\mathbf{R}_{xx}\mathbf{E}^T{+}\mathbf{R}_{nn}\right)^{-1}\mathbf{E}\mathbf{R}_{xx} && (7.18)
\end{aligned}
$$

These latter expression are extremely important; they permit discussion of the solution to a set of linear algebraic equations in the presence of noise using information concerning the statistics of the noise and of the solution. Notice that they are *identical to the least-squares expression* (4.47) *if* $\mathbf{S} = \mathbf{R}_{xx}$, $\mathbf{W} = \mathbf{R}_{nn}$, except there the uncertainty was estimated about the mean solution; here it is taken about the true one. As is generally true of all linear methods, the uncertainty, $\mathbf{P}$, is independent of the actual data, and can be computed in advance should one wish.

From the matrix inversion lemma, Eqs. (7.16, 7.18) can be rewritten

$$
\begin{aligned}
\tilde{\mathbf{x}} &= \left(\mathbf{R}_{xx}^{-1}{+}\mathbf{E}^T\mathbf{R}_{nn}^{-1}\mathbf{E}\right)^{-1}\mathbf{E}^T\mathbf{R}_{nn}^{-1}\mathbf{y} && (7.19)\\
\tilde{\mathbf{n}} &= \mathbf{y}-\mathbf{E}\tilde{\mathbf{x}} && (7.20)\\
\mathbf{P} &= \left(\mathbf{R}_{xx}^{-1}{+}\mathbf{E}^T\mathbf{R}_{nn}^{-1}\mathbf{E}\right)^{-1} && (7.21)
\end{aligned}
$$

Although these alternate forms are algebraically and numerically identical to the alternative least-squares form (4.39), the size of the matrices to be inverted changes from $M \times M$ matrices to $N \times N$, where $\mathbf{E}$ is $M \times N$ (but note that $\mathbf{R}_{nn}$ is $M \times M$; the efficacy of this alternate form may depend upon whether the *inverse* of $\mathbf{R}_{nn}$ is known). Depending upon the relative magnitudes of M, N, one form may be much preferable to the other; finally, (7.21) has an important interpretation we will discuss when we come to recursive methods. Recall, too, the options we had with the SVD of solving $M \times M$ or $N \times N$ problems. Note that in the limit of complete *a priori* ignorance of the solution, $\left\|\mathbf{R}_{xx}^{-1}\right\| \to 0$, Eqs. (7.19, 7.21) reduce to,

$$
\begin{aligned}
\tilde{\mathbf{x}}&{=}\left(\mathbf{E}^T\mathbf{R}_{nn}^{-1}\mathbf{E}\right)^{-1}\mathbf{E}^T\mathbf{R}_{nn}\mathbf{y},\\
\mathbf{P}&{=}\left(\mathbf{E}^T\mathbf{R}_{nn}^{-1}\mathbf{E}\right)^{-1},
\end{aligned}
$$

the conventional weighted least-squares solution, now with $\mathbf{P} = \mathbf{C}_{xx.}$ More generally, the presence of finite $\mathbf{R}_{xx}^{-1}$ introduces a bias into the solution so that $\langle\tilde{\mathbf{x}}\rangle \neq \mathbf{x}$, which however, produces a smaller solution variance than the unbiassed solution has (e.g., the truncated SVD solution).

The solution (7.16-7.18, 7.19-7.21) is an "estimator"; it was found by demanding a solution with the minimum dispersion about the true solution and is found, surprisingly, to be identical with the tapered, weighted least-squares solution when $\mathbf{S} = \mathbf{R}_{xx},\ \mathbf{W} = \mathbf{R}_{nn}$the least-squares objective function weights are chosen, as is commonly done. This correspondence of the two solutions often leads them to be seriously confused. It is essential to recognize that the logic of the derivations are distinct: We were free in the least-squares derivation to use weight matrices which were anything we wished—as long as appropriate inverses existed.

The correspondence of least-squares with what is usually known as minimum variance estimation can be understood by recognizing that the Gauss-Markov estimator was derived by minimizing a quadratic objective function. The least-squares estimate was obtained from minimizing a summation which is a sample *estimate* of the Gauss-Markov objective function when $\mathbf{S} = \mathbf{R}_{xx},\ \mathbf{W} = \mathbf{R}_{nn}.$

7.3. Testing After the Fact. As with any statistical estimator, an essential step is the testing after an apparent solution has been found, that the behavior of $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$ is consistent with the assumed prior statistics reflected in $\mathbf{R}_{xx}$, $\mathbf{R}_{nn}$, and any assumptions about their means or other properties. Such *a posteriori* checks are both necessary and very demanding. One sometimes hears it said that estimation using Gauss-Markov and related methods is "pulling solutions out of the air" because the prior covariance matrices $\mathbf{R}_{xx}$, $\mathbf{R}_{nn}$ often are only poorly known. But producing solutions which pass the test of consistency with the prior covariances can be very difficult. It is also true that the solutions tend to be somewhat insensitive to the details of the prior covariances and it is easy to become overly concerned with the detailed structure of $\mathbf{R}_{xx}$, $\mathbf{R}_{nn}$.

As stated previously, it is also rare to be faced with a situation in which one is truly ignorant of the covariances, true ignorance meaning that arbitrarily large or small numerical values of x_i, n_i would be acceptable. In the box inversions of Chapter 1 (to be revisited extensivley later), deep velocity fields of order 1000 cm/s might be absurd, and their absurdity is readily asserted by choosing $\mathbf{R}_{xx} = (10\text{ cm/s})^2$, which reflects a mild belief that velocities are 0(10 cm/s) with no known correlations with each other. Testing of statistical estimates against prior hypotheses is a highly developed field in applied statistics, and we leave it to the references already listed for their discussion. Should such tests be failed, one must reject the solutions $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$ and ask why they failed—as it usually implies an incorrect model ($\mathbf{E}$, and the assumed statistics of solution and/or noise).

7.4. Use of Basis Functions. A superficially different way of dealing with prior statistical information is often commonly used. Suppose that the indices of x_i refer to a spatial or temporal position, call it r_i, so that $x_i = x(r_i)$. Then it is often sensible to consider expanding the unknown $\mathbf{x}$ in a set of basis functions, F_j, for example, sines and cosines, Chebyschev polynomials, ordinary polynomials, etc. One might write

$$x(r_i) = \sum_{j=1}^{L} \alpha_j F_j(r_i)$$

or

$$\mathbf{x} = \mathbf{F}\boldsymbol{\alpha}, \quad \mathbf{F} = \begin{Bmatrix} F_1(r_1) & F_2(r_1) & \cdots & F_L(r_1) \\ F_1(r_2) & F_2(r_2) & \cdots & F_L(r_2) \\ \cdot & \cdot & \cdot & \cdot \\ F_1(r_N) & F_2(r_N) & \cdots & F_L(r_N) \end{Bmatrix}, \quad \boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \cdot \\ \cdot \\ \cdot \\ \alpha_L \end{bmatrix}^T$$

which, when substituted into (4.2), produces

$$\mathbf{L}\boldsymbol{\alpha} + \mathbf{n} = \mathbf{y}\,, \quad \mathbf{L} = \mathbf{E}\mathbf{F}\,. \tag{7.22}$$

If $L < M < N$, one can convert an underdetermined system into one which is formally overdetermined and, of course, the reverse is possible as well. It should be apparent, however, that the solution to (7.22) will have a covariance structure dictated in large part by that contained in the basis functions chosen, and thus there is no fundamental gain in employing basis functions although they may be convenient, numerically or otherwise. If $\mathbf{P}_{\alpha\alpha}$ denotes the uncertainty of $\boldsymbol{\alpha}$,then

$$\mathbf{P} = \mathbf{F}\mathbf{P}_{\alpha\alpha}\mathbf{F}^T \tag{7.23}$$

is the uncertainty of $\tilde{\mathbf{x}}$. If there are special conditions applying to $\mathbf{x}$, such as boundary conditions at certain positions, r_B, a choice of basis function satisfying those conditions could be more convenient than appending them as additional equations.

Example 16. *The underdetermined system*

$$\begin{Bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \end{Bmatrix} \mathbf{x} + \mathbf{n} = \begin{bmatrix} 1 \\ -1 \end{bmatrix},$$

with noise variance $\langle \mathbf{n}\mathbf{n}^T \rangle = .01\mathbf{I}$, has a solution, if $\mathbf{R}_{xx} = \mathbf{I}$, of

$$\tilde{\mathbf{x}} = \mathbf{E}^T(\mathbf{E}\mathbf{E}^T + .01\mathbf{I})^{-1}\mathbf{y} = \begin{bmatrix} 0 & .4988 & .4988 & 0 \end{bmatrix}^T, \qquad \tilde{\mathbf{n}} = \begin{bmatrix} .0025 & -.0025 \end{bmatrix}^T.$$

If the solution was thought to be large scale and smooth, one might use the covariance

$$\mathbf{R}_{xx} = \begin{Bmatrix} 1 & .999 & .998 & .997 \\ .999 & 1 & .999 & .998 \\ .998 & .999 & 1 & .999 \\ .997 & .998 & .999 & 1 \end{Bmatrix},$$

which produces a solution

$$\tilde{\mathbf{x}} = \begin{bmatrix} 0.2402 \pm 0.028 & 0.2595 \pm 0.0264 & 0.2595 \pm 0.0264 & 0.2402 \pm 0.0283 \end{bmatrix}^T ,$$

$$\tilde{\mathbf{n}} = \begin{bmatrix} 0.0006 & -0.9615 \end{bmatrix}^T ,$$

which has the desired large-scale property. (One might worry a bit about the structure of the residuals; but two equations are wholly inadequate to draw any conclusions.) If one attempts a solution as a first order polynomial,

$$x_i = a + br_i, \quad r_1 = 0,\ r_2 = 1,\ r_3 = 2, \ldots$$

the system will become two equations in the two unknowns a, b:

$$\mathbf{EF}\begin{bmatrix} a \\ b \end{bmatrix} = \begin{Bmatrix} 4 & 6 \\ 0 & 0 \end{Bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} + \mathbf{n} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} ,$$

and if no prior information about the covariance of a, b is provided,

$$[\tilde{a},\ \tilde{b}] = [0.0769,\ 0.1154] ,$$

$$\tilde{\mathbf{x}} = \begin{bmatrix} 0.0769 \pm 0.0077 & 0.1923 \pm 0.0192 & 0.3076 \pm 0.0308 & 0.4230 \pm 0.0423 \end{bmatrix} ,$$

$$\tilde{\mathbf{n}} = [0.0002,\ -1.00] ,$$

which is also large scale and smooth, but clearly different than that from the Gauss-Markov estimator. Although this latter solution has been obtained from a just-determined system, it is not clearly "better." If a linear trend is expected in the solution, then the polynomial expansion is certainly convenient—although such a structure can be imposed through use of $\mathbf{R}_{xx}$ by specifying a growing variance with r_i.

7.5. Determining a Mean Value. Let the measurements of the physical quantity continue to be denoted y_i and suppose that each is made up of an unknown large scale mean m, plus a deviation from that mean of θ_i. Then,

$$m + \theta_i = y_i\,, \quad 1 \le i \le M \tag{7.24}$$

or

$$\mathbf{D}m + \boldsymbol{\theta} = \mathbf{y}\,, \qquad \mathbf{D}^T = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \end{bmatrix}^T , \tag{7.25}$$

and we seek a best estimate, $\tilde{m}$, of m. In (7.24) or (7.25) the unknown $\mathbf{x}$ has become the scalar m, and the deviation of the field from its mean is the noise, that is, $\boldsymbol{\theta} \equiv \mathbf{n}$, whose true mean is zero. The problem is evidently a special case of the use of basis functions, in which only one function—a zeroth–order polynomial, m, is retained.

Set $\mathbf{R}_{nn} = \langle \boldsymbol{\theta}\boldsymbol{\theta}^T \rangle$. If, for example, we were looking for a large-scale mean temperature in a field of oceanic mesoscale eddies, then $\mathbf{R}_{nn}$ is the sum of the covariance of the eddy field plus that of observational errors and any other fields contributing to the difference between y_i and the true mean m. To be general, suppose $\mathbf{R}_{xx} = \langle m^2 \rangle = m_0^2$ and from (7.19),

$$\begin{aligned}\tilde{m} &= \left\{ \frac{1}{m_0^2} + \mathbf{D}^T \mathbf{R}_{nn}^{-1} \mathbf{D} \right\}^{-1} \mathbf{D}^T \mathbf{R}_{nn}^{-1} \mathbf{y} \\ &= \frac{1}{1/m_0^2 + \mathbf{D}^T \mathbf{R}_{nn}^{-1} \mathbf{D}} \mathbf{D}^T \mathbf{R}_{nn}^{-1} \mathbf{y}\end{aligned} \tag{7.26}$$

($\mathbf{D}^T \mathbf{R}_{nn}^{-1} \mathbf{D}$ is a scalar). The expected uncertainty of this estimate is (7.21),

$$P = \left\{ \frac{1}{m_0^2} + \mathbf{D}^T \mathbf{R}_{nn}^{-1} \mathbf{D} \right\}^{-1} = \frac{1}{1/m_0^2 + \mathbf{D}^T \mathbf{R}_{nn}^{-1} \mathbf{D}} \,, \tag{7.27}$$

(a scalar).[44]

The estimates may appear somewhat unfamiliar; they reduce to more common expressions in certain limits. Let the θ_i be uncorrelated, with uniform variance σ^2; $\mathbf{R}_{nn}$ is then diagonal and (7.26) reduces to

$$\tilde{m} = \frac{1}{(1/m_0^2 + M/\sigma^2)\sigma^2} \sum_{i=1}^{M} y_i = \frac{m_0^2}{\sigma^2 + M m_0^2} \sum_{i=1}^{M} y_i \,, \tag{7.28}$$

where the relations $\mathbf{D}^T \mathbf{D} = M$, $\mathbf{D}^T \mathbf{y} = \sum_{i=1}^{M} y_i$ were used. The expected value of the estimate is

$$\langle \tilde{m} \rangle = \frac{m_0^2}{\sigma^2 + M m_0^2} \sum_{i}^{M} \langle y_i \rangle = \frac{m_0^2}{\sigma^2 + M m_0^2} M m \neq m \,, \tag{7.29}$$

that is, it is biassed , as inferred above, unless $\langle y_i \rangle = 0$, implying $m = 0$. $\mathbf{P}$ becomes,

$$P = \frac{1}{1/m_0^2 + M/\sigma^2} = \frac{\sigma^2 m_0^2}{\sigma^2 + M m_0^2} \,. \tag{7.30}$$

Under the further assumption that $m_0^2 \to \infty$,

$$\tilde{m} = \frac{1}{M} \sum_{i=1}^{M} y_i \,, \tag{7.31}$$

$$P = \sigma^2 / M \,, \tag{7.32}$$

which are the ordinary average and its variance (the latter expression is the well-known "square root of M rule" for the standard deviation of an average; recall Eq. (3.5)); $\langle \tilde{m} \rangle$ in (7.31) is readily seen to be the true mean—this estimate has become unbiassed . But the magnitude of (7.32) always exceeds that of (7.30)—acceptance of bias in the estimate (7.28) reduces the uncertainty of the result.

Eqs. (7.26)–(7.27) are the more general estimation rule—accounting through $\mathbf{R}_{nn}$ for correlations in the observations and their irregular distribution. Because many samples are not

[44] See also Bretherton *et al.* (1976).

independent, (7.32) may be extremely optimistic. Eq. (7.27) gives one the appropriate expression for the variance when the data are correlated (that is, when there are fewer degrees of freedom than the number of sample points).[?Example needed.]

The use of the prior estimate, m_0^2, is interesting. Letting m_0^2 go to infinity does not mean that an infinite mean is expected ((7.31) is finite). It is is merely a statement that there is no information whatever, before we start, as to the magnitude of the true average—it could be arbitrarily large (or small and of either sign) and if it came out that way, would be acceptable. Such a situation is, of course, unlikely and even though we might choose not to use information concerning the probable size of the solution, we should remain aware that we could do so (the importance of the prior estimate diminishes as M grows—so that with an infinite amount of data it has no effect at all on the estimate). If a prior estimate of m itself is available, rather than just its mean square, the problem should be reformulated as one for the estimate of the perturbation about this value.

It is very important not to be tempted into making a first estimate of m_0^2 by using (7.31), substituting into (7.28), thinking to reduce the error variance. For the Gauss-Markov theorem to be valid, the prior information must be truly independent of the data being used.

8. Improving Recursively

A common situation arises that one has a solution $\tilde{\mathbf{x}}, \tilde{\mathbf{n}}, \mathbf{P}$, and more information becomes available, often. in the form of further noisy linear constraints. One way of using the new information is to simply combine the old and new equations into one larger system, and re-solve. This approach may well be the best one. Sometimes however, perhaps because the earlier equations have been discarded, or for reasons of storage or both, one prefers to retain the information in the previous solution without re-solving the entire system. So-called recursive methods, in both least-squares and minimum variance estimation provide the appropriate recipes.

Let the original equations be re-labeled so that we can distinguish them from those that come later, in the form,

$$\mathbf{E}(1)\mathbf{x} + \mathbf{n}(1) = \mathbf{y}(1) \tag{8.1}$$

where the noise $\mathbf{n}(1)$ has zero mean and covariance matrix $\mathbf{R}_{nn}(1)$. Let the estimate of the solution to (8.1) from one of the estimators be written as $\tilde{\mathbf{x}}(1)$, with uncertainty $\mathbf{P}(1)$. As a specific example, suppose (8.1) is full-rank overdetermined, and was solved using row weighted least-squares as,

$$\tilde{\mathbf{x}}(1) = \left[\mathbf{E}(1)^T\mathbf{R}_{nn}(1)^{-1}\mathbf{E}(1)\right]^{-1}\mathbf{E}(1)^T\mathbf{R}_{nn}(1)^{-1}\mathbf{y}(1), \tag{8.2}$$

with corresponding $\mathbf{P}(1)$ (column weighting has no effect in the full-rank fully-determined case).

Some new observations, $\mathbf{y}(2)$, are obtained, with the error covariance of the new observations given by $\mathbf{R}_{nn}(2)$, so that the problem for the unknown $\mathbf{x}$ is

$$\begin{Bmatrix}\mathbf{E}(1)\\ \mathbf{E}(2)\end{Bmatrix}\mathbf{x}+\begin{bmatrix}\mathbf{n}(1)\\ \mathbf{n}(2)\end{bmatrix}=\begin{bmatrix}\mathbf{y}(1)\\ \mathbf{y}(2)\end{bmatrix} \tag{8.3}$$

where $\mathbf{x}$ is the same unknown. We assume $\langle\mathbf{n}(2)\rangle=\mathbf{0}$ and $\langle\mathbf{n}(1)\mathbf{n}(2)^T\rangle=\mathbf{0}$, that is, *no correlation of the old and new measurement errors.* A solution to (8.3) should give a better estimate of $\mathbf{x}$ than (8.1) alone, because more observations are available. It is sensible to row weight the concatenated set to

$$\begin{bmatrix}\mathbf{R}_{nn}(1)^{-T/2}\mathbf{E}(1)\\ \mathbf{R}_{nn}(2)^{-T/2}\mathbf{E}(2)\end{bmatrix}\mathbf{x}+\begin{bmatrix}\mathbf{R}_{nn}(1)^{-T/2}\mathbf{n}(1)\\ \mathbf{R}_{nn}(2)^{-T/2}\mathbf{n}(2)\end{bmatrix}=\begin{bmatrix}\mathbf{R}_{nn}(1)^{-T/2}\mathbf{y}(1)\\ \mathbf{R}_{nn}(2)^{-T/2}\mathbf{y}(2)\end{bmatrix}. \tag{8.4}$$

"Recursive weighted least-squares" seeks the solution to (8.4) without inverting the new, larger matrix, by taking advantage of the existing knowledge of $\mathbf{x}(1)$, $\mathbf{P}(1)$ —however they might actually have been obtained. The objective function corresponding to finding the minimum weighted error norm in (8.4) is,

$$\begin{aligned}J&=(\mathbf{y}(1)-\mathbf{E}(1)\mathbf{x})^T\mathbf{R}_{nn}(1)^{-1}(\mathbf{y}(1)-\mathbf{E}(1)\mathbf{x})\\ &\quad+(\mathbf{y}(2)-\mathbf{E}(2)\mathbf{x})\mathbf{R}_{nn}(2)^{-1}(\mathbf{y}(2)-\mathbf{E}(2)\mathbf{x}).\end{aligned} \tag{8.5}$$

Taking the derivatives with respect to $\mathbf{x}$, the normal equations produce a new solution,

$$\begin{aligned}\tilde{\mathbf{x}}(2)&=\left\{\mathbf{E}(1)\mathbf{R}_{nn}(1)^{-1}\mathbf{E}(1)+\mathbf{E}(2)^T\mathbf{R}_{nn}(2)^{-1}\mathbf{E}(2)\right\}^{-1}\\ &\quad\left\{\mathbf{E}(1)^T\mathbf{R}_{nn}(1)^{-1}\mathbf{y}(1)+\mathbf{E}(2)^T\mathbf{R}_{nn}(2)^{-1}\mathbf{y}(2)\right\}.\end{aligned} \tag{8.6}$$

This is the result from the brute-force re-solution. But one can manipulate (8.6) into[45],

$$\begin{aligned}\tilde{\mathbf{x}}(2)&=\tilde{\mathbf{x}}(1)+\\ &\quad\mathbf{P}(1)\mathbf{E}(2)^T\left[\mathbf{E}(2)\mathbf{P}(1)\mathbf{E}(2)^T+\mathbf{R}_{nn}(2)\right]^{-1}[\mathbf{y}(2)-\mathbf{E}(2)\tilde{\mathbf{x}}(1)]\\ &=\tilde{\mathbf{x}}(1)+\mathbf{K}(2)[\mathbf{y}(2)-\mathbf{E}(2)\tilde{\mathbf{x}}(1)] \qquad (8.7)\\ \mathbf{P}(2)&=\mathbf{P}(1)-\mathbf{K}(2)\mathbf{E}(2)\mathbf{P}(1) \qquad (8.8)\end{aligned}$$

where,

$$\mathbf{K}(2)=\mathbf{P}(1)\mathbf{E}(2)^T\left[\mathbf{E}(2)\mathbf{P}(1)\mathbf{E}(2)^T+\mathbf{R}_{nn}(2)\right]^{-1} \tag{8.9}$$

[45] Brogan (1985).

An alternate form, for $\mathbf{P}(2)$, found from the matrix inversion lemma, is

$$\mathbf{P}(2) = \left[\mathbf{P}(1)^{-1} + \mathbf{E}(2)\,\mathbf{R}_{nn}(2)\,\mathbf{E}(2)\right]^{-1}. \tag{8.10}$$

A similar alternative for $\tilde{\mathbf{x}}(2)$ also exists, involving different dimensions of the matrices to be inverted is also available from the matrix inversion lemma, but is generally less useful. (In some large problems, matrix inversion can prove less onerous than matrix multiplication.)

The solution (8.7) is just the least-squares solution to the full set, but rearranged after a bit of algebra. *The original data*, $\mathbf{y}(1)$, a*nd coefficient matrix,* $\mathbf{E}(1)$, *have disappeared, to be replaced by the first solution* $\tilde{\mathbf{x}}(1)$ *and its uncertainty* $\mathbf{P}(1)$. *That is to say, one need not retain the original data and* $\mathbf{E}(1)$ *for the new solution to be computed.* All information is included in $\tilde{\mathbf{x}}(1), \mathbf{P}(1)$. Furthermore, because the new solution depends only upon $\tilde{\mathbf{x}}(1)$, $\mathbf{P}(1)$, the particular methodology originally employed for obtaining them is irrelevant: they might even have been obtained from an educated guess, or from some long previous calculation of arbitrary complexity. If the initial set of equations (8.1) is actually underdetermined, and should it have been solved using the SVD, one must be careful that $\mathbf{P}(1)$ includes the estimated error owing to the missing nullspace. Otherwise, these elements would be assigned zero error variance, and the new data could never affect them.

The structure of the improved solution (8.7) is also interesting and suggestive. It is made up of two terms: the previous estimate plus a term proportional to the difference between the new observations $\mathbf{y}(2)$, and *a prediction of what those observations should have been* were the first estimate the wholly correct one and the new observations perfect. It thus has the form of a "predictor-corrector." The difference between the prediction and the forecast can be called the "prediction error," but recall there is observational noise in $\mathbf{y}(2)$. The new estimate is a weighted average of this difference and the prior estimate, with the weighting depending upon the details of the uncertainty of prior estimate and new data. The behavior of the updated estimate is worth understanding in various limits. For example, suppose the initial uncertainty estimate is diagonal, $\mathbf{P}(1) = \Delta^2\mathbf{I}$. Then,

$$\mathbf{K}(2) = \mathbf{E}(2)^T\left[\mathbf{E}(2)\mathbf{E}(2)^T + \mathbf{R}_{nn}(2)/\Delta^2\right]^{-1}. \tag{8.11}$$

If the norm of $\mathbf{R}_{nn}(2)/\Delta^2$ is small and if the second set of observations is full rank underdetermined, then,

$$\mathbf{K}(2) \longrightarrow \mathbf{E}(2)^T(\mathbf{E}(2)\mathbf{E}(2)^T)^{-1}$$

and

$$\begin{aligned}\tilde{\mathbf{x}}(2) &= \tilde{\mathbf{x}}(1) + \mathbf{E}(2)^T(\mathbf{E}(2)\mathbf{E}(2)^T)^{-1}[\mathbf{y}(2) - \mathbf{E}(2)\tilde{\mathbf{x}}(1)] \\ &= [\mathbf{I} - \mathbf{E}(2)^T(\mathbf{E}(2)\mathbf{E}(2)^T)^{-1}\mathbf{E}(2)]\tilde{\mathbf{x}}(1) + \mathbf{E}(2)^T(\mathbf{E}(2)\mathbf{E}(2)^T)^{-1}\mathbf{y}(2)\end{aligned} \tag{8.12}$$

Now, $[\mathbf{I}-\mathbf{E}(2)^T(\mathbf{E}(2)\mathbf{E}(2)^T)\mathbf{E}(2)^{-1}] = \mathbf{I}_N - \mathbf{V}\mathbf{V}^T = \mathbf{Q}_v\mathbf{Q}_v^T$, where $\mathbf{V}$ is the full-rank singular vector matrix for $\mathbf{E}(2)$, spans the nullspace (see Eq. 5.101) of $\mathbf{E}(2)$). The update thus replaces, in the the first estimate, all the structures given perfectly by the second set of observations, but retains those structures from the first estimate about which the new observations say nothing—an eminently reasonable result.

At the opposite extreme, when the new observations are very noisy compared to the previous ones, $\left\|\mathbf{R}_{nn}/\Delta^2\right\| \to \infty$, $\|\mathbf{K}(2)\| \to 0$, and the first estimate is left unchanged. The general case represents a weighted average of the previous estimate with elements found from the new data, with the weighting depending both upon the relative noise in each, and upon the structure of the observations relative to the structure of $\mathbf{x}$ as represented in $\mathbf{P}(1)$, $\mathbf{R}_{nn}(2)$, $\mathbf{E}(2)$.

The matrix being inverted in (8.9) is the sum of the measurement error covariance $\mathbf{R}_{nn}(2)$, and the error covariance of the "forecast" $\mathbf{E}(2)\tilde{\mathbf{x}}(1)$. To see this, let $\boldsymbol{\gamma}$ be the error component in $\tilde{\mathbf{x}}(1)$, which by definition has covariance $\langle\boldsymbol{\gamma}\boldsymbol{\gamma}^T\rangle = \mathbf{P}(1)$. Then the expected covariance of the error of prediction is $\langle\mathbf{E}(1)\boldsymbol{\gamma}\boldsymbol{\gamma}^T\mathbf{E}^T\rangle = \mathbf{E}(1)\mathbf{P}(1)\mathbf{E}(1)^T$, which appears in $\mathbf{K}(2)$.

It is useful to notice that Eq. (6.11), the solution to the least-squares problem subject to certain perfect constraints imposed by a Lagrange multiplier, can be recovered from the minimum variance solution (8.7) by putting $\mathbf{E}(2) = \mathbf{A}$, $\mathbf{y}(2) = \mathbf{q}$, $\mathbf{R}_{nn}(2) \to 0$. That is, this earlier solution can be conceived of as having been obtained by first solving the conventional least-squares problem, and then being modified by the *later* information that $\mathbf{Ax} = \mathbf{q}$, with very high accuracy.

The possibility of a recursion based on Eqs. 8.7, 8.8 (or 8.10) is obvious—all subscript 1 variables being replaced by subscript 2 variables, which in turn are replaced by subscript 3 variables, etc. The general form would be,

$$\tilde{\mathbf{x}}(n) = \tilde{\mathbf{x}}(n-1) + \mathbf{K}(n)\left[\mathbf{y}(n) - \mathbf{E}(n)\tilde{\mathbf{x}}(n-1)\right] \tag{8.13}$$

$$\mathbf{K}(n) = \mathbf{P}(n-1)\mathbf{E}(n)^T\left[\mathbf{E}(n)\mathbf{P}(n-1)\mathbf{E}(n)^T + \mathbf{R}_{nn}(n)\right]^{-1} \tag{8.14}$$

$$\mathbf{P}(n) = \mathbf{P}(n-1) - \mathbf{K}(n)\mathbf{E}(n)\mathbf{P}(n-1) \tag{8.15}$$

The alternative form for Eq. (8.15) is, from (8.10),

$$\mathbf{P}(n) = \left[\mathbf{P}(n-1)^{-1} + \mathbf{E}(n)\mathbf{R}_{nn}(n)^{-1}\mathbf{E}(n)\right]^{-1}. \tag{8.16}$$

The computational load of the recursive solution needs to be addressed. A least-squares solution does *not* require one to calculate the uncertainty $\mathbf{P}$ (although the utility of $\tilde{\mathbf{x}}$ without such an estimate is unclear). But to use the recursive form, one must have $\mathbf{P}(n-1)$, otherwise the update step, Eq. (8.13) cannot be used. In very large problems, such as appear in fluid flows, the computation of the uncertainty, from 8.15, or 8.16 can become extremely burdensome, if not prohibitive. In such a situation, one might simply store all the data, and do one large, single calculation—if this is feasible. Normally, it will involve less pure computation than will the recursive solution which must repeatedly update $\mathbf{P}\left(n\right)$.

The comparatively simple interpretation of the recursive, weighted least-squares problem will be used in Chapter 4 to derive the Kalman filter and suboptimal filters in a very simple form. It also becomes the key to understanding "assimilation" schemes such as "nudging" and "forcing to climatology," and "robust diagnostic" methods.

Let us confirm directly that the recursive least-squares result is identical to a recursive estimation procedure, if appropriate least-squares weight matrices were used. Suppose there exist two *independent* estimates of an unknown vector $\mathbf{x}$, denoted $\tilde{\mathbf{x}}_a$, $\tilde{\mathbf{x}}_b$ with estimated uncertainties $\mathbf{P}_a$, $\mathbf{P}_b$, respectively. They are either unbiassed , or have the same bias, that is, $\langle\tilde{\mathbf{x}}_a\rangle = \langle\tilde{\mathbf{x}}_b\rangle = \mathbf{x}_B$. How should the two be combined to give a third estimate $\tilde{\mathbf{x}}^+$ with minimum error? Try a linear combination,

$$\tilde{\mathbf{x}}^+ = \mathbf{L}_a\tilde{\mathbf{x}}_a + \mathbf{L}_b\tilde{\mathbf{x}}_b\,. \tag{8.17}$$

If the new estimate is to be unbiassed, or is to retain the prior bias, it follows that,

$$\langle\tilde{\mathbf{x}}^+\rangle = \mathbf{L}_a\langle\tilde{\mathbf{x}}_a\rangle + \mathbf{L}_b\langle\tilde{\mathbf{x}}_b\rangle \tag{8.18}$$

or,

$$\mathbf{x}_B\!= \mathbf{L}_a\mathbf{x}_B\!+\!\mathbf{L}_b\mathbf{x}_B \tag{8.19}$$

or,

$$\mathbf{L}_b\!= \mathbf{I}-\mathbf{L}_a \tag{8.20}$$

Then the uncertainty is,

$$\begin{aligned}\mathbf{P}^+ &= \langle(\tilde{\mathbf{x}}^+ - \mathbf{x})(\tilde{\mathbf{x}}^+ - \mathbf{x})^T\rangle = \langle(\mathbf{L}_a\tilde{\mathbf{x}}_a + (\mathbf{I}-\mathbf{L}_a)\tilde{\mathbf{x}}_b)(\mathbf{L}_a\tilde{\mathbf{x}}_a + (\mathbf{I}-\mathbf{L}_a)\tilde{\mathbf{x}}_b)^T\rangle \\ &= \mathbf{L}_a\mathbf{P}_a\mathbf{L}_a^T + (\mathbf{I}-\mathbf{L}_a)\mathbf{P}_b(\mathbf{I}-\mathbf{L}_a)^T\end{aligned} \tag{8.21}$$

where the independence assumption has been used to set $\langle(\tilde{\mathbf{x}}_a\!-\!\mathbf{x})\,(\tilde{\mathbf{x}}_b\!-\!\mathbf{x})\rangle = 0$. $\mathbf{P}^+$ is positive definite; minimizing its diagonal elements with respect to $\mathbf{L}_a$ yields (after writing out the diagonal elements of the products, just as in the Gauss-Markov theorem derivation),

$$\mathbf{L}_a = \mathbf{P}_b(\mathbf{P}_a + \mathbf{P}_b)^{-1}, \qquad \mathbf{L}_b = \mathbf{P}_a(\mathbf{P}_a + \mathbf{P}_b)^{-1}\,.$$

The new combined estimate is then,

$$\tilde{\mathbf{x}}^+ = \mathbf{P}_b(\mathbf{P}_a + \mathbf{P}_b)^{-1}\tilde{\mathbf{x}}_a + \mathbf{P}_a(\mathbf{P}_a + \mathbf{P}_b)^{-1}\tilde{\mathbf{x}}_b\,.$$

This last expression can be rewritten by adding and subtracting $\tilde{\mathbf{x}}_a$ as,

$$\begin{aligned}\tilde{\mathbf{x}}^{+} &= \tilde{\mathbf{x}}_a+\mathbf{P}_b(\mathbf{P}_a+\mathbf{P}_b)^{-1}\tilde{\mathbf{x}}_a\\ &\quad+\mathbf{P}_a(\mathbf{P}_a+\mathbf{P}_b)^{-1}\tilde{\mathbf{x}}_b-(\mathbf{P}_a+\mathbf{P}_b)(\mathbf{P}_a+\mathbf{P}_b)^{-1}\tilde{\mathbf{x}}_a\\ &= \tilde{\mathbf{x}}_a+\mathbf{P}_a(\mathbf{P}_a+\mathbf{P}_b)^{-1}(\tilde{\mathbf{x}}_b-\tilde{\mathbf{x}}_a)\,. \end{aligned} \tag{8.22}$$

Notice in particular, the re-appearance of a predictor-corrector form relative to $\tilde{\mathbf{x}}_a$.

The uncertainty of the estimate (8.22) is easily evaluated as

$$\mathbf{P}^{+}=\mathbf{P}_a-\mathbf{P}_a(\mathbf{P}_a+\mathbf{P}_b)^{-1}\mathbf{P}_a\,. \tag{8.23}$$

and which, by straightforward application of the matrix inversion lemma, is,

$$\mathbf{P}^{+}=(\mathbf{P}_a^{-1}+\mathbf{P}_b^{-1})^{-1}. \tag{8.24}$$

As is always the case with linear estimation methods, the uncertainty is independent of the observations. Eqs. (8.22-8.24) are the general rules for combining two estimates with uncorrelated errors.

Now suppose that $\tilde{\mathbf{x}}_a$ and its uncertainty are known, but that instead of $\tilde{\mathbf{x}}_b$ there are measurements,

$$\mathbf{E}(2)\mathbf{x}+\mathbf{n}(2)=\mathbf{y}(2), \tag{8.25}$$

with $\langle\mathbf{n}(2)\rangle=0$, $\langle\mathbf{n}(2)\mathbf{n}(2)^T\rangle=\mathbf{R}_{nn}(2)$. From this second set of observations, we *estimate* the solution, using the minimum variance estimator (7.19, 7.21) with no use of the solution variance; that is, let $\|\mathbf{R}_{xx}^{-1}\|\to 0$. The reason for suppressing $\mathbf{R}_{xx}$, which logically could come from $\mathbf{P}_a$, is to maintain the independence of the previous and the new estimates. Then,

$$\tilde{\mathbf{x}}_b=(\mathbf{E}^T\mathbf{R}_{nn}\,(2)^{-1}\,\mathbf{E})^{-1}\mathbf{E}^T\mathbf{R}_{nn}(2)^{-1}\mathbf{y} \tag{8.26}$$

$$\mathbf{P}_b=(\mathbf{E}^T\mathbf{R}_{nn}\,(2)^{-1}\,\mathbf{E})^{-1}\,. \tag{8.27}$$

Substituting (8.26), (8.27) into (8.22), (8.23), and using the matrix inversion lemma gives

$$\tilde{\mathbf{x}}^{+}=\tilde{\mathbf{x}}_a+\mathbf{P}_a\mathbf{E}(2)[\mathbf{E}(2)\mathbf{P}_a\mathbf{E}(2)^T+\mathbf{R}_{nn}(2)]^{-1}(\mathbf{y}-\mathbf{E}(2)\tilde{\mathbf{x}}_a)\,, \tag{8.28}$$

which is the same as (8.13), and thus *a recursive minimum variance estimate coincides with a corresponding weighted least-squares recursion.* The new covariance may also be confirmed to be (8.15 or 8.16). Notice that if $\tilde{\mathbf{x}}_a$ was itself estimated from an earlier set of observations, that those data have disappeared from the problem, with all the information derived from them contained in $\tilde{\mathbf{x}}_a$ and $\mathbf{P}_a$. Thus, again, earlier data can be wholly discarded after use.

9. A Recapitulation

This chapter has not exhausted the possibilities for inverse methods, and the techniques will be extended in several directions in the next Chapters. Given the lengthy nature of the discussion so far, some summary of what has been accomplished may be helpful.

The focus is on making inferences about parameters or fields, $\mathbf{x}$, $\mathbf{n}$ satisfying linear relationships of the form

$$\mathbf{E}\mathbf{x} + \mathbf{n} = \mathbf{y}\,.$$

Such equations arise as we have seen, from both "forward" and "inverse" problems, but the techniques for estimating $\mathbf{x}$, $\mathbf{n}$ and their uncertainty are useful whatever the physical origin of the equations. Two methods for estimating $\mathbf{x}$, $\mathbf{n}$ have been the focus of the chapter: least-squares (including the singular value decomposition) and the Gauss-Markov or minimum variance technique. Least-squares, in any of its many guises, is a very powerful method—but its power and ease of use have (judging from the published literature) led many investigators into serious confusion about what they are doing—invoking the language of minimum variance estimation while solving their problem with least-squares.

An attempt is made therefore, to emphasize the two distinct roles of least-squares: as a method of *approximation*, and as a method of *estimation*. It is only in the second formulation that it can be regarded as an inverse method. A working definition of an inverse method is a technique able to estimate unknown parameters or fields of a model, while producing an estimate of the uncertainties of the results. Solution plus uncertainty are the fundamental requirements. There are many desirable additional features of inverse methods which can prove extremely important. Among these features are: (1) Separation of nullspace uncertainties from observational noise uncertainties, (2) the ability to rank the data in its importance to the solution, (3) the ability to use prior statistical knowledge, (4) understanding of solution structures in terms of data structure, (5) the ability to trade resolution against variance. (The list is not exhaustive. For example, in Chapter 3 we will briefly examine the use of inequality information.) As with all estimation methods, one also trades computational load against the need for information (the SVD is a powerful form of least-squares, but requires more computation that do other forms). The Gauss-Markov approach has the strength of forcing explicit use of prior statistical information and is directed at the central goal of obtaining $\mathbf{x}$, $\mathbf{n}$ with smallest mean-square error, and for this reason might well be regarded as the default methodology for linear inverse problems. It has the added advantage that we know we can obtain precisely the same result with appropriate versions of least-squares, including the SVD, permitting the use of least-squares algorithms, but at the risk of losing sight of what we are actually attempting.

A number of different procedures for producing estimates of the solution to a set of noisy simultaneous equations of arbitrary dimension have been described here. The reader may wonder which of the variants makes the most sense to use in practice. Because, in the presence of noise one is dealing with a statistical estimation problem, there is no single "best" answer, and one must be guided by model context and goals. A few general remarks might be helpful.

In any problem where data are to be used to make inferences about physical parameters, one typically needs some approximate idea of just how large the solution is likely to be and how large the residuals probably are. In this nearly agnostic case, where almost nothing else is known, and the problem is very large, the weighted, tapered least-squares solution is a good first choice—it is easily and efficiently computed and coincides with the Gauss-Markov and tapered SVD solutions, if the weight matricesare the appropriate covariances. Sparse matrix methods for its solution exist should that be necessary[46]. Coincidence with the Gauss-Markov solution means one can reinterpret it as a minimum-variance or maximum-likelihood solution (the Chapter Appendix) should one wish.

It is a comparatively easy matter to vary the trade-off parameter, α^2, to explore the consequences of any errors in specifying the noise and solution variances. Once a value for α^2 is known, the tapered SVD can then be computed to understand the relationships between solution and data structures, their resolution and their variance. For problems of small to moderate size (the meaning of "moderate" is constantly shifting, but it is difficult to examine and interpret matrices of more than about 500×500), the SVD, whether in the truncated or tapered forms is probably the method of choice—because it provides the fullest information about data and its relationship to the solution. Its only disadvantages are that one can easily be overwhelmed by the available information, particularly if a range of solutions must be examined, and it cannot take advantage of sparsity in large problems. The SVD has a flexibility beyond even what we have discussed—one could, for example, change the degree of tapering in each of the terms of (5.149)–(5.150) should there be reason to repartition the variance between solution and noise, or some terms could be dropped out of the truncated form at will—should the investigator know enough to justify it.

To the extent that either or both of $\mathbf{x}, \mathbf{n}$ have expected structures expressible through covariance matrices, these structures can be removed from the problem through the various weight matrices and/or the Cholesky decomposition. The resulting problem is then one in completely unstructured (equivalent to white noise) elements $\mathbf{x}, \mathbf{n}$. In the resulting scaled and rotated systems, one can use the simplest of all objective functions. Covariance, resolution etc., in the original spaces of $\mathbf{x}, \mathbf{n}$ is readily recovered by appropriately applying the weight matrices to the results of the scaled/rotated space.

Both ordinary weighted least-squares and the SVD applied to row- and column-weighted equations are best thought of as approximation, rather than estimation, methods and thus have

[46] Paige & Saunders (1982)

a lot to recommend them. In particular, the truncated SVD does not produce a minimum variance estimate the way the tapered version can. The tapered SVD (along with the Gauss-Markov estimate, or the tapered least-squares solutions) produce the minimum variance property by tolerating a bias in the solution. Whether the bias is more desirable than a larger uncertainty is a decision the user must make. But the reader is warned against the belief that there is any single best method whose determination should take precedence over understanding the problem physics.

10. Appendix to Chapter. Maximum Likelihood

The estimation procedure used in this book is primarily based upon the idea of minimizing the variance of the estimate about the true value. Alternatives exist. Given a set of observations with known joint probability density, one can base of method for estimating various sample parameters upon a principle of "maximum likelihood." This very general and powerful principle attempts to find those estimated parameters which render the actual observations the most likely to have occurred. By way of motivation, consider the simple case of uncorrelated jointly normal stationary time series, x_i, where,

$$\langle x_i \rangle = m, \; \langle (x_i - m)(x_j - m) \rangle = \sigma^2 \delta_{ij}.$$

The corresponding joint probability density for $\mathbf{x} = [x_1, x_2, ..., x_N]$ can be written,

$$\begin{aligned} p_{\mathbf{x}}(\mathbf{X}) &= \frac{1}{(2\pi)^{N/2}\sigma^N} \times \\ &\exp\left\{ -\frac{1}{2\sigma^2} \left[(X_1 - m)^2 + (X_2 - m)^2 + ... + (X_N - m)^2 \right] \right\} \end{aligned} \tag{10.1}$$

Substitution of the observed values, $X_1 = x_1, X_2 = x_2, ...$ into Eq. (10.1) permits evaluation of the probability that these particular values occurred. Denote the corresponding probability density as L. One can demand those values of m, σ rendering the value of L as large as possible. L will be a maximum if $\log(L)$ is as large as possible: that is we seek to maximize,

$$\log(L) = -\frac{1}{2\sigma^2}\left[(x_1 - m)^2 + (x_2 - m)^2 + ... + (x_N - m)^2 \right] + N \log(\sigma) + \frac{N}{2} \log(2\pi)$$

with respect to m, σ. Setting the corresponding partial derivatives to zero and solving produces,

$$\tilde{m} = \frac{1}{N} \sum_{i=1}^{M} x_i, \; \tilde{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \tilde{m})^2.$$

That is, the usual sample mean, and biassed sample variance maximize the probability of the observed data actually occurring. A similar calculation is readily carried out using correlated normal, or any random variables with a different probability density.

Likelihood estimation, and its close cousin, Bayesian methods, are general powerful estimation methods which can be used as an alternative to almost everything covered in this book.[47] Some will prefer that route, but the methods used here are more intuitive and adequate for a very wide range of problems.

11. Appendix to Chapter. Relationship to Differential Operators and Green Functions

Adjoints appear prominently in the theory of differential operators and are usually defined independent of any optimization problem. Many of the concepts are those used in defining Green functions.

Consider an example. Suppose we want to solve an ordinary differential equation,

$$\frac{du\left(\xi\right)}{d\xi}+\frac{d^{2}u\left(\xi\right)}{d\xi^{2}}=\rho\left(\xi\right), \tag{11.1}$$

subject to boundary conditions on $u\left(\xi\right)$ at $\xi=0,L$. To proceed, seek first a solution to

$$\alpha\frac{\partial v\left(\xi,\xi_{0}\right)}{\partial\xi}+\frac{\partial^{2}v\left(\xi,\xi_{0}\right)}{\partial\xi^{2}}=\delta\left(\xi_{0}-\xi\right) \tag{11.2}$$

where α is arbitrary for the time being. Multiply (11.1) by v, and (11.2) by u and subtract:

$$\begin{aligned} & v\left(\xi,\xi_{0}\right)\frac{du\left(\xi\right)}{d\xi}+v\left(\xi,\xi_{0}\right)\frac{d^{2}u\left(\xi\right)}{d\xi^{2}}-u\left(\xi\right)\alpha\frac{\partial v\left(\xi,\xi_{0}\right)}{\partial\xi}-u\left(\xi\right)\frac{\partial^{2}v\left(\xi,\xi_{0}\right)}{\partial\xi^{2}} \\ = \; & v\left(\xi,\xi_{0}\right)\rho\left(\xi\right)-u\left(\xi\right)v\left(\xi,\xi_{0}\right). \end{aligned} \tag{11.3}$$

Integrate this last equation over the domain,

$$\begin{aligned} & \int_{0}^{L}\left\{v\left(\xi,\xi_{0}\right)\frac{du\left(\xi\right)}{d\xi}+\right. \\ & \left. v\left(\xi,\xi_{0}\right)\frac{d^{2}u\left(\xi\right)}{d\xi^{2}}-u\left(\xi\right)\alpha\frac{\partial v\left(\xi,\xi_{0}\right)}{\partial\xi}-u\left(\xi\right)\frac{\partial^{2}v\left(\xi,\xi_{0}\right)}{\partial\xi^{2}}\right\}d\xi \qquad (11.4) \\ & =\int_{0}^{L}\left\{v\left(\xi,\xi_{0}\right)\rho\left(\xi\right)-u\left(\xi\right)\delta\left(\xi_{0}-\xi\right)\right\}d\xi, \qquad (11.5) \end{aligned}$$

or,

$$\begin{aligned} & \int_{0}^{L}\frac{d}{d\xi}\left\{v\frac{du}{d\xi}-\alpha u\frac{dv}{d\xi}\right\}d\xi+\int_{0}^{L}\left\{u\frac{d^{2}v\left(\xi,\xi_{0}\right)}{d\xi^{2}}-u\frac{d^{2}v\left(\xi,\xi_{0}\right)}{d\xi^{2}}\right\}d\xi \\ = \; & \int_{0}^{L}v\left(\xi,\xi_{0}\right)\rho\left(\xi\right)d\xi-u\left(\xi_{0}\right) \end{aligned} \tag{11.6}$$

Choose $\alpha=-1$; then the first term on the left hand-side is integrable, as,

$$\int_{0}^{L}\frac{d}{d\xi}\left\{uv\right\}d\xi=\left.uv\right|_{0}^{L}, \tag{11.7}$$

[47] See especially, van Trees (1968).

as is the second term on the left,

$$\int_0^L \frac{d}{d\xi}\left\{u\frac{dv}{d\xi}-v\frac{du}{d\xi}\right\}d\xi=\left[u\frac{dv}{d\xi}-v\frac{du}{d\xi}\right]_0^L \tag{11.8}$$

and thus,

$$u(\xi_0)=\int_0^L v(\xi,\xi_0)\,\rho(\xi)\,d\xi+uv|_0^L+\left[u\frac{dv}{d\xi}-v\frac{du}{d\xi}\right]_0^L \tag{11.9}$$

Because the boundary conditions on v were not specified, we are free to choose them such that $v=0$ on $\xi=0,L$ such that e.g., the boundary terms reduce simply to $[udv/d\xi]_0^L$, which is then known.

Here, v is the adjoint solution to Eq. (11.2) with $\alpha=-1$, defining the adjoint equation to (11.1); it was found by requiring that the terms on the left-hand-side of Eq. (11.6) should be exactly integrable. v is also the problem Green function (although the Green function is sometimes defined so as to satisfy the forward operator, rather than the adjoint one). Textbooks show that for a general differential operator, $\mathcal{L}$, the requirement that v should render the analogous terms integrable is that,

$$u^T\mathcal{L}v=v^T\mathcal{L}^Tu \tag{11.10}$$

where here the superscript T denotes the adjoint, Eq. (7.19) defines the adjoint operator (compare to (6.26a)).

12. Exercises

Exercise 3. *Using an eigenvector/eigenvalue analysis, solve (a)*

$$\left\{\begin{matrix}1&1&-2\\1&2&-1\\-2&-1&6\end{matrix}\right\}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}=\begin{bmatrix}1\\0\\1\end{bmatrix}, \tag{12.1}$$

and (b)

$$\left\{\begin{matrix}1&1&-2\\1&2&-1\\1.5&2&-2.5\end{matrix}\right\}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}=\begin{bmatrix}1\\0\\1\end{bmatrix} \tag{12.2}$$

Exercise 4. *(a) Find the ranges and null spaces of*

$$\mathbf{A}=\left\{\begin{matrix}2&-1&1\\3&2&1\end{matrix}\right\} \tag{12.3}$$

and calculate the solution and data resolution matrices. (b) Let there be a set of observations $\mathbf{y}$, *such that*

$$\mathbf{Ax}+\mathbf{n}=\begin{bmatrix}1\\2\end{bmatrix} \tag{12.4}$$

This problem is clearly formally undetermined. Find the solution which minimizes

$$J = \mathbf{x}^T\mathbf{x} \tag{12.5}$$

and compare it to the SVD solution with null space set to zero. What is the uncertainty of this solution? (c) Now consider instead

$$\mathbf{A} = \left\{ \begin{array}{cc} 2 & 3 \\ -1 & 2 \\ 1 & 1 \end{array} \right\}, \tag{12.6}$$

and the formally overdetermined problem

$$\mathbf{Ax} + \mathbf{n} = \left[\begin{array}{c} 1 \\ 2 \\ -1 \end{array} \right] \tag{12.7}$$

and find the least-squares solution which minimizes $\mathbf{n}^T\mathbf{n}$. *What is the uncertainty of this solution? How does the solution compare to the SVD solution? (d) For an arbitrary* $\mathbf{A}$, *solve the least-squares problem of minimizng*

$$J = \mathbf{x}^T\mathbf{x} + \boldsymbol{\alpha}^{-2}\mathbf{n}^T\mathbf{n} \tag{12.8}$$

and re-write the solution in terms of its SVD. Discuss what happens to the small singular value contributions.

Exercise 5. *There is one observation*

$$x + n_1 = 1 \tag{12.9}$$

and a priori statistics $< n >=< x >= 0, < n^2 >= 1/2, < x^2 >= 1/2$. *(a)What is the best estimate of* x, n*? (b) A second measurement becomes available,*

$$x + n_2 = 3 \tag{12.10}$$

with $< n_2 >= 0, < n_2^2 >= 4$. *What is the new best estimate of* x *and what is its estimated uncertainty. Are the various a priori statistics consistent with the final result?*

Exercise 6. *Two observations of unknown* x *produce the apparent results*

$$x = 1 \tag{12.11}$$

$$x = 3 \tag{12.12}$$

Produce a reasonable value for x *under the assumption that (a) both observations are equally reliable, and (b) that the second observation is much more reliable (but not infinitely so) than the first (make some reasonable numerical assumption about what "reliable" means and state what you are doing). Can you re-write eqs. (12.11,12.12) in a more sensible form?*

EXERCISE 7. *Two observations of 3 unknowns, x, y, z produce the apparent result,*

$$
\begin{aligned}
x - y - z &= 1 && (12.13)\\
x - y - z &= 3 && (12.14)
\end{aligned}
$$

Discuss what if, anything, might be inferred from such a peculiar result. You can make some sensible assumptions about what is going on, but say what they are.

EXERCISE 8. *The temperature along an oceanic transect is believed to satisfy a linear rule, $\theta = ar + b$, where r is the distance from a reference point, and a, b are constants. Measurements of θ at sea, called y, produce the following values, $r = 0, y = 10; r = 1, y = 9.5; r = 2, y = 11.1, r = 3, y = 12$. (a) Using ordinary least-squares, find an estimate of a, b and the noise in each measurement, and their standard errors. (b) Solve it again using the SVD and discuss, via the resolution matrices, which of the observations, if any proved most important. Is the solution fully resolved?*

EXERCISE 9. *Consider the system of equations*

$$\left\{\begin{array}{ccc} 1 & 2 & 1 \\ 1 & 2.1 & 1 \end{array}\right\} \mathbf{x} + \mathbf{n} = \left[\begin{array}{c} 1 \\ 2 \end{array}\right]. \qquad (12.15)$$

Using the SVD, compare the solutions at ranks $1, 2$ for the two cases of

$$\mathbf{R} = \mathbf{I}_2, \mathbf{R} = \left\{\begin{array}{cc} 1 & 0.99999 \\ 0.99999 & 1 \end{array}\right\}. \qquad (12.16)$$

How do the rank 1 solutions differ in their treatment of the noise? What is the difference in the solutions at rank 2?

EXERCISE 10. *Figure 14*

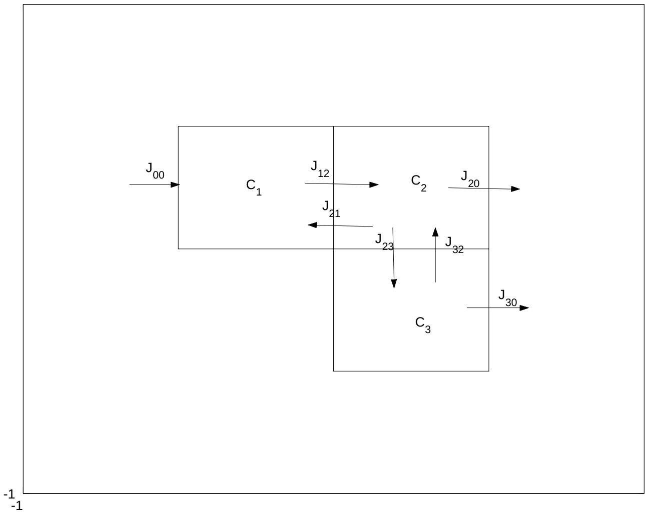


FIGURE 14. Special case of a tracer flux problem with only three reservoirs (plus an external one). J_{ij} represent flows from box i to box j and the C_i are tracer concentrations.

depicts a simple "box-model". There are concentrations C_i in each of three boxes and the mass flux from box i to box j is $J_{ij} > 0$. Box "0" corresponds to externally imposed conditions. (a) Write the simultaneous equations for mass conservation in each box. (b) Let the concentration source or sink in box i be denoted q_i. Write the simultaneous equations for concentration steady-state in each box. (c) Initially all J_{ij} are thought to be about 8 (this is a not very sophisticated way of dealing with the positivity constraint on J_{ij}) and measurements show $C_0 = 5, C_1 = 3, C_3 = 1, q_1 = 20 \pm 2, q_2 = -2 \pm 2, q_3 = 8 \pm 10$. Assuming the measurements of C_i are perfect, make a better estimate of J_{ij},by finding the various corrections ΔJ_{ij}. (d) Assuming $< \Delta J_{ij} >= 0, < \Delta J_{ij}^2 >= 10$, find a solution using the truncated and tapered SVD and the Gauss-Markov Theorem. Find the uncertainty of the estimates. (e) Solve the problem by linear programming without using the a priori *variances, but enforcing the positivity constraints on the J_{ij}.*

EXERCISE 11. *For the Laplace-Poisson equation $\nabla^2\phi = \rho$ with Dirichlet boundary conditions in a square domain, put it into discrete form and code it on a computer so that it can be written,*

$$\mathbf{Ax} = \mathbf{b}. \tag{12.17}$$

Choose any reasonable dimension for the number of grid points or finite elements or basis functions. Confirm that $\mathbf{A}$ is square. (a) For any reasonable boundary conditions ϕ_b and values of ρ, solve (12.17) as a forward problem (b) Add some random noise to ϕ_b and solve it again. (c) Omit any knowledge of ρ over some part of the domain and find at least one possible solution (you could use least-squares). (d) Omit any knowledge of ϕ_b over some part of the domain and *find at least one possible solution. (e) Suppose ϕ from (a) is known over part of the domain, use that knowledge to help improve the solutions in (b-d).*

EXERCISE 12. *Consider the simultaneous equations, $\mathbf{Ax} = \mathbf{y}$,*

$$\left\{\begin{array}{lll} 1 & 1 & -1 \\ 2 & 1 & 1 \\ -1 & 0 & -2 \end{array}\right\} \left[\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array}\right] = \left[\begin{array}{c} 1 \\ 2 \\ 2 \end{array}\right]$$

(a) Using a numerical routine for symmetric matrix eigenvalue/eigenvector problems (i.e., do not *use a singular value decomposition program such as MATLAB's SVD), find the singular value decomposition for the matrix $\mathbf{A}$. (b) Find the null space and ranges of $\mathbf{A}, \mathbf{A}^T$ (c) Using the singular vectors and singular values, find the general solution to the equations and explain the behavior of this solution. Are there any residuals? (d) Find the resolution matrix for the solution and for the "data", $\mathbf{y}$.*

EXERCISE 13. *You have five data points, $y_t = 1, -2, -3, -2, -1$,$t = 0, 1, 3, 4, 5$ and you have reason to believe they are given by a reduced Fourier series*

$$y_t = a\cos(2\pi t/6) + b\sin(2\pi t/3) + n_t \tag{12.18}$$

where n_t is noise. Solve this problem for estimates of $a_,$ b, n_t in three ways (a) As an ordinary least-squares problem. (You can use a matrix inversion routine if you wish.). (b) As an underdetermined problem in 7 unknowns. (c) By the singular value decomposition (you may use an svd routine if you want). Explain the differences among the solutions. (d) The noise variance is believed to be $< n_t^2 >= 1.5$. Make an estimate of the uncertainty in your estimates of a, b.

EXERCISE 14. *(a) Set up the Neumann problem on a grid and show explicitly that there is a solution and "observation" null space. Interpret them. (b) Let the normal boundary condition be $\partial\phi/\partial n = 3$ everywhere. Is there any difficulty? What is its character, and how might it be dealt with?*

EXERCISE 15. *For the Neumann problem, write the model equations with error terms, and solve the problem with additional information providing estimates of ϕ_{ij} at several grid points (rendering the problem formally overdetermined).*

CHAPTER 4

The Time-Dependent Inverse Problem: State Estimation

1. Background

The discussion so far has treated models and data that most naturally represent a static world. Much data however, represent systems that are changing to some degree. Many familiar differential equations describe phenomena that are intrinsically time-dependent; a good example is the ordinary wave equation,

$$\frac{1}{c^2}\frac{\partial^2 x\,(t,r)}{\partial t^2}-\frac{\partial^2 x(t,r)}{\partial r^2}=0. \tag{1.1}$$

One may well wonder if the methods described in Chapter 2 have any use with data thought to be described by (1.1). One way to answer the question is to recognize that t is simply another coordinate, and can be regarded e.g., as the counterpart of one of the space coordinates encountered in the previous discussion of two dimensional partial differential equations. From this point of view, time dependent systems are nothing but versions of the systems already developed. (The statement is even more obvious for the simpler equation,

$$\frac{d^2 x\,(t)}{dt^2}=q\,(t)\,. \tag{1.2}$$

That the coordinate is labelled t is a detail.)

On the other hand, time often has a somewhat different flavor to it than does a spatial coordinate because it has an associated direction. The most obvious example occurs when one has data up to and including some particular time t, and one asks for a *forecast* of some elements of the system at some future time $t' > t$. Even this role of time is not unique: one could imagine a completely equivalent spatial forecast problem, in which e.g., one required extrapolation of the map of an ore body beyond some area in which measurements exist. In state estimation, time does not introduce truly novel problems. The main issue is really a computational one: problems in two or more spatial dimensions, when time dependent, typically generate system dimensions which are simply too large for conventionally available computer systems. Simply on the basis of computational load, one seeks state estimation algorithms that are more efficient in some way, than what can be achieved with the methods used so far. Consider as an example,

$$\frac{\partial\phi}{\partial t}=\kappa\nabla^2\phi, \tag{1.3}$$

a two dimensional generalization of the Laplace equation (a diffusion equation). Using a one-sided time difference, and the discrete form of the Laplacian in Eq. (2.2, Chapter 1), one has

$$\begin{aligned}&\frac{\phi_{ij}\left(\left(n+1\right)\Delta t\right)-\phi_{ij}\left(n\Delta t\right)}{\Delta t}= \\ &\kappa\left\{\phi_{i+1,j}\left(n\Delta t\right)-2\phi_{i,j}\left(n\Delta t\right)+\phi_{i-1,j}\left(n\Delta t\right)+\phi_{i,j+1}\left(n\Delta t\right)-2\phi_{i,j}\left(n\Delta t\right)+\phi_{i,j-1}\left(n\Delta t\right)\right\}\end{aligned} \tag{1.4}$$

If there are M^2 elements defining ϕ_{ij} at each time $n\Delta t$, then the number of elements over the entire time span of T time steps, would be TM^2 and which grows rapidly as the number of time steps increases. Typically the relevant observation numbers also grows rapidly through time. On the other hand, the operation,

$$\mathbf{x}=\mathrm{vec}\left(\phi_{ij}\left(n\Delta t\right)\right), \tag{1.5}$$

renders Eq. (1.4) in the familiar form

$$\mathbf{A}_1\mathbf{x}=\mathbf{0}, \tag{1.6}$$

and with some boundary conditions, some initial conditions and/or observations, and a big enough computer, one could proceed with any of the methods in Chapter 2 completely unchanged. There are however, many times when $T,\ M^2$ become so large, that even the biggest extant computer is inadquate. One then seeks methods, which can take advantage of special structures built into time evolving equations to reduce the computational load. (Note however, that $\mathbf{A}_1$ is sparse.)

This chapter is in no sense exhaustive; many entire books are devoted to the material and its extensions, which are important for understanding and practical use. The intention is to lay out the fundamental ideas, which are generalizations of methods already described in Chapters 2 and with the hope that they will permit the reader to penetrate the wider literature. Several very useful textbooks are available for readers who are not deterred by discussions in contexts differing from their own applications.[1] Most of the methods now being used in fields like oceanography and meteorology have been known for years under the general heading of control theory and control engineering. The experience in these latter fields is very helpful; the main issues in applications to fluid problems concern the size of the models and data encountered: they are typically many orders of magnitude larger than anything contemplated by engineers, and one requires novel computational approaches. In meteorology, estimation for forecasting is called *assimilation*.[2]

There are several notation systems in wide use. I have chosen here to use one that appears to be both simple and adequate to our needs, and it is taken directly from the control theory

[1] I have found those by Liebelt (1967), Gelb (1974), Bryson and Ho (1975), Brown (1983), and Anderson and Moore (1979) to be especially helpful

[2] Daley (1991)

literature. Meteorologists have tended to go their own idiosyncratic way[3], with some loss of transparency and in the ability to easily import ideas from other fields.

2. Some Basic Ideas and Notation

2.1. Models. In the context of this chapter, by "models" is meant statements about the connections between the system variables in some place at some time, and those in all other places and times. Maxwell's equations are a model of the behavior of time-dependent electromagnetic disturbances. These equations can be used to connect the magnetic and electric fields everywhere in space and time. Other physical systems are described by the Schrodinger, elastic wave, or fluid-dynamical equations. Static situations are simply special limits, e.g., for an electrostatic field in a container with known boundary conditions.

A useful concept is that of the system "state". By that is meant the information at a single moment in time required to fully describe the system a small time step into the future. So for example, the time evolution of a system described by the tracer diffusion equation,

$$\frac{\partial C\left(\mathbf{r},t\right)}{\partial t}-\kappa\nabla^2 C\left(\mathbf{r},t\right)=0,$$

inside a closed container can be calculated with arbitrary accuracy at time $t+\Delta t$, if one knows $C\left(\mathbf{r},t\right)$, and the boundary conditions $C_B\left(t\right)$,as $\Delta t\rightarrow 0$. $C\left(\mathbf{r},t\right)$ is the state variable, with the boundary conditions being regarded as separate externally provided variables (but the distinction is, as we will see, to some degree an arbitrary one) In practice, such quantities as initial and boundary conditions, container shape, etc. are always imperfectly known, typically arise from observation, and the problems are identical to those already considered.

Consider any model, whether time dependent or steady, but rendered in discrete time. The "state vector" $\mathbf{x}\left(t\right)$ is defined as those elements of the model employed to describe fully the physical state of the system at any time and all places as required by the model in use. For the discrete Laplace/Poisson equation in Chapter 1, $\left[\phi_{ij}\right]$ is the state vector. In an ocean general circulation model, the state vector might consist of three components of velocity, pressure and density at each of millions of grid points, and it will be a function of time, $\mathbf{x}(t)$, as well. (One might want to regard the complete description,

$$\mathbf{x}_B=\left[\mathbf{x}\left(1\Delta t\right),\mathbf{x}\left(2\Delta t\right),..,\mathbf{x}\left(T\Delta t\right)\right]^T, \tag{2.1}$$

as the state vector, but by convention, it refers to the subvectors, $\mathbf{x}\left(n\Delta t\right)$, each of which is sufficient to compute any future one, given the boundary conditions.)

[3] Ide et al. (1997)

Consider a partial differential equation[4]

$$\frac{\partial}{\partial t}(\nabla_h^2 p) + \beta \frac{\partial p}{\partial \eta} = q\left(\xi, \eta, t\right) , \tag{2.2}$$

Suppose it is solved by an expansion

$$p(\xi,\, \eta,\, t) = \sum_{j=1}^{N/2} a_{2j}(t) \cos(\mathbf{k}_j \cdot \mathbf{r}) + a_{2j-1}(t) \sin(\mathbf{k}_j \cdot \mathbf{r}). \tag{2.3}$$

$[\mathbf{k}_j = (k_\xi,\, k_\eta),\, \mathbf{r} = (\xi,\, \eta)]$, then $\mathbf{a}(t) = \left[a_1(t)\; a_2(t) \cdots a_{2j-1}(t) \ldots\right]^T$ is a discrete description and becomes the $N-$dimensional state vector[5] Any adequate discretization can provide the state vector, it is non-unique, and careful choice can greatly simplify calculations.

In the most general terms, we can write any discrete model as a set of functional relations

$$\begin{gathered} L\big(\mathbf{x}(0), \ldots, \mathbf{x}(t-\Delta t),\, \mathbf{x}(t),\, \mathbf{x}(t+\Delta t), \ldots \mathbf{x}(t_f) \ldots,\, \mathbf{B}(t)\mathbf{q}(t)\,, \\ \mathbf{B}(t)\mathbf{q}(t+\Delta t), \ldots, t\big) = 0 \end{gathered} \tag{2.4}$$

where $\mathbf{B}(t)\mathbf{q}(t)$ represents a general, canonical, form for boundary and initial conditions/sources/sinks.[6] We almost always choose the time units so that $\Delta t = 1$. The static system equation

$$\mathbf{Ax} = \mathbf{q} \tag{2.5}$$

is a special case. In practice, the collection of relationships (2.4) always can be rewritten as a time-stepping rule–for example

$$\mathbf{x}(t+1) = \mathbf{L}\big(\mathbf{x}(t),\, \mathbf{B}(t)\mathbf{q}(t),\, t\big), \quad \Delta t = 1, \tag{2.6}$$

or, if the model is linear,

$$\mathbf{x}(t+1) = \mathbf{A}(t)\,\mathbf{x}(t) + \mathbf{B}(t)\,\mathbf{q}(t)\,. \tag{2.7}$$

If the model is time invariant, $\mathbf{A}(t) = \mathbf{A}$, and $\mathbf{B}$ may also be time invariant. These relationships have complete analogues in the continuous-time case; for example, (2.7) would be

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{A}_1(t)\,\mathbf{x}(t) + \mathbf{B}(t)\,\mathbf{q}(t)\,. \tag{2.8}$$

A time-dependent model is a set of rules for computing the state vector at time $t+1$ from knowledge of its values at time t and the externally imposed forces and boundary conditions. It is generally true that any linear discretized model can be put into this canonical form, although it may take some work. By the same historical conventions described in Chapter 1, solution

[4] Oceanographers will recognize this equation as that describing so-called barotropic Rossby waves. For present purposes, it is just one of many such equations one is faced with.

[5] Gaspar and Wunsch (1989)

[6] Construction of computer codes for the fluid dynamical equations is a large and complex subject in its own right. Readers interested in the general problem of oceanic numerical models can consult textbooks such as Haidvogel and Beckmann ().

of systems like (2.6), subject to appropriate initial and boundary conditions, constitutes the forward or direct problem.

Example: The Straight Line. The straight-line model, discussed in Chapter 1, can be put into the time-evolving framework. Let $\xi(t)$, satisfy the rule,

$$\frac{d^2\xi}{dt^2} = 0\,, \tag{2.9}$$

which can be discretized as

$$\xi(t+\Delta t) - 2\xi(t) + \xi(t-\Delta t) = 0\,. \tag{2.10}$$

Putting

$$x_1(t) = \xi(t)\,,\ x_2(t) = \xi(t-\Delta t)\,,$$

one has

$$\mathbf{x}(t+\Delta t) = \mathbf{A}\mathbf{x}(t)$$

where

$$\mathbf{A} = \begin{Bmatrix} 2 & -1 \\ 1 & 0 \end{Bmatrix}\,,$$

which is of the standard form (2.7), with $\mathbf{B} = \mathbf{0}$.

Example: The Mass-Spring Oscillator. The elementary mass-spring oscillator satisfies the differential equation

$$m\frac{d^2\xi(t)}{dt^2} + r\frac{d\xi(t)}{dt} + k\xi(t) = q(t)$$

where r is a damping constant. A simple one-sided time discretization produces

$$\begin{aligned} m\big(\xi(t+\Delta t) - 2\xi(t) + \xi(t-\Delta t)\big) + r\Delta t\big(\xi(t) - \xi(t-\Delta t)\big) + k(\Delta t)^2\,\xi(t) \\ = q(t)\,(\Delta t)^2 \end{aligned}$$

or

$$\begin{aligned} \xi(t+\Delta t) = {} & \left(2 - \frac{r\Delta t}{m} - \frac{k(\Delta t)^2}{m}\right)\xi(t) \\ & + \left(\frac{r\Delta t}{m} - 1\right)\xi(t-\Delta t) + (\Delta t)^2\,\frac{q(t)}{m}\,, \end{aligned}$$

which is

$$\begin{aligned} \begin{bmatrix} \xi(t+\Delta t) \\ \xi(t) \end{bmatrix} = {} & \begin{Bmatrix} 2 - \frac{r}{m}\Delta t - \frac{k}{m}(\Delta t)^2 & \frac{r\Delta t}{m} - 1 \\ 0 & 1 \end{Bmatrix} \begin{bmatrix} \xi(t) \\ \xi(t-\Delta t) \end{bmatrix} \\ & + \begin{bmatrix} (\Delta t)^2\,\frac{q(t)}{m} \\ 0 \end{bmatrix} \end{aligned} \tag{2.11}$$

and is the canonical form with $\mathbf{A}$ independent of time,

$$\mathbf{x}(t) = \begin{bmatrix}\xi(t) & \xi(t-\Delta t)\end{bmatrix}^T, \qquad \mathbf{B}(t)\mathbf{q}(t) = \begin{bmatrix}(\Delta t)^2\,q(t)/m & 0\end{bmatrix}^T.$$

Example: A Linear Difference Equation. A difference equation important in time-series analysis[7] is,

$$\xi(t) + a_1\xi(t-1) + a_2\xi(t-2) + \cdots + a_N\xi(t-N) = \eta(t) \tag{2.12}$$

where $\eta(t)$ is a zero-mean, white-noise process [Equation (2.12) is an example of an autoregressive process (AR)]. To put this into the canonical form, write[8],

$$\begin{aligned} x_1(t) &= \xi(t-N) \\ x_2(t) &= \xi(t-N+1) \\ &\vdots \\ x_N(t) &= \xi(t-1) \\ x_N(t+1) &= -a_1\,x_{N-1}(t) - a_2\,x_{N-2}(t) \cdots - a_N\,x_1(t) + \eta(t)\,. \end{aligned}$$

It follows that $x_1(t+1) = x_2(t)$, etc., or

$$\mathbf{x}(t+1) = \left\{ \begin{matrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ -a_N & -a_{N-1} & -a_{N-2} & \cdots & -a_2 & -a_1 \end{matrix} \right\} \mathbf{x}(t) + \begin{bmatrix} 0 \\ 0 \\ \cdot \\ 1 \end{bmatrix} \eta(t)\,. \tag{2.13}$$

Coefficients of the form in (2.13) are known as *companion matrices*; Equation (2.13) connects this Chapter with the field of time-series analysis. Here, $\mathbf{B}(t) = [0 \quad 0 \quad \cdot \quad 1]^T$, $\mathbf{q}(t) = \eta(t)$.

Given that most time-dependent models can be written as in (2.6) or (2.7), the forward model solution involves marching forward from known initial conditions at $t = 0$, subject to specified boundary values. So, for example, the linear model (2.7), with given initial conditions $\mathbf{x}(0) = \mathbf{x}_0$, involves the sequence,

$$\begin{aligned} \mathbf{x}(1) &= \mathbf{A}(0)\,\mathbf{x}_0 + \mathbf{B}(0)\,\mathbf{q}(0) \\ \mathbf{x}(2) &= \mathbf{A}(1)\,\mathbf{x}(1) + \mathbf{B}(1)\,\mathbf{q}(1) \\ &= \mathbf{A}(1)\,\mathbf{A}(0)\,\mathbf{x}_0 + \mathbf{A}(1)\,\mathbf{B}(0)\,\mathbf{q}(0) + \mathbf{B}(1)\,\mathbf{q}(1) \\ &\vdots \\ \mathbf{x}(t_f) &= \mathbf{A}(t_f-1)\,\mathbf{x}(t_f-1) + \mathbf{B}(t_f-1)\,\mathbf{q}(t_f-1) \\ &= \mathbf{A}(t_f-1)\,\mathbf{A}(t_f-2)\ldots\mathbf{A}(0)\,\mathbf{x}_0 + \ldots. \end{aligned}$$

Most of the basic ideas can be understood in the notationally simplest case of time-independent **A**, **B**, and that is usually the situation we will address with little loss of generality. Figure 1a

[7] Box et al. (1994)

[8] Luenberger (1979)

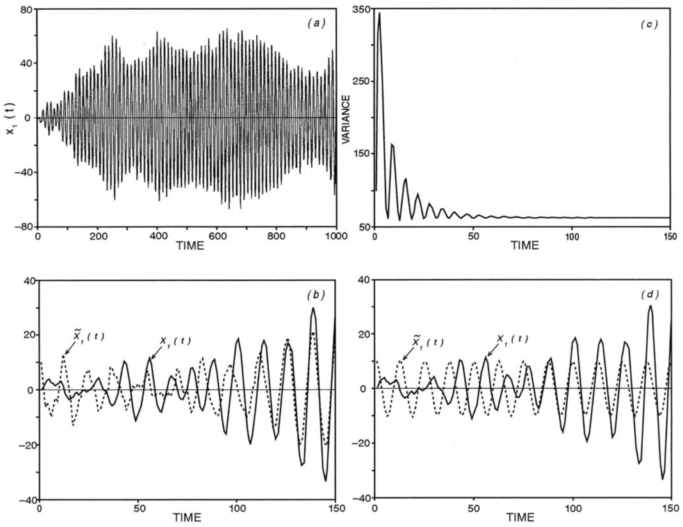


FIGURE 1. (a) The time history of $x_1(t)$ from the state vector of a linearized discrete oscillator driven by white noise forcing (the control). Here, $\Delta t = 1$, $k = 0.25$, $r = 0$, $Q = 1$, $\langle u(t) \rangle = 0$. The large excursions possible with a zero-mean small variance forcing are a simple illustration of the random walk phenomenon. (b) Early portion of the time history $x_1(t)$ (solid line) displayed in (a), and the Kalman filter estimate of it (dashed). The filter was started with $\tilde{\mathbf{x}}(0) = [10, 10]^T$ (correct value is $\mathbf{x}(0) = [0, 0]^T$), $\mathbf{P}(0) = 100\mathbf{I}$, $\mathbf{E} = [1, 0]$, $\mathbf{Q} = 1$, and $\mathbf{R} = 1000$. The controls were treated as completely unknown. The system gradually converges to the true state and is consistent within one standard error of the truth. (c) The variance $P_{11}(t)$ of the Kalman filter estimate displayed in (b). Note the initial oscillatory behavior derived from the dynamics and then the asymptote as the incoming data stream just balances the uncertainties introduced by the unknown $u(t)$. (d) The same as (b) except that the data were available only every 25 time-steps ($t = 25, 50, \ldots$). Convergence still takes place but is much slower.

depicts the time history for the harmonic oscillator, with the parameter choice $\Delta t = 1$, $k = 0.25$, $m = 1$, $r = 0$, so that

$$\mathbf{A} = \begin{Bmatrix} 1.75 & -1 \\ 1 & 0 \end{Bmatrix}, \qquad \mathbf{Bq}(t) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} u(t)$$

where $\langle u(t)^2 \rangle = 1$. The initial conditions were $\mathbf{x}(0) = \begin{bmatrix} \xi(0) & \xi(-1) \end{bmatrix}^T$. The general $t^{1/2}$ growth of the displacement of a randomly forced, undamped oscillator is apparent.

If the model is a "black box" in which $\mathbf{A}$ is unknown, it can be found in a number of ways. Suppose that $\mathbf{Bq}(t) = \mathbf{0}$, and Equation (2.7) is solved N times, starting at time t_0, subject to

$\mathbf{x}^{(i)}(t_0) =$ column i of $\mathbf{I}$—that is, the model is stepped forward for N different initial conditions corresponding to the N-different problems of unit initial condition at a single grid or boundary point, with zero initial conditions everywhere else. Let each column of $\mathbf{G}(t, t_0)$ correspond to the appropriate value of $\mathbf{x}(t)$—that is,

$$\begin{aligned}
\mathbf{G}(t_0, t_0) &= \mathbf{I} \\
\mathbf{G}(t_0 + 1, t_0) &= \mathbf{A}(t_0)\mathbf{G}(t_0, t_0) \\
\mathbf{G}(t_0 + 2, t_0) &= \mathbf{A}(t_0 + 1)\mathbf{G}(t_0 + 1, t_0) = \mathbf{A}(t_0 + 1)\mathbf{A}(t_0) \\
&\vdots \\
\mathbf{G}(t_0 + t, t_0) &= \mathbf{A}(t_0 + t - 1)\mathbf{A}(t_0 + t - 2) \cdots \mathbf{A}(t_0)\,.
\end{aligned}$$

We refer to $\mathbf{G}(t, t_0)$ as a *unit solution;* it is closely related to the Green function discussed in Chapter 2. The solution for arbitrary initial conditions is then,

$$\mathbf{x}(t) = \mathbf{G}(t, t_0)\mathbf{x}(t_0)\,, \tag{2.14}$$

and the modification for $\mathbf{Bq} \neq 0$ is straightforward.

$\mathbf{A}$ is necessarily square. It is also often true that $\mathbf{A}^{-1}$ exists, and a generalized inverse can be used if necessary. If $\mathbf{A}^{-1}$ can be computed, one can contemplate the possibility (important later) of running a model backward in time, for example as,

$$\mathbf{x}(t) = \mathbf{A}^{-1}\mathbf{x}(t + 1) - \mathbf{A}^{-1}\mathbf{B}(t)\,\mathbf{q}(t)\,.$$

Such a computation may be inaccurate if carried on for long times, but the same may well be true of the forward model.

Some attention must be paid to the structure of $\mathbf{B}(t)\,\mathbf{q}(t)$. The partitioning into these elements is not unique and can be done to suit one's convenience. The dimension of $\mathbf{B}$ is that of the size of the state vector by the dimension of $\mathbf{q}$, which typically would reflect the number of independent degrees of freedom in the forcing/boundary conditions. (*Forcing* is hereafter used to include boundary conditions, sources and sinks, and anything normally prescribed externally to the model.) Consider the model grid points displayed in Figure 2.. Suppose that the boundary grid points are numbered 1–5, 6, 10, 46–50, and all others are interior. If there are no interior forces, and all boundary values have a time history $q(t)$, then we could take,

$$\mathbf{B} = [1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0 \cdots 1 \quad 1]^T\,, \tag{2.15}$$

where the 1−s occur at the boundary points, and the zeros at the interior ones.

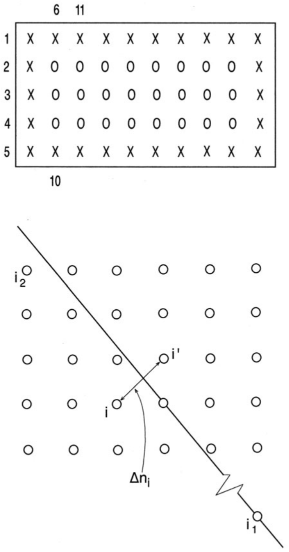


FIGURE 2. (a) Simple numerical grid for use of discrete form of model; × denote boundary grid points, and o are interior ones. Numbering is sequential down the columns, as shown. (b) Tomographic velocity integral is assumed given between i_1, i_2.

Suppose, instead, that boundary grid point 2 has values $q_1(t)$, all other boundary conditions are zero, and interior point 7 has a forcing history $q_2(t)$ and all others are unforced; then

$$\mathbf{B}\mathbf{q}(t) = \left\{\begin{matrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ \cdot & \cdot \\ 0 & 0 \end{matrix}\right\} \begin{bmatrix} q_1(t) \\ q_2(t) \end{bmatrix} . \tag{2.16}$$

A time-dependent $\mathbf{B}$ would correspond to time-evolving positions at which forces were prescribed—a somewhat unusual situation. It would be useful, for example, if one were driving a fluid model with a heat flux or stress in the presence of a prescribed moving ice cover. One could also impose initial conditions using a time-dependent $\mathbf{B}(t)$, which would vanish after $t = 0$.

As with steady models, we need to be careful about understanding the propagation of errors in time and space. If we have some knowledge of the initial oceanic state, $\tilde{\mathbf{x}}(0)$, and are doing an experiment at a later time t, the prior information—the estimated initial conditions—carries information in addition to what we are currently measuring. We seek to combine the two sets of information. How does information propagate forward in time? Formally, the rule (2.6) tells us exactly what to do. But because there are always errors in $\tilde{\mathbf{x}}(0)$, we need to be careful about assuming that a model computation of $\tilde{\mathbf{x}}(t)$ is useful. Depending upon the details of the model, one can qualitatively distinguish the behavior of the errors through time. (1) The model has decaying components. If the amplitudes of these components are partially erroneous, then for large enough t, these elements will have diminished, perhaps to the point where they are negligible. (2) The model has neutral components. At time t, the erroneous elements have amplitudes no larger than they were at $t = 0$. (3) The model has unstable components; at time t any erroneous parts may have grown to swamp everything else computed by the model.

Realistic models, particularly fluid ones, can contain all three types of behavior simultaneously. It thus becomes necessary to determine which elements of the forecast $\tilde{\mathbf{x}}(t)$ can be used to help estimate the system state by combination with new data, and which should be suppressed as partially or completely erroneous. Simply assuming all components are equally accurate can be a disastrous recipe.

Before proceeding, we reiterate the point that time need not be accorded a privileged position. From the inclusive state vector, $\mathbf{x}_B$ defined in Eq. (2.1). Then models of the form (2.7) can be written in the *whole-domain* form,

$$\mathbf{A}_B \mathbf{x}_B = \mathbf{d}_B$$

$$\mathbf{A}_B = \left\{ \begin{matrix} -\mathbf{A} & \mathbf{I} & \mathbf{0} & \cdot & \cdot & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathbf{A} & \mathbf{I} & \mathbf{0} & \cdot & \mathbf{0} & \mathbf{0} \\ \cdot & \cdot & \cdot & \cdot & \cdot & -\mathbf{A} & \mathbf{I} \end{matrix} \right\}, \quad \mathbf{d}_B = \begin{bmatrix} \mathbf{Bq}(0) \\ \mathbf{Bq}(1) \\ \vdots \end{bmatrix},$$

which is no different, except for its possibly enormous size, from that of a static system and can be handled by any of the methods of earlier chapters if the computational capacity is sufficient. Note the block-banded nature of $\mathbf{A}_B$.

2.2. How to Find the Matrix $\mathbf{A}(t)$. Most modern large-scale time-evolving models, even if completely linear, are written in terms of computer codes, typically in languages such as Fortran90 or C/C++. The state transition matrix is not normally explicitly constructed, instead individual elements of $x_i(t)$ are explicitly time-stepped to produce $x_i(t+1)$, usually using various vectorizations. For many purposes, $\mathbf{A}(t)$ is never explicitly required, as all one cares about is its operation on $\mathbf{x}(t)$, and the forward model code does the equivalent. But often, as will be seen below, the availability of $\mathbf{A}(t)$ can be very useful, and the question arises as to how one might obtain $\mathbf{A}(t)$ from the existing model code?

Several methods exist, but consider now only the case of a steady model, with no time-dependence in any of the governing model matrices $(\mathbf{A},\mathbf{B},\mathbf{\Gamma})$.For maximum simplicity, set $\mathbf{B}=\mathbf{\Gamma}=\mathbf{0}$.

(1) Define $N-$independent initial condition vectors $\mathbf{x}_0^{(i)}$, $1\leq i\leq N$, and form a matrix,

$$\mathbf{X}_0=\left\{\mathbf{x}_0^{(i)}\right\}.$$

Time step-the model once, equivalent to,

$$\mathbf{X}_1=\mathbf{A}\mathbf{X}_0$$

and invert:

$$\mathbf{A}=\mathbf{X}_1\mathbf{X}_0^{-1}.$$

The inverse will exist by the assumption of independence (spanning set) in the initial condition vectors. One must run the model $N-$times in this approach. A special, useful case, would be when $\mathbf{x}_o^{(i)}=\delta_{ij}$ (unit perturbation initial conditions), in which case the solution $\mathbf{X}_1$ would be the Green function.

The changes from $\mathbf{X}_0$ to $\mathbf{X}_1$ may be too small for adequate numerical accuracy, and one might use multiple time-steps, computing

$$\mathbf{X}_n=\mathbf{A}^n\mathbf{X}_0$$

which would determine $\mathbf{A}^n$, and $\mathbf{A}$ itself can be found by one of the matrix root algorithms.[9]

(2) Suppose the statistics of the solutions are known, e.g.,

$$\mathbf{R}\left(0\right)=\left\langle\mathbf{x}\left(t\right)\mathbf{x}\left(t\right)\right\rangle,\ \mathbf{R}\left(1\right)=\left\langle\mathbf{x}\left(t+1\right)\mathbf{x}\left(t\right)\right\rangle,$$

perhaps because the model has been run many times—making it possible to estimate these from stored output. Then noting,

$$\left\langle\mathbf{x}\left(t+1\right)\mathbf{x}\left(t\right)\right\rangle=\mathbf{A}\left\langle\mathbf{x}\left(t\right)\mathbf{x}\left(t\right)\right\rangle,$$

or

$$\mathbf{R}\left(1\right)=\mathbf{A}\mathbf{R}\left(0\right),$$

and

$$\mathbf{A}=\mathbf{R}\left(1\right)\mathbf{R}\left(0\right)^{-1}.$$

That is to say, knowledge of these covariances is equivalent to knowledge of the model itself (and vice-versa).[10] Again, multiple time steps can be used if necessary to infer $\mathbf{A}^n$.

Note that determination of $\mathbf{B},\mathbf{\Gamma}$, can be done analogously using a spanning set of $\mathbf{q}^{(i)},\mathbf{u}^{(i)}$ respectively, as boundary conditions, setting $\mathbf{x}\left(0\right)=0$.

[9] Menemenlis and Wunsch (1997); Stammer and Wunsch (1996)

[10] von Storch, et al. (1988).

(3) Automatic differentiation (AD) tools exist[11] which can take computer code (at the moment only Fortran, but other languages will become available soon), for the forward model, and produce by analysis of the code, equivalent computer code (Fortran) for construction of $\mathbf{A}$. Some codes preferentially produce $\mathbf{A}^T$, but simple transposition then can be employed.

If the model is fully time-dependent, then $\mathbf{A}\left(t\right)$ has be deduced at each separate time-step, as above. For some purposes, one might seek temporal averages, so defining and $\bar{\mathbf{A}}$, as

$$\bar{\mathbf{A}}^n = \mathbf{A}\left(0\right)\mathbf{A}\left(1\right)..\mathbf{A}\left(n-2\right)\mathbf{A}\left(n-1\right).$$

2.3. Observations and Data. Here, observations are introduced into the modeling discussion so that they stand on an equal footing with the set of model equations (2.6) or (2.7). Observations will be represented as a set of linear simultaneous equations at time t,

$$\mathbf{E}(t)\,\mathbf{x}(t) + \mathbf{n}(t) = \mathbf{y}(t)\,, \tag{2.17}$$

a straightforward generalization of the previous static systems where t did not appear explicitly; here, $\mathbf{E}$ is sometimes called the *design* or *observation* matrix. The notation used in Chapter 2 to discuss recursive estimation was deliberately chosen to be the same as here.

The requirement that the observations be linear combinations of the state-vector elements can be relaxed if necessary, but most common observations are of that form. [An obvious exception would be the situation in which the state vector included velocity components, $u(t)$, $v(t)$, but an instrument measuring speed, $\sqrt{(u(t)^2+v(t)^2)}$, would be a nonlinear relation between $y_i(t)$ and the state vector. Such systems are usually handled by some form of linearization.[12]]

To be specific, the noise $\mathbf{n}(t)$ is supposed to have zero mean and known second-moment matrix,

$$\langle \mathbf{n}(t)\rangle = 0, \qquad \langle \mathbf{n}(t)\,\mathbf{n}(t)^T\rangle = \mathbf{R}(t)\,. \tag{2.18}$$

But

$$\langle \mathbf{n}(t)\,\mathbf{n}(t')^T\rangle = \mathbf{0}, \quad t \neq t'\,. \tag{2.19}$$

That is, the observational noise should not be correlated from one measurement time to another; there is a considerable literature on how to proceed when this crucial assumption fails (called the *colored-noise problem*[13]). Unless specifically stated otherwise, we will assume that (2.19) is valid.

The matrix $\mathbf{E}(t)$ can accommodate almost any form of linear measurement. If, at some time, there are no measurements, then $\mathbf{E}(t)$ vanishes, along with $\mathbf{R}(t)$. If a single element $x_i(t)$ is measured, then $\mathbf{E}(t)$ is a row vector that is zero everywhere except in column i, where it is

[11] Giering and Kaminski (1997), Marotzke et al. (1999).

[12] For example, Bryson and Ho (1975, p. 351).

[13] For example, Stengel (1986).

1. It is particularly important to recognize that many measurements are weighted averages of the state-vector elements. Some measurements—for example, tomographic ones[14] as described in Chapter 1—are explicitly spatial averages (integrals) obtained by measuring some property along a ray travelling between two points; see Fig. 2. Any such data representing spatially filtered versions of the state vector can be written

$$y(t) = \sum \alpha_j x_j(t)\,. \tag{2.20}$$

Point observations often occur at positions not coincident with model grid positions (although many models, e.g., spectral ones, do not use grids). Then (2.17) is an interpolation rule, possibly either very simple or conceivably a full-objective mapping calculation, of the value of the state vector at the measurement point. Often the number of model grid points vastly exceeds the number of the data grid points; thus, it is convenient that the formulation (2.17) demands interpolation from the dense model grid to the sparse data positions. (In the unusual situation where the data density is greater than the model grid density, one can restructure the problem so the interpolation goes the other way.) More complex filtered measurements exist. In particular, one may have measurements of a state vector only in specific wavenumber bands; but such *band-passed* observations are automatically in the form (2.17).

As with the model, the observations of the combined state vector can be concatenated into a single observational set

$$\mathbf{E}_B\mathbf{x}_B + \mathbf{n}_B = \mathbf{y}_B \tag{2.21}$$

where

$$\mathbf{E}_B = \left\{\begin{matrix} 0 & 0 & 0 & \cdot & 0 \\ 0 & \mathbf{E}(1) & 0 & \cdot & 0 \\ 0 & 0 & \mathbf{E}(2) & \cdot & \cdot \\ \cdot & \cdot & \cdot & 0 & \mathbf{E}(t_f) \end{matrix}\right\}, \quad \mathbf{n}_B = \begin{bmatrix} 0 \\ \mathbf{n}(1) \\ \vdots \\ \mathbf{n}(t_f) \end{bmatrix}, \quad \mathbf{y}_B = \begin{bmatrix} 0 \\ \mathbf{y}(1) \\ \vdots \\ \mathbf{y}(t_f) \end{bmatrix}.$$

$\mathbf{E}_B$ is also block-banded and often very sparse. If the size is no problem, the concatenated model and observations could be dealt with using any of the methods of Chapter 2. The rest of this chapter can be thought of as an attempt to produce from the model/data combination the same type of estimates as were found useful in Chapter 2, but exploiting the special structure of matrices $\mathbf{A}_B$ and $\mathbf{E}_B$ so as to avoid having to store them in the computer all at once, which may be physically impossible.

The unit solution formulation of Chapter 2, Eq. (**??**) leads to a particularly simple reduced form if, for example, only the initial conditions $\mathbf{x}(0)$ are unknown, one has immediately

$$\mathbf{y}(t) = \mathbf{E}(t)\mathbf{G}(t,0)\mathbf{x}(0) + \mathbf{n}(t)\,, \quad 1 \le t \le t_f\,,$$

[14] Munk, Worcester, & Wunsch (1995).

which are readily solved in whole-domain form for $\mathbf{x}(0)$. If only a subset of the $\mathbf{x}(0)$ are thought to be nonzero, the columns of $\mathbf{G}$ need to be computed only for those elements.[15]

3. Estimation

3.1. Model and Data Consistency. In many scientific fields, the central issue is to make estimates of what is going on from the combination of a model with measurements. The model is intended to encompass all one's theoretical knowledge about how the system behaves, and the data are the complete observational knowledge of the same system. If done properly, inferences from the model/data combination should be no worse, and may well be very much better, than those made from either alone. It is the latter possibility that motivates the development of state estimation procedures. "Best-estimates" made by combining models with observations are often used to forecast a system (e.g., to land an airplane), but this is by no means the major application.

Such model/data problems are ones of statistical inference with a host of specific subtypes. Some powerful techniques are available, but like any powerful tools (a chain saw, for example), they can be dangerous to the user! In general, one is confronted with a two-stage problem. Stage 1 involves developing a suitable model that is likely to be consistent with the data. "Consistent" means that within the estimated data errors (and obtaining the data errors is itself a serious modeling problem), the model is likely to be able to describe the features of interest. One can go very badly wrong here, before any computation takes place. If one attempts to model elastic wave propagation using the equations of fluid dynamics, estimation methods will commonly produce some kind of "answer", but one which would be nonsensical. Model failure can of course, be much more subtle, in which some supposed secondary element (e.g., a time-dependence) proves to be critical to a description of the data. Good technique alerts users to the presence of such failures, along with clues as to what should be changed in the model. But these issues do not however, apply only to the model. The assertion that a particular data set carries a signal of a particular kind, can prove to be false in a large number of ways. A temperature signal thought to represent the seasonal cycle might prove, on careful examination to be dominated by higher or lower frequency structures, and thus its use with an excellent model of annual variation might prove disastrous. Whether this is regarded as a data or as a model issue is evidently somewhat arbitrary.

Thus stage 1 of any estimation problem has to involve understanding of whether the data and the model are physically and statistically consistent. If they are not, one should stop. Often where they are believed to be generally consistent up to certain quantitative adjustments, one can combine the two stages. A model may have adjustable parameters (turbulent mixing coefficients;

[15] A method exploited by Stammer and Wunsch (1996).

boundary condition errors; etc.) which could bring the model and the data into consistency, and then the estimation procedure becomes, in part, an attempt to find those parameters in addition to the state. Alternatively, the data error covariance, $\mathbf{R}(t)$, may be regarded as incompletely known, and one might seek as part of the state estimation procedure to improve one's estimate of it. (Problems like this one are taken up here under the subject of "adaptive filtering.")

Assuming for now that the model and data are likely to prove consistent, one can address what might be thought of as a form of interpolation: Given a set of observations in space and time as described by Equation (2.17), use the dynamics as described by the model (2.6) or (2.7) to estimate various state-vector elements at various times of interest. Yet another, less familiar, problem recognizes that some of the forcing terms $\mathbf{B}(t)\mathbf{q}(t)$ are partially or wholly unknown (e.g., no one believes that the windstress boundary conditions over the ocean are perfectly known), and one might seek to estimate them from whatever ocean observations are available and from the known model dynamics. Many real problems contain both these issues..

The forcing terms—representing boundary conditions as well as interior sources/sinks and forces—almost always need to be divided into two elements: the known and the unknown parts. The latter will often be perturbations about the known values. Thus, rewrite (2.7) in the modified form

$$\mathbf{x}(t+1) = \mathbf{A}(t)\,\mathbf{x}(t) + \mathbf{B}(t)\,\mathbf{q}(t) + \Gamma(t)\,\mathbf{u}(t), \tag{3.1}$$

where now $\mathbf{B}(t)\,\mathbf{q}(t)$ represent the known forcing terms and $\Gamma(t)\,\mathbf{u}(t)$ the unknown ones, which we will generally refer to as the *controls*, or *control terms.* $\Gamma(t)$ is known and plays the same role for $\mathbf{u}(t)$ as does $\mathbf{B}(t)$ for $\mathbf{q}(t)$. Usually $\mathbf{B}(t)$, $\Gamma(t)$ will be treated as time independent, but this simplification is not necessary. Almost always, we can make some estimate of the size of the control terms, as for example,

$$\langle \mathbf{u}(t) \rangle = 0, \qquad \langle \mathbf{u}(t)\,\mathbf{u}(t)^T \rangle = \mathbf{Q}(t)\,. \tag{3.2}$$

The controls have a second, somewhat different role: They can also represent the model error. Most models are inaccurate to a degree—approximations are often made to the equations describing any particular physical situation. One can expect that the person who constructed the model has some idea of the size and structure of the physics or chemistry, or biology,... that has been omitted or distorted in the model construction. In this context, $\mathbf{Q}$ represents the covariance of the model error, and the control terms represent the missing physics. The assumption $\langle \mathbf{u}(t) \rangle = \mathbf{0}$ must be critically examined in this case, and in the event of failure, some modification of the model must be made or the variance artificially modified to attempt to accommodate what becomes a model bias error. But the most serious problem is that models are rarely produced with *any* quantitative description of their accuracy beyond one or two examples of comparison with known solutions, and one is left to make pure guesses at $\mathbf{Q}(t)$. Getting beyond such guesses is again a problem of adaptive estimation addressed later.

Collecting the standard equations of model and data:

$$\mathbf{x}\left(t+1\right) = \mathbf{A}\left(t\right)\mathbf{x}\left(t\right) + \mathbf{B}\mathbf{q}\left(t\right) + \boldsymbol{\Gamma}\mathbf{u}\left(t\right), \; 0 \le t \le t_f - 1 \tag{3.3}$$

$$\mathbf{E}\left(t\right)\mathbf{x}\left(t\right) + \mathbf{n}\left(t\right) = \mathbf{y}\left(t\right), \quad 1 \le t \le t_f \tag{3.4}$$

$$\mathbf{n}\left(t\right) = \mathbf{0}, \quad \left\langle \mathbf{n}\left(t\right)\mathbf{n}\left(t\right)^T \right\rangle = \mathbf{R}\left(t\right), \quad \left\langle \mathbf{n}\left(t\right)\mathbf{n}\left(t'\right)\right\rangle = \mathbf{0}, \; t \neq t' \tag{3.5}$$

$$\left\langle \mathbf{u} \right\rangle = \mathbf{0}, \quad \left\langle \mathbf{u}\left(t\right)\mathbf{u}\left(t\right)\right\rangle = \mathbf{Q}\left(t\right) \tag{3.6}$$

$$\tilde{\mathbf{x}}\left(0\right) = \mathbf{x}_0, \quad \left\langle \left(\tilde{\mathbf{x}}\left(0\right) - \mathbf{x}\left(0\right)\right)\left(\tilde{\mathbf{x}}\left(0\right) - \mathbf{x}\left(0\right)\right)^T \right\rangle = \mathbf{P}\left(0\right) \tag{3.7}$$

where t_f defines the endpoint of the interval of interest. The last equation, (3.7), treats the initial conditions of the model as a special case—the uncertain initialization problem, where $\mathbf{x}(0)$ is the true initial condition and $\tilde{\mathbf{x}}(0) = \mathbf{x}_0$ is the value actually used but with uncertainty $\mathbf{P}(0)$. Alternatively, one could write

$$\mathbf{E}(0)\mathbf{x}\left(0\right) + \mathbf{n}\left(0\right) = \mathbf{x}_0, \; \mathbf{E}\left(0\right) = \mathbf{I}, \; \left\langle \mathbf{n}\left(0\right)\mathbf{n}\left(0\right)^T \right\rangle = \mathbf{P}\left(0\right), \tag{3.8}$$

and include the initial conditions as a special case of the observations—recognizing explicitly that one often obtains initial conditions from observations.

This general form permits one to grapple with reality. In the spirit of ordinary least squares and its intimate cousin, minimum-error variance estimation, consider the general problem of finding state vectors and controls, $\mathbf{u}\left(t\right)$, that minimize an objective function,

$$\begin{aligned} J = &\left(\mathbf{x}(0) - \mathbf{x}_0\right)^T \mathbf{P}(0)^{-1}\left(\mathbf{x}(0) - \mathbf{x}_0\right) \\ &+ \sum_{t=1}^{t_f} \left(\mathbf{E}(t)\,\mathbf{x}(t) - \mathbf{y}(t)\right)^T \mathbf{R}(t)^{-1}\left(\mathbf{E}(t)\,\mathbf{x}(t) - \mathbf{y}(t)\right) \\ &+ \sum_{t=0}^{t_f-1} \mathbf{u}(t)^T \mathbf{Q}(t)^{-1}\mathbf{u}(t), \end{aligned} \tag{3.9}$$

subject to the model, (3.3, 3.6) As written here, this choice of an objective function is somewhat arbitrary, but perhaps reasonable, as the direct analogue to those used in Chapter 2. It seeks a state vector $\mathbf{x}(t)$, $0 \le t \le t_f$, and a set of controls, $\mathbf{u}(t)$, $0 \le t \le t_f - 1$, that satisfy the model and that agree with the observations to an extent determined by the weight matrices $\mathbf{R}(t)$ and $\mathbf{Q}(t)$, respectively. From the previous discussions of least squares and minimum-error variance estimation, the minimum-square requirement Eq.(3.9) will produce a solution identical to that derived from minimum variance estimation by the specific choice of the weight matrices as the corresponding prior uncertainties, $\mathbf{R}(t)$, $\mathbf{Q}(t)$, $\mathbf{P}(0)$. In a Gaussian system, it also proves to be the maximum likelihood estimate. The introduction of the controls, $\mathbf{u}(t)$, into the objective function, represents an acknowledgment that arbitrarily large controls (forces) would not usually be an acceptable solution; they should be consistent with $\mathbf{Q}\left(t\right)$.

Much of the rest of this chapter will be directed at solving the problem of finding the minimum of J subject to the model. Notice that J involves the state vector, the controls, and the observations over the entire time period under consideration, $0 \leq t \leq t_f$. This type of objective function is the one usually of most interest to scientists attempting to understand their system—in which data are stored and employed over a finite time. In some other applications, most notably forecasting and which is taken up immediately below, one has only the past measurements available; this situation proves to be a special case of the more general one.

Although we will not keep repeating the warning each time an objective function such as Eq. (3.9) is encountered, the reader is reminded of the general message: The assumption that the model and observations are consistent and that the minimum of the objective function produces a meaningful and useful estimate must always be tested after the fact. That is, at the minimum, $\tilde{\mathbf{u}}(t)$ must prove consistent with $\mathbf{Q}(t)$, and $\tilde{\mathbf{x}}(t)$ must produce residuals consistent with $\mathbf{R}(t)$. Failure of these and other posterior tests should lead to rejection of the model. As always, one can thus reject a model [which includes $\mathbf{Q}(t)$, $\mathbf{R}(t)$] on the basis of a failed consistency with observations. One can never prove a model "correct," merely "consistent."[16]

3.2. The Kalman Filter. We begin with a special case. Suppose that by some means at time $t = 0$ we have an unbiased estimate, $\tilde{\mathbf{x}}(0)$, of the state vector with known uncertainty $\mathbf{P}(0)$. At time $t = 1$, observations from Eq. (3.4) are available. How would the information available best be used to estimate $\mathbf{x}(1)$?

The model permits a forecast of what $\mathbf{x}(1)$ should be, were $\tilde{\mathbf{x}}(0)$ known perfectly,

$$\tilde{\mathbf{x}}(1,-) = \mathbf{A}\tilde{\mathbf{x}}(0) + \mathbf{B}\mathbf{q}(0), \tag{3.10}$$

where the unknown control terms have been replaced by the best estimate we can make of them—their mean, which is zero, and $\mathbf{A}$ has been assumed to be time independent. A minus sign has been introduced into the argument of $\tilde{\mathbf{x}}(1,-)$ to show that *no data have yet been used to make the estimate* at $t = 1$, in a notation we will generally use. How good is this forecast?

Suppose the erroneous components of $\tilde{\mathbf{x}}(0)$ are,

$$\boldsymbol{\gamma}(0) = \tilde{\mathbf{x}}(0) - \mathbf{x}(0)\,, \tag{3.11}$$

[16] Some modelers like to claim "validation" or "verification" of their models. The fallacy of this idea is discussed by Oreskes et al. (1994).

then the erroneous components of the forecast are,

$$\begin{aligned}\boldsymbol{\gamma}(1) \equiv \tilde{\mathbf{x}}(1,-) - \mathbf{x}(1) &= \mathbf{A}\tilde{\mathbf{x}}(0) + \mathbf{B}\mathbf{q}(0) - \big(\mathbf{A}\mathbf{x}(0) + \mathbf{B}\mathbf{q}(0) + \Gamma\mathbf{u}(0)\big) \\ &= \mathbf{A}\boldsymbol{\gamma}(0) - \Gamma\mathbf{u}(0)\,,\end{aligned} \tag{3.12}$$

that is, composed of two distinct elements: the propagated erroneous portion of $\tilde{\mathbf{x}}(0)$, and the unknown control term. Their second moments are

$$\begin{aligned}\langle\boldsymbol{\gamma}(1)\,\boldsymbol{\gamma}(1)^T\rangle &= \langle\big(\mathbf{A}\boldsymbol{\gamma}(0) - \Gamma\mathbf{u}(0)\big)\big(\mathbf{A}\boldsymbol{\gamma}(0) - \Gamma\mathbf{u}(0)\big)^T\rangle \\ &= \mathbf{A}\langle\boldsymbol{\gamma}(0)\,\boldsymbol{\gamma}(0)^T\rangle\mathbf{A}^T + \Gamma\langle\mathbf{u}(0)\,\mathbf{u}(0)^T\rangle\Gamma^T \\ &= \mathbf{A}\mathbf{P}(0)\mathbf{A}^T + \Gamma\mathbf{Q}(0)\Gamma^T \\ &\equiv \mathbf{P}(1,-)\end{aligned} \tag{3.13}$$

by the definitions of $\mathbf{P}(0)$, $\mathbf{Q}(0)$ and the assumption that the unknown controls are not correlated with the error in the state estimate at $t = 0$. We now have an estimate of $\mathbf{x}(1)$ with uncertainty $\mathbf{P}(1,-)$ and a set of observations,

$$\mathbf{E}(1)\,\mathbf{x}(1) + \mathbf{n}(1) = \mathbf{y}(1)\,. \tag{3.14}$$

To combine the two sets of information, we use the recursive least-squares solution Ch. 2 Eqs. (8.7, 8.8). By assumption, the uncertainty in $\mathbf{y}(1)$ is uncorrelated with that in $\tilde{\mathbf{x}}(1,-)$. Making the appropriate substitutions into those equations,

$$\begin{aligned}\tilde{\mathbf{x}}(1) &= \tilde{\mathbf{x}}(1,-) + \mathbf{K}(1)\left[\mathbf{y}(1) - \mathbf{E}(1)\,\tilde{\mathbf{x}}(1,-)\right], \\ \mathbf{K}(1) &= \mathbf{P}(1,-)\,\mathbf{E}(1)^T\left[\mathbf{E}(1)\,\mathbf{P}(1,-)\,\mathbf{E}(1)^T + \mathbf{R}(1)\right]^{-1}\end{aligned} \tag{3.15}$$

with new uncertainty

$$\mathbf{P}(1) = \mathbf{P}(1,-) - \mathbf{K}(1)\,\mathbf{E}(1)\,\mathbf{P}(1,-)\,. \tag{3.16}$$

Thus there are four steps:

(1) Make a forecast using the model (3.3) with the unknown control terms $\mathbf{\Gamma u}$ set to zero.

(2) Calculate the uncertainty of this forecast, (3.13), which is made up of two separate terms.

(3) Do a weighted average (3.15) of the forecast with the observations, the weighting being chosen to reflect the relative uncertainties.

(4) Compute the uncertainty of the final weighted average 3.16..

Such a computation is called a *Kalman filter*.[17]. It is conventionally given a more formal derivation. **K** is called the *Kalman gain*. At the stage where the forecast (3.10) has already

[17] Kalman (1960). Kalman's derivation was for this discrete case. The continuous case, derived later, is known as the Kalman-Bucy filter and is a much more complicated object.

been made, the problem was reduced to finding the minimum of the objective function,

$$\begin{aligned} J = & \left[\tilde{\mathbf{x}}(1,-) - \mathbf{x}(1)\right]^T \mathbf{P}(1,-)^{-1} \left[\tilde{\mathbf{x}}(1,-) - \mathbf{x}(1)\right] \\ & + \left[\mathbf{y}(1) - \mathbf{E}(1)\,\mathbf{x}(1)\right]^T \mathbf{R}(1)^{-1} \left[\mathbf{y}(1) - \mathbf{E}(1)\,\mathbf{x}(1)\right], \end{aligned} \tag{3.17}$$

which is a special case of the objective function used to define the recursive least-squares algorithm in Ch. 2. In this final stage, the explicit model has disappeared, being present only implicitly through the uncertainty $\mathbf{P}(1,-)$. Notice that the model is being satisfied exactly; in the terminology introduced in Chapter 2, it is a hard constraint. But again, as was true with the static models, the hard constraint description is somewhat misleading, as the presence of the terms in $\mathbf{u}$ means that model errors are permitted.

A complete recursion can now be defined through the Equations (3.10)–(3.16), replacing all the $t = 0$ variables with $t = 1$ variables, the $t = 1$ variables becoming $t = 2$ variables, etc. In terms of arbitrary t, the recursion is

,

$$\tilde{\mathbf{x}}\,(t,-) = \mathbf{A}\,(t-1)\,\tilde{\mathbf{x}}\,(t-1) + \mathbf{B}\,(t-1)\,\mathbf{q}\,(t-1) \tag{3.18}$$

$$\mathbf{P}\,(t,-) = \mathbf{A}\,(t-1)\,\mathbf{P}\,(t-1)\,\mathbf{A}\,(t-1) + \mathbf{\Gamma}\mathbf{Q}\,(t-1)\,\mathbf{\Gamma}^T \tag{3.19}$$

$$\tilde{\mathbf{x}}\,(t) = \tilde{\mathbf{x}}\,(t,-) + \mathbf{K}\,(t)\,[\mathbf{y}\,(t) - \mathbf{E}\,(t)\,\tilde{\mathbf{x}}\,(t,-)] \tag{3.20}$$

$$\mathbf{K}\,(t) = \mathbf{P}\,(t,-)\,\mathbf{E}\,(t)^T \left[\mathbf{E}\,(t)\,\mathbf{P}\,(t,-)\,\mathbf{E}\,(t)^T + \mathbf{R}\,(t)\right]^{-1} \tag{3.21}$$

$$\mathbf{P}\,(t) = \mathbf{P}\,(t,-) - \mathbf{K}\,(t)\,\mathbf{E}\,(t)\,\mathbf{P}\,(t,-)\,, \quad 1 \le t \le t_f \tag{3.22}$$

These equations are those for the complete Kalman filter. Note that some authors prefer to write it for $\tilde{\mathbf{x}}\,(t+1,-)$ in terms of $\tilde{\mathbf{x}}\,(t)$, etc. The matrix inversion lemma permits one to rewrite Eq. (3.22) as,

$$\mathbf{P}\,(t) = \left[\mathbf{P}\,(t,-)^{-1} + \mathbf{E}\,(t)^T\,\mathbf{R}\,(t)^{-1}\,\mathbf{E}\,(t)\right]^{-1}, \tag{3.23}$$

and an alternate form for the gain is[18]

$$\mathbf{K}\,(t) = \mathbf{P}\,(t)\,\mathbf{E}\,(t)^T\,\mathbf{R}\,(t)^{-1}, \tag{3.24}$$

and other rearrangements are possible too. These re-written forms are often important for computational efficiency and accuracy.

If observations are not available at some time step, t, the best estimate reduces to that from the model forecast alone, $\mathbf{K}\,(t) = 0$, $\mathbf{P}\,(t) = \mathbf{P}\,(t,-)$ and one simply proceeds. Typically in such situations, the error variances will grow from the accumulation of the unknown $\mathbf{u}\,(t)$, at least,

[18] Stengel (1986, Eq. 4.3-22)

until such times as an observation does become available. If $\mathbf{u}(t)$ is purely random, the system will undergo a form of random walk.[19]

Example: Straight Line. Let us reconsider the problem of fitting a straight line to data, as discussed in Chapter 2, but now in the context of a Kalman filter, using the canonical form derived from (3.18-3.22). "Data" were generated as depicted in Figure 3—a straight line plus noise. The observation equation is

$$y(t) = x_1(t) + n(t)\,,$$

that is, $\mathbf{E}(t) = \{1 \quad 0\}$, $R(t) = 100$. The model was assumed perfect, $Q(t) = 0$, but the initial state estimate was set erroneously as $\mathbf{x}(0) = [10 \quad 10]^T$ with an uncertainty

$$\mathbf{P}(0) = \begin{Bmatrix} 100 & 0 \\ 0 & 100 \end{Bmatrix}.$$

The result of the computation for the fit is shown in Figure 3 for 50 time steps. Figure 4 shows the standard error, $\sqrt{P_{11}(t')}$ of $x_1(t)$, and its decline as observations accumulate.

If the state vector is redefined to consist of the two model parameters a, b, then $\mathbf{x} = [a \quad b]^T$ and $\mathbf{A} = \mathbf{I}$. Now the observation matrix is $\mathbf{E} = [1 \quad t]$–that is, time-dependent. Notice the radical change in the state vector from a time-varying one to a constant. The same grossly incorrect estimates $\tilde{\mathbf{x}}(0) = [10 \quad 10]^T$ were used, with the same $\mathbf{P}(0)$ (the correct values are $a = 1$, $b = 2$) and with the time histories of the estimates depicted in Figure 4a. At the end of 100 time steps, we have $\tilde{a} = 1.85 \pm 2.0$, $\tilde{b} = 2.0 \pm 0.03$, both of which are consistent with the correct values. For reasons the reader might wish to think about, the uncertainty of the intercept is much greater than for the slope.

Example: The Mass-Spring Oscillator. Consider the mass-spring oscillator described earlier with time history in Figure 1a. It was supposed that the initial conditions were perfectly known, but that the forcing was completely unknown. Noisy observations of $x_1(t)$ were provided at every time step with a noise variance $R = 9$. The Kalman filter was computed by (3.18-3.22) and used to estimate the position at each time step. The result for part of the time history is in Figure 1b, showing the true value and the estimated value. The time history of the uncertainty of $x_1(t)$ is depicted in Figure 1c and is reaching an asymptote. Overall, the filter manages to track the position of the oscillator within two standard deviations.

It was then supposed that the same noisy observations were available but only at every 25th time step. In general, the presence of the model error, or control uncertainty, accumulates over the 25 time steps as the model is run forward without observations. The expected error of such a system is shown for 150 time steps in Figure 5e. Notice (1) the growing envelope as uncertainty accumulates faster than the observations can reduce it; (2) the periodic nature of

[19] Feller (195x).

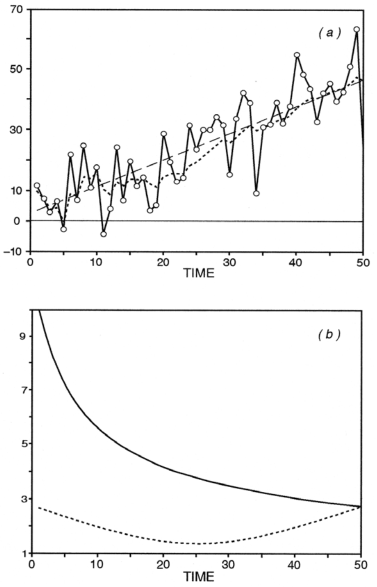


FIGURE 3. (a) For the model of a straight line in time, the "0" are the observations, $\mathbf{y}$, and the dotted line is the Kalman filter estimate for the first 50 time steps. There is an observation at each time-step. The filter does a reasonable job of tracking the straight line in the presence of noisy observations, converging to it at the end. The dashed line is the result of the RTS smoother applied to the results of the Kalman filter. Now the result is very near the correct straight-line state estimate. (b) Standard error of the Kalman filter estimate of x_1 (solid line), and the standard error of the smoothed estimate (dotted line). The correct values and the estimated values are generally within one standard error of each other, as they should be most of the time. Notice that the filter and smoother covarance begin with identical values at the terminal time with the smoother error reaching a minimum near the center of the time interval, increasing again toward $t = 0$ owing to the very large initial estimate of $\mathbf{P}(0)$. Again, the best estimates are, unsurprisingly, in the center of the observation period.

the error within the growing envelope; and (3) that the envelope appears to be asymptoting to a fixed upper bound for large t. The true and estimated time histories for a portion of the time history are shown in Figure 1d. With fewer available observations, as expected, the misfit of the estimated and true values is larger than with data at every time step. At every 25th point, the error norm drops as observations become available.

If the observation is that of the velocity $x_1(t+1) - x_2(t+1) = \xi(t+1) - \xi(t)$, then $\mathbf{E} = \{1 \quad -1\}$. A portion of the time history of the Kalman filtered estimate with a velocity observation available only at every 25th point may be seen in Figure 5f. Velocity observations are

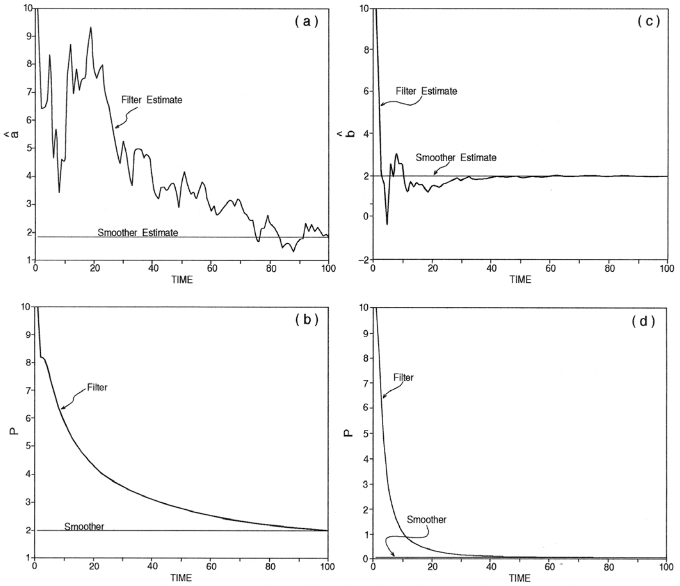


FIGURE 4. (a) Estimate of straight-line intercept, a, when the state vector was defined to be $\mathbf{x} = [a, b]^T$. The initial estimate was 10, when the correct value was 1. Notice that the smoothed estimate is a constant over the entire domain, consistent with the dynamical model for $\mathbf{x}$. (b) Standard errors of the filter and smoother for estimates in (a). (c) Filter and smoother estimates of straight-line slope (correct value is $b = 2$) when slope was part of the state vector. Convergence is much more rapid than for the intercept. (d) The standard errors of the filtered and smoothed estimate in (c).

evidently useful for estimating position, owing to the connection between velocity and position provided by the model and is a simple example of how observations of almost anything can be used to improve a state estimate.

The Kalman filter does *not* produce the minimum of the objective function Eq. (3.9) because the data from times later than t are not being used to make estimates of the earlier values of the state vector or of $\mathbf{u}(t)$. At each step, the Kalman filter is instead minimizing an objective function of the form Eq. (3.17), where t corresponds to the previous $t = 0$, and $t+1$ corresponds to $t = 1$. To obtain the needed minimum we have to consider what is called the *smoothing problem*, to which we will turn in a moment.

But the Kalman filter is extremely important in practice for many problems. In particular, if one must literally make a forecast (e.g., such filters are used to help land airplanes or, in a primitive way, to forecast the weather), then the future data are simply unavailable, and the state estimate made at time $t+1$, using data up to and including time $t+1$, is the best one can

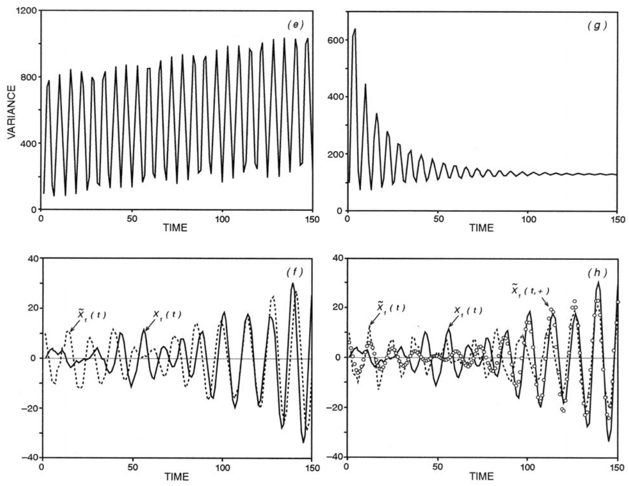


FIGURE 5. (e) The variance $P_{11}(t)$ for the Kalman filter estimate in Figure 1a, when data were available only every 25th time-step. An error asymptote is still achieved, but only after a longer time than shown and only in the envelope, with the values oscillating between data availability. Values are considerably higher than in 1c. (f) The Kalman filter estimate for the first part of Figure 1a, with the same parameters, except here the observation is supposed to be of the "velocity": $y(t) = x_1(t) - x_2(t) + n(t)$, ($\mathbf{E} = [1, -1]$), $\mathbf{R} = 1000$, at every time-step. Convergence to the correct state still takes place but is slow. (g) The variance, $P_{11}(t)$, for the filter estimate in (f), which can be compared to that in 1c. The slower convergence to the correct solution in (f) is consistent with the slower asymptote and the higher ultimate value of the uncertainty that is achieved. (h) Smoothed estimate, $\tilde{x}_1(t,+)$ (shown as open circles, from the RTS smoother started from $t = t_f = 150$ from the Kalman filter estimate $\tilde{\mathbf{x}}(150)$ in Figure 1b. The correct value, $x_1(t)$, is the solid curve. As expected in a smoothing algorithm, the smoother significantly improves agreement with the true values in the center of the interval. Its value and uncertainty agree with that of the Kalman filter at $t = t_f$ and somewhat improves the values at the time origin.

do. Some history of the idea of the filter, its origins in the work of Wiener and Kolmogoroff . (Much of the discussion in the wider literature is in terms of continuous time. The continuous form of the filter is usually known as the *Kalman-Bucy filter*.[20])

[20] See, for example, Bucy & Joseph (1968). The Kalman filter is so important that entire books are devoted to its theory and practice, and many thousands of papers have been published on it and its extensions. A history with a number of applications can be found in Sorenson (1985). Ghil, Cohn, Tavantzis, Bube, and Isaacson, (1981) discuss it in a meteorological context; Miller (1986), Wunsch (1988b), Miller and Cane (1989), Fu, Vazquez, and

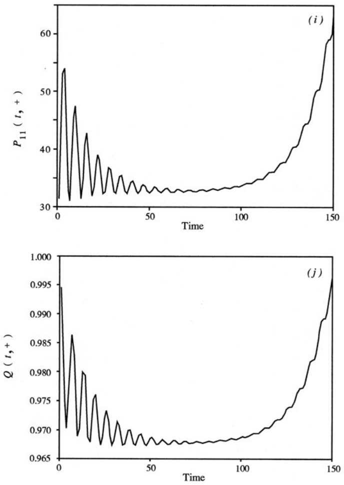


FIGURE 6. (i) Uncertainty, $P_{11}(t,+)$, for the RTS smoother estimate in Figure 5h showing the minimum expected error in the center of the interval. This uncertainty should be compared to that of the filter in Figure 1c. (j) Uncertainty estimate $Q(t,+)$, for the control estimate, $\tilde{u}(t,+)$, from the RTS smoother. [$\tilde{u}(t,+)$ is not shown, being a set of random numbers, as is $u(t)$, and is uninteresting. The two do agree within the uncertainty shown. Note that the data error here was very large, and hence the final uncertainty is little reduced over the initial value of 1.

For estimation, the Kalman filter is only a first step owing to its failure to use data from the formal future. It also raises questions about computational feasibility. As with all recursive estimators, the uncertainties $\mathbf{P}(t,-)$, $\mathbf{P}(t)$ must be available so as to form the weighted averages. If the state vector contains N elements, then the model, Eq. (3.18), requires multiplying an N-dimensional vector by an $N \times N$ matrix at each time step. The covariance update (3.19) requires updating each of N columns of $\mathbf{P}(t)$ in the same way, and then doing it again [i.e., in practice, one forms $\mathbf{AP}(t)$, transposes it, and forms $\mathbf{A}\big(\mathbf{AP}(t)\big)^T$, equivalent to running the model $2N$ times at each time step]. In many applications, this covariance update step dominates the calculation and renders it impractical.

The Kalman filter is nonetheless of crucial importance. Someday we will need to make forecasts (some limited forecasting of El Niño is already attempted; navies forecast acoustic features for antisubmarine warfare maneuvers; storm surge forecasting,[21] etc.). That the Kalman filter is the desired result is of great theoretical importance, and it becomes a guide in understanding

Perigaud (1991), and Gaspar and Wunsch (1989) discuss oceanographic applications; and Ghil and Malanotte-Rizzoli (1991) and Bennett (1992) provide reviews in the wider field of geophysical fluid dynamics.

[21] Heemink and Kloosterhuis (1990).

potentially more practical, suboptimal methods. Furthermore, it is also a central element of most algorithms that do solve the problem of using all the data, as well as their suboptimal approximations.

The Kalman filter was derived heuristically as a simple generalization of the ideas used in Chapter 2. Unsurprisingly, the static inverse results are readily obtained from the filter in various limits. As one example, consider the nearly noise-free case in which both process and observation noise are very small, i.e. $\|\mathbf{Q}\|, \|\mathbf{R}\| \to 0$. Then if $\mathbf{P}(t+1,-)$ is nearly diagonal, $\mathbf{P}(t+1,-) \sim \delta^2\mathbf{I}$,

$$\mathbf{K}(t+1) \longrightarrow \mathbf{E}^T\left(\mathbf{E}\mathbf{E}^T\right)^{-1},$$

assuming existence of the inverse and,

$$\begin{aligned}\tilde{\mathbf{x}}(t) \sim\ &\mathbf{A}\tilde{\mathbf{x}}(t-1) + \mathbf{B}\mathbf{q}(t-1)\\ &+ \mathbf{E}^T(\mathbf{E}\mathbf{E}^T)^{-1}\Big\{\mathbf{y}(t) - \mathbf{E}\big[\mathbf{A}\tilde{\mathbf{x}}(t-1) + \mathbf{B}\mathbf{q}(t-1)\big]\Big\}\\ =\ &\mathbf{E}^T(\mathbf{E}\mathbf{E}^T)^{-1}\mathbf{y}(t)\\ &+ \left[\mathbf{I} - \mathbf{E}^T\left(\mathbf{E}\mathbf{E}^T\right)^{-1}\mathbf{E}\right]\left[\mathbf{A}\tilde{\mathbf{x}}(t-1) + \mathbf{B}\mathbf{q}(t-1)\right].\end{aligned} \tag{3.25}$$

$\mathbf{E}^T\left(\mathbf{E}\mathbf{E}^T\right)^{-1}\mathbf{y}(t)$ is just the expression in Ch 2. for the direct estimate of $\mathbf{x}(t)$ from a set of full-rank, underdetermined, noise-free observations. It is the static estimate we would use at time t if no dynamics were available to permit the use of prior observations. Equation 3.25 will be recognized as identical to Ch 2. Eq. (8.18) from the recursive least-squares solution. The only difference is that here the dynamical evolution equation transmits information about previous observations to help improve the static estimate. The columns of $\mathbf{I} - \mathbf{E}^T\left(\mathbf{E}\mathbf{E}^T\right)^{-1}\mathbf{E}$ are the nullspace of $\mathbf{E}$ and (3.25) thus employs only those elements of the forecast lying in the nullspace of the observations—a sensible result given that the observations here produce perfect estimates of components of $\mathbf{x}(t+1)$ in the range of $\mathbf{E}$. Thus, in this particular limit, the Kalman filter computes from the noise-free observations those elements of $\mathbf{x}(t+1)$ that it can, and for those which cannot be so determined, it forecasts them from the dynamics. The reader ought to examine other limiting cases–retaining process and/or observational noise–including the behavior of the error covariance propagation.

A number of more general reformulations of the equations into algebraically equivalent forms are particularly important. In one form, one works not with the covariances, $\mathbf{P}(t+1,-), \ldots$, but with their inverses, the so-called information matrices, $\mathbf{P}(t+1,-)^{-1}$, etc. This *information filter* form may be more efficient if, for example, the information matrices are banded and sparse while the covariance matrices are not. In another formulation, one uses the square roots of the covariance matrices rather than the matrices themselves. The *square root filter* can be of great importance as there is a tendency for the computation of the updated values of $\mathbf{P}$ to become nonpositive-definite owing to round-off errors and other problems. The square root formulation guarantees positive definite covariances[22].

Example: It is interesting to apply some of these expressions to the simple problem of finding the mean of a set of observations. The model is of an unchanging scalar mean,

$$x(t) = x(t-1)$$

observed in the presence of noise,

$$y(t) = x(t) + n(t)$$

where $\langle n(t)^2 \rangle = R$, so $E = 1$, $A = 1$. In contrast to the situation in Chapter 2, the machinery we have developed here requires that the noise be uncorrelated: $\langle n(t)n(t') \rangle = 0$, $t \neq t'$, although as already mentioned, methods exist to overcome this restriction. Suppose that the initial estimate of the mean is 0—that is, $\tilde{x}(0) = 0$, with uncertainty $P(0)$. Then (3.19) is $P(t+1,-) = P(t)$, and the Kalman filter uncertainty, in the form (3.23), is

$$\frac{1}{P(t+1)} = \frac{1}{P(t)} + \frac{1}{R},$$

a difference equation, with known initial condition, whose solution by inspection is

$$\frac{1}{P(t)} = \frac{t}{R} + \frac{1}{P(0)}.$$

Using (3.20) with $\mathbf{E} = 1$, and successively stepping forward (Bryson & Ho, 1975, p. 363, or Brown, 1983, p. 218) produces

$$\tilde{x}(t) = \frac{R}{R + tP(0)} \left\{ \frac{P(0)}{R} \sum_{j=1}^{t} y(j) \right\}, \tag{3.26}$$

whose limit as $t \rightarrow \infty$ is

$$\tilde{x}(t) \longrightarrow \frac{1}{t} \sum_{j=1}^{t} y(j),$$

the simple average, with uncertainty $P(t) \rightarrow 0$, as $t \rightarrow \infty$. If there is no useful estimate available of $P(0)$, rewrite Eq. (3.26) as,

$$\tilde{x}(t) = \frac{R}{R/P(0) + t} \left\{ \frac{1}{R} \sum_{j=1}^{t} y(j) \right\}, \tag{3.27}$$

and take the agnostic limit, $1/P(0) \rightarrow 0$, or

$$\tilde{x}(t) = \frac{1}{t} \sum_{j=1}^{t} y(j), \tag{3.28}$$

which is again wholly conventional.

[22] Anderson and Moore (1979) discuss these and other variants of the Kalman filter equations.

The Kalman filter permits one to make an optimal forecast from a linear model, subject to the accuracy of the various assumptions being made. When data are available, the state-averaging step produces the best estimate at that time. In between the times when data are available, the state estimate evolves according to the governing dynamics in the model. The system state trajectory thus follows a dynamically consistent pathway in between observations. At the time of observation however, the state trajectory jumps, perhaps dramatically, owing to the employment of the data. These jumps are not normally consistent with the model dynamics, and the resulting estimates, $\tilde{\mathbf{x}}\,(t)$, can be problematic for dynamical understanding of how the physics are evolving. To obtain a consistent state trajectory, we need to solve the problem as originally stated.

3.3. The Smoothing Problem. Minimization of J in Eq. (3.9) subject to the model is still the goal. Begin the discussion by again considering a one-step process.[23] Consider a problem where there are only two time steps involved, $t = 0,\ 1$. There is an initial estimate $\tilde{\mathbf{x}}(0)$, $\tilde{\mathbf{u}}(0) \equiv 0$ with uncertainties $\mathbf{P}(0)$, $\mathbf{Q}(0)$ for the initial state and control vectors, respectively, a set of measurements at time-step 1, and the model. A suitable objective function is,

$$\begin{aligned} J = {} & \big(\tilde{\mathbf{x}}(0,+) - \tilde{\mathbf{x}}(0)\big)^T \mathbf{P}(0)^{-1} \big(\tilde{\mathbf{x}}(0,+) - \tilde{\mathbf{x}}(0)\big) \\ & + \big(\tilde{\mathbf{u}}(0,+) - \tilde{\mathbf{u}}(0)\big)^T \mathbf{Q}(0)^{-1} \big(\tilde{\mathbf{u}}(0,+) - \tilde{\mathbf{u}}(0)\big) \\ & + \big(\mathbf{y}(1) - \mathbf{E}(1)\,\tilde{\mathbf{x}}(1)\big)^T \mathbf{R}(1)^{-1} \big(\mathbf{y}(1) - \mathbf{E}(1)\,\tilde{\mathbf{x}}(1)\big)\,, \end{aligned} \tag{3.29}$$

subject to the model

$$\tilde{\mathbf{x}}(1) = \mathbf{A}(0)\,\tilde{\mathbf{x}}(0,+) + \mathbf{B}(0)\mathbf{q}(0) + \Gamma\tilde{\mathbf{u}}(0,+)\,, \tag{3.30}$$

with the weight matrices again chosen as the inverses of the prior covariances. A minimizing solution to this objective function would produce a new estimate of $\mathbf{x}(0)$, denoted $\tilde{\mathbf{x}}(0,+)$, with error covariance $\mathbf{P}(0,+)$; the + denotes use of future observations, $\mathbf{y}(1)$, in the estimate. On the other hand, we would still denote the estimate at $t = 1$ as $\tilde{\mathbf{x}}(1)$, coinciding with the Kalman filter estimate, because only data prior to and at the same time would have been used. The estimate $\tilde{\mathbf{x}}(1)$ must be given by Eq. (3.15), but it remains to improve $\tilde{\mathbf{u}}(0)$, $\tilde{\mathbf{x}}(0)$.

The basic issue can be understood by observing that the initial estimates $\tilde{\mathbf{u}}(0)$, $\tilde{\mathbf{x}}(0)$ (the former is usually taken to be zero) lead to a forecast that disagrees with the final best estimate $\tilde{\mathbf{x}}(1)$. If either of $\tilde{\mathbf{u}}(0)$, $\tilde{\mathbf{x}}(0)$ were known perfectly, the forecast discrepancy could be ascribed to the other one, permitting ready computation of the required value. In practice, both are somewhat uncertain, and the modification must be partitioned between them. One would not be surprised to find that the partitioning proves to be proportional to their initial uncertainty.

[23] Adapted here from Bryson & Ho, 1975, Chapter 13, whose notation is unfortunately somewhat difficult.

To find the stationary point (we will not trouble to prove it a minimum rather than a maximum), set the differential of J with respect to $\tilde{\mathbf{x}}(0,+)$, $\tilde{\mathbf{x}}(1)$, $\tilde{\mathbf{u}}(0,+)$ to zero,

$$\begin{aligned}\frac{dJ}{2} &= d\tilde{\mathbf{x}}(0,+)^T\mathbf{P}(0)^{-1}\left[\tilde{\mathbf{x}}(0,+)-\tilde{\mathbf{x}}(0)\right]\\ &+ d\tilde{\mathbf{u}}(0,+)^T\mathbf{Q}(0)^{-1}\left[\tilde{\mathbf{u}}(0,+)-\tilde{\mathbf{u}}(0)\right]\\ &- d\tilde{\mathbf{x}}(1)^T\mathbf{E}(1)^T\mathbf{R}(1)^{-1}\left[\mathbf{y}(1)-\mathbf{E}(1)\tilde{\mathbf{x}}(1)\right]=0\,.\end{aligned} \tag{3.31}$$

The coefficients of the differentials cannot be set to zero separately because they are connected via the model, Equation (3.30), which provides a relationship

$$d\tilde{\mathbf{x}}(1)=\mathbf{A}(0)\,d\tilde{\mathbf{x}}(0,+)+\Gamma(0)\,d\tilde{\mathbf{u}}(0,+)\,. \tag{3.32}$$

Eliminating $d\tilde{\mathbf{x}}(1)$,

$$\begin{aligned}\frac{dJ}{2} = d\tilde{\mathbf{x}}(0,+)^T&\Big[\mathbf{P}(0)^{-1}\big(\tilde{\mathbf{x}}(0,+)-\tilde{\mathbf{x}}(0)\big)\\ &\quad-\mathbf{A}(0)^T\mathbf{E}(1)^T\mathbf{R}(1)^{-1}\big(\mathbf{y}(1)-\mathbf{E}(1)\tilde{\mathbf{x}}(1)\Big]\\ +\ d\tilde{\mathbf{u}}(0,+)^T&\Big[\mathbf{Q}(0)^{-1}\big(\tilde{\mathbf{u}}(0,+)-\tilde{\mathbf{u}}(0)\big)\\ &\quad+\Gamma^T(0)\mathbf{E}(1)^T\mathbf{R}(1)^{-1}\big(\mathbf{y}(1)-\mathbf{E}(1)\tilde{\mathbf{x}}(1)\big)\Big]\,.\end{aligned} \tag{3.33}$$

Now the differential vanishes, producing a stationary value of J only if the coefficients of $d\tilde{\mathbf{x}}(0,+)$, $d\tilde{\mathbf{u}}(0,+)$ separately vanish, yielding

$$\tilde{\mathbf{x}}(0,+)=\tilde{\mathbf{x}}(0)+\mathbf{P}(0)\mathbf{A}(0)^T\mathbf{E}(1)^T\mathbf{R}(1)^{-1}\big(\mathbf{y}(1)-\mathbf{E}(1)\tilde{\mathbf{x}}(1)\big) \tag{3.34}$$

$$\tilde{\mathbf{u}}(0,+)=\tilde{\mathbf{u}}(0)+\mathbf{Q}(0)\Gamma(0)^T\mathbf{E}(1)^T\mathbf{R}(1)^{-1}\big(\mathbf{y}(1)-\ \mathbf{E}(1)\tilde{\mathbf{x}}(1)\big) \tag{3.35}$$

and

$$\begin{aligned}\tilde{\mathbf{x}}(1)&=\tilde{\mathbf{x}}(1,-)\\ &\quad+\mathbf{P}(1,-)\mathbf{E}(1)^T\left[\mathbf{E}(1)\mathbf{P}(1,-)\mathbf{E}(1)^T+\mathbf{R}(1)\right]^{-1}\big(\mathbf{y}(1)-\mathbf{E}(1)\tilde{\mathbf{x}}(1,-)\big)\end{aligned} \tag{3.36}$$

using the previous definitions of $\tilde{\mathbf{x}}(1,-)$, $\mathbf{P}(1,-)$, which is recognized as the Kalman filter estimate. At this point we are essentially done: An estimate has been produced not only of $\mathbf{x}(1)$, but an improvement has been made in the prior estimate of $\mathbf{x}(0)$ using the future measurements, and we have estimated the control term. Notice the corrections to $\tilde{\mathbf{u}}(0)$, $\tilde{\mathbf{x}}(0)$ are proportional to $\mathbf{P}(0)$, $\mathbf{Q}(0)$, respectively, as anticipated. We still need to examine the uncertainties of these latter quantities.

First rewrite the estimates (3.34)–(3.35) as

$$\begin{aligned}\tilde{\mathbf{x}}(0,+)&=\tilde{\mathbf{x}}(0)+\mathbf{L}(1)\,\big(\tilde{\mathbf{x}}(1)-\tilde{\mathbf{x}}(1,-)\big)\,,\quad \mathbf{L}(1)=\mathbf{P}(0)\mathbf{A}(0)^T\mathbf{P}(1,-)^{-1}\,,\\ \tilde{\mathbf{u}}(0,+)&=\tilde{\mathbf{u}}(0)+\mathbf{M}(1)\,\big(\tilde{\mathbf{x}}(1)-\tilde{\mathbf{x}}(1,-)\big)\,,\quad \mathbf{M}(1)=\mathbf{Q}(0)\Gamma(0)^T\mathbf{P}(1,-)^{-1}\,,\end{aligned} \tag{3.37}$$

which can be done by extended, but uninteresting, algebraic manipulation. The importance of these latter two expressions is that both $\tilde{\mathbf{x}}(0,+)$, $\tilde{\mathbf{u}}(0,+)$ are expressed in terms of their prior estimates in a weighted average with the difference between the prediction of the state at $t=1$, $\tilde{\mathbf{x}}(1,-)$ and what was actually estimated there following the data use, $\tilde{\mathbf{x}}(1)$. [But the data do not appear explicitly in (3.37).] Then it is also possible to show that

$$\begin{aligned}\mathbf{P}(0,+) &= \mathbf{P}(0) + \mathbf{L}(1)\big(\mathbf{P}(1) - \mathbf{P}(1,-)\big)\mathbf{L}(1)^T \\ \mathbf{Q}(0,+) &= \mathbf{Q}(0) + \mathbf{M}(1)\big(\mathbf{P}(1) - \mathbf{P}(1,-)\big)\mathbf{M}(1)^T\,.\end{aligned} \tag{3.38}$$

Based upon this one-step derivation, a complete recursion for any time interval can be inferred. Suppose the Kalman filter has been run all the way to a terminal time t_f. The result is $\tilde{\mathbf{x}}(t_f)$ and its variance $\mathbf{P}(t_f)$. With no future data available, $\tilde{\mathbf{x}}(t_f)$ cannot be further improved. At time t_f-1, we have an estimate $\tilde{\mathbf{x}}(t_f-1)$ with uncertainty $\mathbf{P}(t_f-1)$, which could be improved by knowledge of the future observations at t_f. But this situation is precisely the one addressed by the objective function Eq. (3.29) with t_f replacing $t=1$, and t_f-1 replacing $t=0$. Now having improved the estimate at t_f-1 and calling it $\tilde{\mathbf{x}}(t_f-1,+)$ with uncertainty $\mathbf{P}(t_f-1,+)$, this new estimate is used to improve the prior estimate $\tilde{\mathbf{x}}(t_f-2)$, and we step all the way back to $t=0$. The complete recursion is

$$\begin{aligned}\tilde{\mathbf{x}}(t,+) &= \tilde{\mathbf{x}}(t) + \mathbf{L}(t+1)\left[\tilde{\mathbf{x}}(t+1,+) - \tilde{\mathbf{x}}(t+1,-)\right], \\ \mathbf{L}(t+1) &= \mathbf{P}(t)\mathbf{A}(t)^T\,\mathbf{P}(t+1,-)^{-1}\end{aligned} \tag{3.39}$$

$$\begin{aligned}\tilde{\mathbf{u}}(t,+) &= \tilde{\mathbf{u}}(t) + \mathbf{M}(t+1)\left[\tilde{\mathbf{x}}(t+1,+) - \tilde{\mathbf{x}}(t+1,-)\right], \\ \mathbf{M}(t+1) &= \mathbf{Q}(t)\Gamma(t)^T\mathbf{P}(t+1,-)^{-1}\,,\end{aligned} \tag{3.40}$$

$$\mathbf{P}(t,+)=\mathbf{P}(t)+\mathbf{L}(t+1)\big[\mathbf{P}(t+1,+)-\mathbf{P}(t+1,-)\big]\mathbf{L}(t+1)^T, \tag{3.41}$$

$$\mathbf{Q}(t,+)=\mathbf{Q}(t)+\mathbf{M}(t+1)\big[\mathbf{P}(t+1,+)-\mathbf{P}(t+1,-)\big]\mathbf{M}(t+1)^T, \tag{3.42}$$

with $$\tilde{\mathbf{x}}(t_f,+) \equiv \tilde{\mathbf{x}}(t_f),\ \mathbf{P}(t_f,+) \equiv \mathbf{P}(t_f)\,.$$

This recipe, which uses the Kalman filter on a first forward sweep to the end of the available data, and which then successively improves the prior estimates by sweeping backwards, is called the *RTS algorithm*, for Rauch, Tung, and Striebel (1965). The particular form has the advantage that the data are not involved in the backward sweep, since all the available information has been used in the filter calculation. It does have the potential disadvantage of requiring the storage at each time step of $\mathbf{P}(t)$, and the inversion of $\mathbf{P}(t,-)$. $\mathbf{P}(t,-)$ is readily recomputed from (3.19) and need not be stored. By direct analogy with the one-step objective function, the recursion Eqs. (3.39)–(3.42) is seen to be the solution to the minimization of the objective function Eq. (3.30) subject to the model. Most important, assuming consistency of all assumptions, the resulting state vector trajectory $\tilde{\mathbf{x}}\,(t,+)$ now satisfies the model which was assumed to apply at all times, and along with the control vector estimate $\tilde{\mathbf{u}}\,(t)$ lends itself to scientific analysis.

It is possible to examine limiting cases of the RTS smoother much as with the Kalman filter. For example, suppose again that $\mathbf{Q}$ vanishes, and $\mathbf{A}^{-1}$ exists. Then,

$$\mathbf{L}(t+1) \longrightarrow \mathbf{P}(t)\mathbf{A}^T\left(\mathbf{A}\mathbf{P}(t)\mathbf{A}^T\right)^{-1} = \mathbf{A}^{-1} \tag{3.43}$$

for diagonal $\mathbf{P}(t)$, and Equation (3.39) becomes

$$\tilde{\mathbf{x}}(t,+) \longrightarrow \mathbf{A}^{-1}\left[\tilde{\mathbf{x}}(t+1,+) - \mathbf{B}\mathbf{q}(t)\right], \tag{3.44}$$

a sensible backward estimate obtained by simply solving

$$\tilde{\mathbf{x}}(t+1) = \mathbf{A}\tilde{\mathbf{x}}(t) + \mathbf{B}\mathbf{q}(t) \tag{3.45}$$

for $\tilde{\mathbf{x}}(t)$. Other limits are also illuminating but are left to the reader.

Example: The Straight Line. The smoother result for the straight-line model (2.10) is shown in Figures 3 and 4 for both forms of state vector. The time-evolving estimate is now a nearly perfect straight line, whose uncertainty (e.g., Figure 3) has a terminal value at $t = 100$ equal to that for the Kalman filter estimate, as it must, and reaches a minimum near the middle of the estimation period, before growing again toward $t = 0$, where the initial uncertainty was very large. In the case where the state vector consisted of the constant intercept and slope of the line, both smoothed estimates are seen, in contrast to the filter estimate, to conform very well to the known true behavior. Again, the system has more information available to determine the slope value than it does the intercept. It should be apparent that the best-fitting, straight-line solution of Chapter 2 is also the solution to the smoothing problem, but with the data and model handled all at once, a whole-domain method, rather than sequentially.

Example: The Mass-Spring Oscillator. Figure 5h shows the state estimate for the mass-spring oscillator made from a smoothing computation run backward from $t = 150$, and its variance is shown in Figure 5i. On average (but not everywhere), the smoothed estimate is closer to the correct value than is the filtered estimate, as expected. The standard error is also smaller for the smoothed estimate. Figure 5j displays the variance, $\mathbf{Q}_{11}(t)$, of the estimate one can make of the scalar control variable $\mathbf{u}(\mathbf{t})$. $\tilde{\mathbf{u}}(\mathbf{t})$ is not shown because the actual value was white noise, and the plot is uninteresting [estimate and truth do agree consistent with $\mathbf{Q}_{11}(t)$].

Example: Tracer Problem. Consider a problem stimulated by the need to extract information from transient tracers, C, in a fluid, which are assumed to satisfy an equation like

$$\frac{\partial C}{\partial t} + \mathbf{v}\cdot\nabla C - \kappa\nabla^2 C = -\lambda C + q(\mathbf{r},t) \tag{3.46}$$

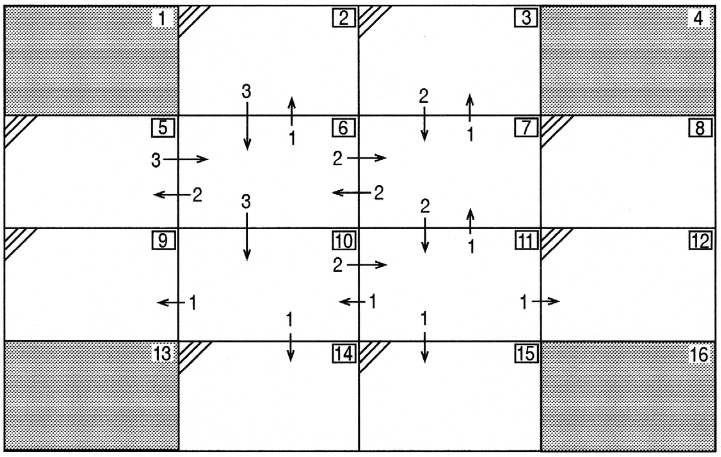


FIGURE 7. Tracer box model where J_{ij} represent fluxes between boxes and are chosen to be mass conserving. Boxes with shaded corners are boundary boxes with externally prescribed concentrations. Numbers in upper right corner are used to identify the boxes.

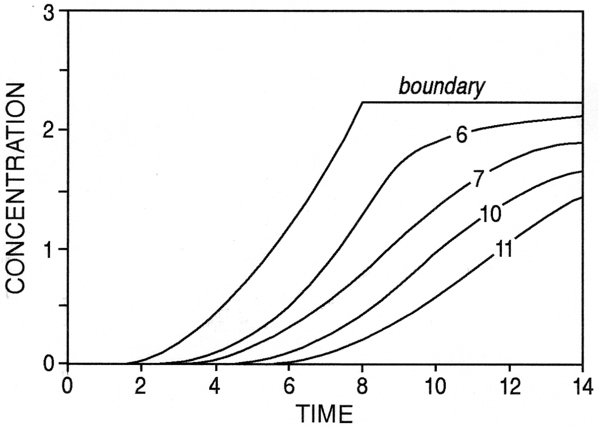


FIGURE 8. Time histories of the forward computation in which boundary concentrations shown were prescribed, and values computed for boxes 6,7,10,11. These represent the "truth".

where q represents sources/sinks and λ is a decay rate if the tracer is radioactive. To have a simple model that will capture the structure of this problem, the fluid is divided into a set of boxes as depicted in Figure 7. The flow field, as depicted there, is represented by exchanges between boxes given by the $J_{ij} \geq 0$. That is, the J_{ij} are just a simplified representation of the effects of advection and mixing on a dye C (the relationship between such simple parameterizations and more formal and elaborate finite-difference schemes. Here, it will only be remarked that J_{ij} are

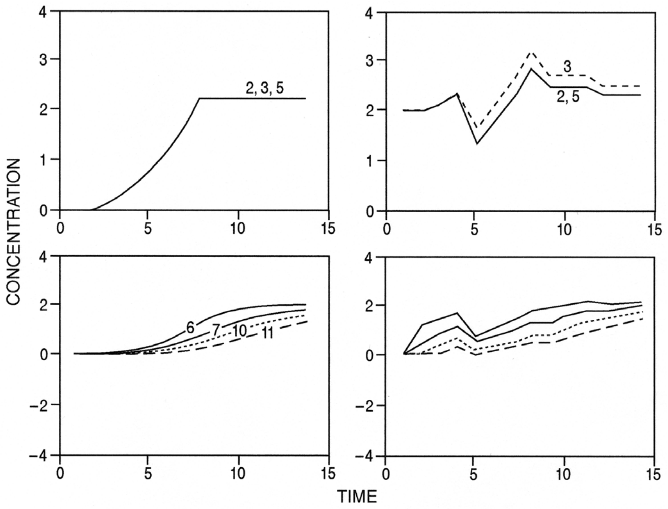


FIGURE 9. The left two panels show the correct values for boundary boxes (upper panel) and interior ones (lower panel). The right panels show the results of the Kalman filter for boundary (upper panel) and interior boxes (lower panel) when noisy observations were provided at $t = 5, 9, 12$. At the observation times, estimates are pulled toward the observations but tend to diverge afterward. By the time the last observations are used, estimated and correct values are quite close. Although not displayed here, there is an uncertainty estimate at all times.

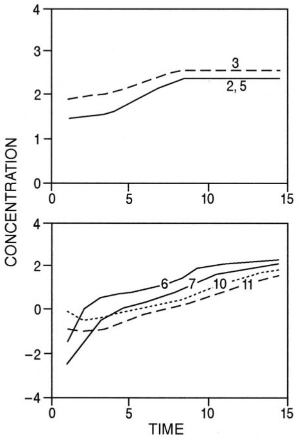


FIGURE 10. Smoothed estimates of boundary (left panel) and interior (right panel) values of tracers.

chosen to be mass conserving so that the sum over all J_{ij} entering and leaving a box vanishes.

The discrete analogue of (3.46) is taken to be

$$\begin{aligned}C_i(t+1) &= C_i(t) - \lambda\Delta t C_i(t)\\ &\quad - \frac{\Delta t}{V}\sum_{j\in N(i)} C_i(t)J_{ij} + \frac{\Delta t}{V}\sum_{j\in N(i)} C_j(t)J_{ji}\end{aligned} \tag{3.47}$$

where the notation $j \in N(i)$ denotes an index sum over the neighboring boxes to box i, V is a volume for the boxes, and Δt is the time step. This model can easily be put into the canonical form

$$\mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{q}(t) + \Gamma\mathbf{u}(t)\,, \qquad \mathbf{Q} = \mathbf{0}\,, \tag{3.48}$$

with the state vector being the vector of box concentrations $\mathbf{x}\,(t) = [[C_i(t), C_i(t-1)]]^T$.

A forward computation was run with initial concentrations everywhere of 0, using the boundary conditions depicted in Figure 8, resulting in interior box values as shown. Based upon these correct values, noisy "observations" of the interior boxes only were constructed at times $t = 5$, 9, 12.

An initial estimate of tracer concentrations at $t = 0$ was taken (correctly) to be zero, but this estimate was given a large variance [diagonal $\mathbf{P}(0)$ with large norm]. The boundary box concentrations were set erroneously to $C = 2$ for all t and held at that value. A Kalman filter computation was run as shown in Figure 9. Initially, the interior box concentration estimates rise erroneously (owing to the dye leaking in from the high nonzero concentrations in the boundary boxes). At $t = 5$, the first set of observations becomes available, and the combined estimate is driven much closer to the true values. By the time the last set of observations is used, the estimated and correct concentrations are quite close, although the time history of the interior is somewhat in error. The RTS algorithm was then applied to generate the smoothed histories shown in Figure 10 and to estimate the boundary concentrations (the controls). As expected, the smoothed estimates are closer to the true time history than are the filtered ones (the uncertainty estimates are not shown, but the results are consistent with the "truth" within statistical expectation). Unless further information is provided, no other estimation procedure could do better, given that the model is the correct one.

Example: Altimetry. Consider Equation (2.2). Write the solution to the equation as in (2.3) so that the state vector consisted of the expansion coefficients,[24]

$$\mathbf{x}(t) = \begin{bmatrix}a_1(t) & a_2(t) & \cdots & a_{2N-1}(t) & a_{2N}(t)\end{bmatrix}^T .$$

As is usually true in real, as opposed to textbook, problems, most of the effort lay in specifying the covariances $\mathbf{P}(0)$, $\mathbf{R}$, $\mathbf{Q}$. Figure 11 shows an estimate of the initial values, $\tilde{\mathbf{x}}(0,+)$ from six months of data. Figure 12 displays the filtered and smoothed estimates of some of the wave amplitudes and phases over the observational duration. The smoothed estimates are indeed much

[24] Gaspar and Wunsch (1996).

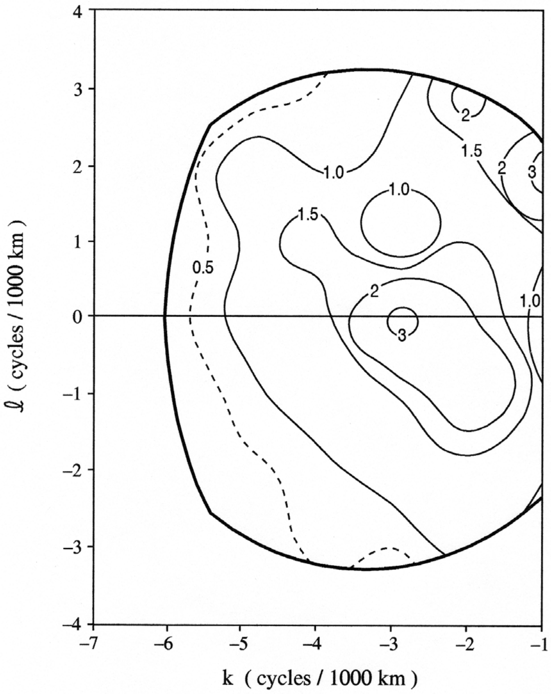


FIGURE 11. Estimates of the initial values of the seasurface elevation model amplitudes from the RTS smoother, contoured in wavenumber space.

smoother than those from the filter, as one would expect from what is basically an interpolation computation as opposed to the extrapolation of the pure filter.

There are many versions of smoother algorithms chosen for special purposes (e.g., successively improving an estimate at a fixed time as data continue to accumulate—the so-called fixed-point smoother—or to achieve various trade-offs of computations versus storage requirements). Consider one other approach to smoothing. Suppose the Kalman filter has been run forward to some time t_c, producing an estimate $\tilde{\mathbf{x}}(t_c)$ with uncertainty $\mathbf{P}(t_c)$. Now suppose, perhaps on the basis of some further observations, that at a *later* time t_f an independent estimate $\tilde{\mathbf{x}}(t_f)$ has been made, with uncertainty $\mathbf{P}(t_f)$. The independence is crucial—we suppose this latter estimate is made without using any observations at time t_c or earlier so that any errors in $\tilde{\mathbf{x}}(t_c)$ and $\tilde{\mathbf{x}}(t_f)$ are uncorrelated.

Let us run the model *backward* in time from t_f to $t_f - 1$:

$$\tilde{\mathbf{x}}_b(t_f - 1) = \mathbf{A}^{-1}\tilde{\mathbf{x}}(t_f) - \mathbf{A}^{-1}\mathbf{B}\mathbf{q}(t_f - 1) \tag{3.49}$$

where the subscript b denotes a backward-in-time estimate. The reader may object that running a model backward in time will often be an unstable operation, and this objection needs to be

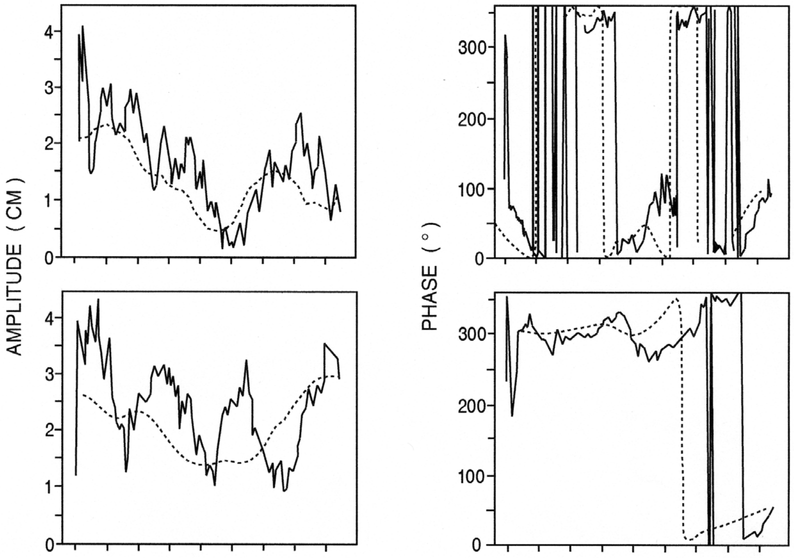


FIGURE 12. Time histories from the Kalman filter (solid lines) and RTS smoother (dashed) of two typical modal amplitudes (left panels), and phases (right panels) from six months of real altimetric data.

addressed, but ignore it for the moment. The uncertainty of $\tilde{\mathbf{x}}(t_f-1)$ is

$$\mathbf{P}_b(t_f-1) = \mathbf{A}^{-1}\mathbf{P}(t_f)\mathbf{A}^{-T} + \mathbf{A}^{-1}\Gamma\mathbf{Q}(t_f-1)\Gamma^T\mathbf{A}^{-T} \tag{3.50}$$

in direct analogy to the forward model computation. This backward computation can be continued to time t_c, at which point we will have an estimate, $\tilde{\mathbf{x}}_b(t_c)$, with uncertainty $\mathbf{P}_b(t_c)$.

The two independent estimates of $\mathbf{x}(t_c)$ can be combined to make an improved estimate using the relations Chapter 2, Eq. (8.22),

$$\tilde{\mathbf{x}}(t_c,+) = \tilde{\mathbf{x}}(t_c) + \mathbf{P}(t_c)\big(\mathbf{P}(t_c)+\mathbf{P}_b(t_c)\big)^{-1}\big(\tilde{\mathbf{x}}_b(t_c)-\tilde{\mathbf{x}}(t_c)\big) \tag{3.51}$$

and

$$\begin{aligned}\mathbf{P}(t_c) &= \Big\langle \big(\tilde{\mathbf{x}}(t_c,+)-\mathbf{x}(t_c)\big)\,\big(\tilde{\mathbf{x}}(t_c,+)-\mathbf{x}(t_c)\big)^T \Big\rangle \\ &= \big(\mathbf{P}(t_c)^{-1}+\mathbf{P}_b(t_c)^{-1}\big)^{-1}.\end{aligned} \tag{3.52}$$

This estimate is the same as obtained from the RTS algorithm because the same objective function, model, and data have been employed.

Running the model backward may indeed be unstable if it contains any dissipative terms. A forward model may be unstable too, if there are unstable modes of motion, either real or numerical artifacts. But the expressions in Eqs. (3.51), (3.52) are stable, because the computation of $\mathbf{P}_b(t)$ and its use in the updating expression (3.52) automatically downweight unstable elements whose errors will be very large, which will not carry useful information about the earlier state.

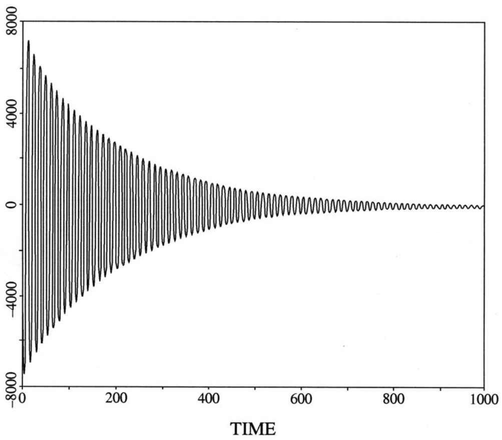


FIGURE 13. Mass-spring oscillator model with friction ($r = 0.01$) run backwards in time from conditions specified at $t = 1000$.The system is unstable, and small uncertainties in the starting conditions would amplify. but the Kalman filter run backwards remains stable because its error estimate grows too, systematically downweighting the model forecast relative to any data that become available at earlier times. A model with unstable elements in the forward direction would behave analogously when integrated in time with growing estimated model forecast error.

The same situation would occur if the forward model had unstable elements–these instabilities would amplify slight errors in the statement of their initial conditions, rendering them difficult to estimate from observations at later times (see Figure 13). Examination of the covariance propagation equation and the filter gain matrix shows that these elements are suppressed in the Kalman filter estimate, with correspondingly large uncertainties. Thus, the filter/smoother formalism properly accounts for unstable, and hence difficult-to-calculate, parameters by estimating their uncertainty as very large, thus handling very general ill-conditioning. In practice, one needs to be careful, for numerical reasons, of the pitfalls in computing and using matrices that may have norms growing exponentially in time. But the conceptual problem is solved.

As with the Kalman filter, it is possible to rewrite the RTS smoother expressions (3.39)–(3.42) in various ways for computational efficiency, storage reduction, and improved accuracy.[25]

The dominant computational load in the smoother is again the calculation of the updated covariance matrices, whose size is square of the state-vector dimension, at every time step, leading to efforts to construct simplified algorithms that retain most of the virtues of the filter/smoother combination but with reduced load. For example, it may have already occurred to the reader that in some of the examples displayed, the state vector uncertainties, $\mathbf{P}$, in both the

[25] Gelb, 1974; Bryson & Ho, 1975; Anderson & Moore, 1979; Goodwin & Sin, 1984; Sorenson, 1985

filter and the smoother appear to approach rapidly a steady state. This asymptotic behavior in turn means that the gain matrices, $\mathbf{K}$, $\mathbf{L}$, $\mathbf{M}$ will also achieve a steady state, implying that one no longer needs to undertake the updating steps–fixed gains can be used. Such steady-state operators are known as *Wiener filters* and *smoothers* and if achieved, they represent a potentially very large computational savings. One needs to understand the circumstances under which such steady states can be expected to appear, and we will examine the problem in Section xx.

We turn now instead to another approach to reducing the computational load–the so-called adjoint methods. We will demonstrate how they work and their equivalence to smoothing.

4. Control Problems: The Pontryagin Principle and Adjoint Methods

4.1. Lagrange Multiplier Constraints. The results of the last section are recursive schemes for computing a filtered and then a smoothed estimate. As with recursive least squares, the necessity to combine two pieces of information to make an improved estimate demands knowledge of the uncertainty of the information.

Because the covariance computation will usually dominate and potentially overwhelm the filter/smoother algorithms, it is very attractive, at least superficially, to find algorithms that do not require the covariances—that is, which employ the entire time domain of observations simultaneously, a whole-domain method. The algorithms that emerge are best known in the context of *control theory.* Essentially, there is a more specific focus upon determining the $\mathbf{u}(t)$: the control variables. In many problems, one literally wishes to drive a system in desirable ways. Although oceanographers and meteorologists can drive their models in ways analogous to controlling a machine tool, they do not ever control the real ocean or atmosphere. It will help the reader who further explores these methods to recognize that we are still doing *estimation*, combining observations and models, but sometimes using algorithms best known under the control rubric.

To see the possibilities, consider again a two-point objective function

$$\begin{aligned} J = &\left(\tilde{\mathbf{x}}(0,+) - \tilde{\mathbf{x}}(0)\right)\mathbf{P}(0)^{-1}\left(\tilde{\mathbf{x}}(0,+) - \tilde{\mathbf{x}}(0)\right) \\ &+ \left(\tilde{\mathbf{u}}(0,+) - \tilde{\mathbf{u}}(0)\right)^T\mathbf{Q}(0)^{-1}\left(\tilde{\mathbf{u}}(0,+) - \tilde{\mathbf{u}}(0)\right) \\ &+ \left(\mathbf{y}(1) - \mathbf{E}(1)\tilde{\mathbf{x}}(1)\right)^T\mathbf{R}(1)^{-1}\left(\mathbf{y}(1) - \mathbf{E}(1)\tilde{\mathbf{x}}(1)\right) \\ &- 2\boldsymbol{\mu}(1)^T\left(\tilde{\mathbf{x}}(1) - \mathbf{A}\tilde{\mathbf{x}}(0,+) - \mathbf{B}\mathbf{q}(0) - \Gamma\tilde{\mathbf{u}}(0,+)\right). \end{aligned} \tag{4.1}$$

where $\mathbf{P}$, etc., are just weight matrices without necessarily having a statistical significance. We wish to find the minimum of J subject to (3.30). To use a variant method, append the model equations as done in Chapter 2 [compare with Ch. 2 Eq. (4.59)], with a vector of Lagrange multipliers $\mu(1)$, for a new objective function As with the filter and smoother, the model is being imposed as a hard constraint, but with the control term rendering the result indistinguishable from a soft one. In contrast to the approach in the last section, the presence of the Lagrange multiplier permits treating the differentials as independent; taking the derivatives of J with respect to $\tilde{\mathbf{x}}(0,+)$, $\tilde{\mathbf{x}}(1)$, $\tilde{\mathbf{u}}(0,+)$, $\boldsymbol{\mu}(1)$ and setting them to zero,

$$\mathbf{P}(0)^{-1}\big(\tilde{\mathbf{x}}(0,+) - \tilde{\mathbf{x}}(0)\big) + \mathbf{A}^T\boldsymbol{\mu}(1) = 0 \tag{4.2}$$

$$\mathbf{E}^T\mathbf{R}(1)^{-1}\big(\mathbf{y}(1) - \mathbf{E}\tilde{\mathbf{x}}(1)\big) + \boldsymbol{\mu}(1) = 0 \tag{4.3}$$

$$\mathbf{Q}(0)^{-1}\big(\tilde{\mathbf{u}}(0,+) - \tilde{\mathbf{u}}(0)\big) + \Gamma^T\boldsymbol{\mu}(1) = 0 \tag{4.4}$$

$$\tilde{\mathbf{x}}(1) - \mathbf{A}\tilde{\mathbf{x}}(0,+) - \mathbf{B}\mathbf{q}(0) - \Gamma\tilde{\mathbf{u}}(0,+) = 0\,. \tag{4.5}$$

Because the objective function in (4.1) is identical with that used with the smoother, and because the identical dynamical model has been imposed [in one case through the differentials in (3.32), in the other explicitly through the Lagrange multipliers], Equations (4.2)–(4.5) must produce the identical solution to that produced by the smoother. A demonstration that Equations (4.2)–(4.5) can be manipulated into the form (3.37) is an exercise in matrix identities.[26] As with smoothing algorithms, the finding of the solution of (4.2)–(4.5) can be arranged in a number of different ways, trading computation against storage, coding ease, convenience, etc.

Let us show explicitly the identity of smoother and adjoint solution for a restricted case—that for which the initial conditions are known exactly, so that $\tilde{\mathbf{x}}(0)$ is not modified by the later observations. For the one-term smoother, the result is obtained by dropping (4.2), as $\mathbf{x}(0)$ is no longer an adjustable parameter. Without further loss of generality, we may put $\tilde{\mathbf{u}}(0) = \mathbf{0}$, and set $\mathbf{R}(1) = \mathbf{R}$, reducing the system to

$$\tilde{\mathbf{x}}(1) = \mathbf{A}\tilde{\mathbf{x}}(0) + \mathbf{B}\mathbf{q}(0) + \Gamma\tilde{\mathbf{u}}(0,+) \tag{4.6}$$

$$\tilde{\mathbf{u}}(0,+) = -\mathbf{Q}(0)\Gamma^T\boldsymbol{\mu}(1) \tag{4.7}$$

$$= \mathbf{Q}(0)\Gamma^T\mathbf{E}^T\mathbf{R}^{-1}\big[\mathbf{y}(1) - \mathbf{E}(1)\tilde{\mathbf{x}}(1)\big]\,. \tag{4.8}$$

Eliminating $\tilde{\mathbf{u}}(0,+)$ from (4.6) produces

$$\tilde{\mathbf{x}}(1) = \mathbf{A}\tilde{\mathbf{x}}(0) + \mathbf{B}\mathbf{q}(0) + \Gamma\mathbf{Q}(0)\Gamma^T\mathbf{E}^T\mathbf{R}^{-1}\big[\mathbf{y}(1) - \mathbf{E}\tilde{\mathbf{x}}(1)\big]\,. \tag{4.9}$$

With no initial error in $\mathbf{x}(0)$, $\mathbf{P}(1,-) = \boldsymbol{\Gamma}\mathbf{Q}(0)\boldsymbol{\Gamma}^T$ and *defining*,

$$\tilde{\mathbf{x}}(1,-) \equiv \mathbf{A}\mathbf{x}(0) + \mathbf{B}\mathbf{q}(0) \tag{4.10}$$

so that (4.9) can be written

$$\big[\mathbf{I} + \mathbf{P}(1,-)\mathbf{E}^T\mathbf{R}^{-1}\mathbf{E}\big]\tilde{\mathbf{x}}(1) = \tilde{\mathbf{x}}(1,-) + \mathbf{P}(1,-)\mathbf{E}^T\mathbf{R}^{-1}\mathbf{y}(1) \tag{4.11}$$

[26] Some guidance is provided by Bryson and Ho (1975, pp. 390–5) or Liebelt (1967). In particular, Bryson and Ho (1975) introduce the Lagrange multipliers (their equations 13.2.7–13.2.8) simply as an intermediate numerical device for solving the smoother equations.

or [factoring $\mathbf{P}(1,-)$]

$$\begin{aligned}\tilde{\mathbf{x}}(1) = &\left\{\mathbf{P}(1,-)^{-1} + \mathbf{E}^T\mathbf{R}^{-1}\mathbf{E}\right\}^{-1}\mathbf{P}(1,-)^{-1}\tilde{\mathbf{x}}(1,-)\\ &+ \left\{\mathbf{P}(1,-)^{-1} + \mathbf{E}^T\mathbf{R}^{-1}\mathbf{E}\right\}^{-1}\mathbf{E}^T\mathbf{R}^{-1}\mathbf{y}(1)\,.\end{aligned}$$

Applying the matrix inversion lemma in the form Ch. 2. Eq. 3.24 to the first term on the right, and in the form Ch. 2, Eq. 3.25 to the second term on the right,

$$\begin{aligned}\tilde{\mathbf{x}}(1) = &\Big\{\mathbf{P}(1,-) - \mathbf{P}(1,-)\mathbf{E}^T\left[\mathbf{E}\mathbf{P}(1,-)\mathbf{E}^T + \mathbf{R}\right]^{-1}\mathbf{E}\mathbf{P}(1,-)\Big\}\\ &\cdot \mathbf{P}(1,-)^{-1}\tilde{\mathbf{x}}(1,-) + \mathbf{P}\mathbf{E}^T\left[\mathbf{R} + \mathbf{E}\mathbf{P}(1,-)\mathbf{E}^T\right]^{-1}\mathbf{y}(1)\end{aligned} \tag{4.12}$$

or

$$\tilde{\mathbf{x}}(1) = \tilde{\mathbf{x}}(1,-) + \mathbf{P}(1,-)\mathbf{E}^T\left[\mathbf{E}\mathbf{P}(1,-)\mathbf{E}^T + \mathbf{R}\right]^{-1}\left(\mathbf{y}(1) - \mathbf{E}\tilde{\mathbf{x}}(1,-)\right). \tag{4.13}$$

This last result is the ordinary Kalman filter estimate, as it must be, but it results here from the adjoint formalism.

We need to consider the adjoint approach for the entire interval $0 \le t \le t_f$. Let us start with the objective function (3.9) and append the model consistency demand using Lagrange multipliers,

$$\begin{aligned}J = &\left(\mathbf{x}(0) - \tilde{\mathbf{x}}(0)\right)^T\mathbf{P}(0)^{-1}\left(\mathbf{x}(0) - \tilde{\mathbf{x}}(0)\right)\\ &+ \sum_{t=1}^{t_f}\left(\mathbf{E}(t)\mathbf{x}(t) - \mathbf{y}(t)\right)^T\mathbf{R}(t)^{-1}\left(\mathbf{E}(t)\mathbf{x}(t) - \mathbf{y}(t)\right)\\ &+ \sum_{t=0}^{t_f-1}\mathbf{u}(t)^T\mathbf{Q}^{-1}\mathbf{u}(t)\\ &- 2\sum_{t=1}^{t_f}\boldsymbol{\mu}(t)^T\left[\mathbf{x}(t) - \mathbf{A}\mathbf{x}(t-1) - \mathbf{B}\mathbf{q}(t-1) - \Gamma\mathbf{u}(t-1)\right].\end{aligned} \tag{4.14}$$

Note the differing lower limits of summation.

Setting all the derivatives to zero gives the normal equations,

$$\tfrac{1}{2}\frac{\partial J}{\partial \mathbf{u}(t)} = \mathbf{Q}^{-1}\mathbf{u}(t-1) + \Gamma^{T}\boldsymbol{\mu}(t) = 0\,, \quad 0 \le t \le t_f - 1 \tag{4.15}$$

$$\tfrac{1}{2}\frac{\partial J}{\partial \boldsymbol{\mu}(t)} = \mathbf{x}(t) - \mathbf{A}\mathbf{x}(t-1) - \mathbf{B}\mathbf{q}(t-1) - \Gamma\mathbf{u}(t-1) = 0, \tag{4.16}$$

$$1 \le t \le t_f$$

$$\tfrac{1}{2}\frac{\partial J}{\partial \mathbf{x}(0)} = \mathbf{P}(0)^{-1}\big(\mathbf{x}(0) - \tilde{\mathbf{x}}(0)\big) + \mathbf{A}^{T}\boldsymbol{\mu}(1) = 0\,, \tag{4.17}$$

$$\frac{1}{2}\frac{\partial J}{\partial \mathbf{x}\,(t)} = \mathbf{E}\,(t)\,\mathbf{R}\,(t)^{-1}\,[\mathbf{E}\,(t)\,\mathbf{x}\,(t) - \mathbf{y}\,(t)] - \boldsymbol{\mu}\,(t) + \mathbf{A}^{T}\boldsymbol{\mu}\,(t+1)\,, \tag{4.18}$$

$$1 \le t \le t_f$$

$$\tfrac{1}{2}\frac{\partial J}{\partial \mathbf{x}(t_f)} = \mathbf{E}(t_f)^{T}\mathbf{R}(t_f)^{-1}\big[\mathbf{E}(t_f)\mathbf{x}(t_f) - \mathbf{y}(t_f)\big] - \boldsymbol{\mu}(t_f) = 0 \tag{4.19}$$

where the derivatives for $\mathbf{x}(t)$, at $t = 0$, $t = t_f$, have been computed separately for clarity. The so-called adjoint model is now given by (4.18). An equation count shows that the number of equations is exactly equal to the number of unknowns $[\mathbf{x}(t),\ \mathbf{u}(t),\ \boldsymbol{\mu}(t)]$. With a large enough computer, we could contemplate solving them by brute force. But for real oceanic models with large time spans and large state vectors, even the biggest supercomputers are easily swamped, and one needs to find other methods.

Eq. (4.18), when re-written slightly is the adjoint evolution equation,

$$\boldsymbol{\mu}\,(t) = \mathbf{A}^{T}\boldsymbol{\mu}\,(t+1) + \mathbf{E}\,(t)\,\mathbf{R}\,(t)^{-1}\,[\mathbf{E}\,(t)\,\mathbf{x}\,(t) - \mathbf{y}\,(t)]\,, \tag{4.20}$$

in which the model/data misfit appears as a "source term." It is sometimes said that time runs backwards in this equation, with $\boldsymbol{\mu}\,(t)$ being computed most naturally from $\boldsymbol{\mu}\,(t+1)$ and the source term. Eq. (4.19) provides a "starting condition" in this view. But in fact, time has no particular direction here, as the equations govern a time interval, $1 \le t \le t_f$, and indeed if $\mathbf{A}^{-1}$ exists, there is no problem in rewriting Eq. (4.18) so that $\boldsymbol{\mu}\,(t+1)$ is given in terms of $\mathbf{A}^{-T}\boldsymbol{\mu}\,(t)$.

The Lagrange multipliers—that is, the adjoint model—have the same interpretation that they did for the steady models described in Chapter 2—that is, as a measure of the objective function sensitivity to the data,

$$\tfrac{1}{2}\frac{\partial J'}{\partial \mathbf{B}\mathbf{q}(t)} = \boldsymbol{\mu}(t+1)\,. \tag{4.21}$$

The physics of the adjoint model, as in Chapter 2, are again represented by the matrix $\mathbf{A}^{T}$. For a forward model that is both linear and self-adjoint ($\mathbf{A}^{T} = \mathbf{A}$), the adjoint solution would have the same physics as the state vector. If the model is not self-adjoint (the usual situation), the evolution of the adjoint may have a radically different interpretation than $\mathbf{x}(t)$. Insight into that physics is the road to understanding of information flow in the ocean. For example, if one employed a large general circulation model to compute the oceanic flux of heat, and wished to understand the extent to which the result was sensitive to the wind forcing at various places, or to a prescribed flux somewhere, the adjoint solution carries that information. In the future, one expects to see display and discussion of the results of the adjoint model on a nearly equal footing with that of the forward model.

4.2. Terminal Constraint Problem: Open Loop Control. Consider first the adjoint approach in the context of the simple chemical box model already described and depicted in Figure 7. The following idealized situation was considered. At $t = 0$, the tracer concentrations in the boxes are known to vanish—that is, $\mathbf{x}(0) = \mathbf{x}_0 = \mathbf{0}$ (the initial conditions are supposedly known exactly). At $t = t_f$, the region is surveyed, and the concentrations $\mathbf{y}(t_f) = \mathbf{E}(t_f)\mathbf{x}(t_f) + \mathbf{n}(t_f)$, $\mathbf{E}(t_f) \equiv \mathbf{I}$, $\langle \mathbf{n}(t)\rangle = \mathbf{0}$, $\langle \mathbf{n}(t_f)\mathbf{n}(t_f)^T\rangle = \mathbf{R}$ are known. No other observations are available. The question posed is: If the boundary conditions are all unknown a priori—that is $\mathbf{Bq} \equiv \mathbf{0}$—what boundary conditions would produce the observed values at t_f within the estimated error bars?

The problem is an example of a *terminal constraint control problem*–it seeks controls (forces, etc.) able to drive the system from an observed initial state to within a given tolerance of a required terminal state[27]. But in the present context, we interpret the result as an *estimate* of the actual boundary condition. For this special case, take the objective function,

$$\begin{aligned} J = &\left(\mathbf{x}(t_f) - \mathbf{x}_d\right)^T \mathbf{R}(t_f)^{-1}\left(\mathbf{x}(t_f) - \mathbf{x}_d\right) + \sum_{t=0}^{t_f-1} \mathbf{u}^T(t)\mathbf{Q}^{-1}\mathbf{u}(t) \\ &- 2\sum_{0}^{t_f-1} \boldsymbol{\mu}(t)^T \left[\mathbf{x}(t) - \mathbf{A}\mathbf{x}(t-1) - \mathbf{B}\mathbf{q}(t-1) - \Gamma\mathbf{u}(t-1)\right], \end{aligned} \tag{4.22}$$

and the governing normal equations are ,

$$\boldsymbol{\mu}(t-1) = \mathbf{A}^T\boldsymbol{\mu}(t) \tag{4.23}$$

$$\boldsymbol{\mu}(t_f) = \mathbf{R}^{-1}\left(\mathbf{x}(t_f) - \mathbf{x}_d\right) \tag{4.24}$$

$$\mathbf{Q}^{-1}\mathbf{u}(t-1) = -\Gamma^T\boldsymbol{\mu}(t) \tag{4.25}$$

plus the model. Eliminating

$$\mathbf{u}(t-1) = -\mathbf{Q}\Gamma^T\boldsymbol{\mu}(t) \tag{4.26}$$

[27] Luenberger (1979)

and substituting into the model, the system to be solved is

$$\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t-1) - \Gamma\mathbf{Q}\Gamma^T\boldsymbol{\mu}(t)\,, \qquad \mathbf{x}(0) = \mathbf{x}_0 \equiv \mathbf{0}\,, \tag{4.27}$$

$$\boldsymbol{\mu}(t-1) = \mathbf{A}^T\boldsymbol{\mu}(t)\,, \quad 1 \le t \le t_f - 1\,, \tag{4.28}$$

$$\boldsymbol{\mu}(t_f) = \mathbf{R}^{-1}\big(\mathbf{x}(t_f) - \mathbf{x}_d\big)\,. \tag{4.29}$$

The coupled problem must be solved for $\mathbf{x}(t)$, $\boldsymbol{\mu}(t)$ but with the initial conditions on the state vector provided at $t = 0$ and those for the Lagrange multipliers provided at the terminal time t_f, in terms of the still unknown $\mathbf{x}(t_f)$ [Equation (4.29)], recognizing that the estimated terminal state and the desired one will almost always differ.

This present problem can be solved in straightforward fashion without having to deal with the giant set of simultaneous equations by exploiting the special structure present in them. Using (4.29), step backward in time from t_f via (4.28) to produce

$$\begin{aligned}\boldsymbol{\mu}(t_f-1) &= \mathbf{A}^T\mathbf{R}^{-1}\big(\mathbf{x}(t_f) - \mathbf{x}_d\big)\\ &\;\;\vdots\\ \boldsymbol{\mu}(0) &= \mathbf{A}^{(t_f)T}\mathbf{R}^{-1}\big(\mathbf{x}(t_f) - \mathbf{x}_d\big)\end{aligned} \tag{4.30}$$

[?check this?] so $\boldsymbol{\mu}(t)$ is given in terms of the known $\mathbf{x}_d$ and the still unknown $\mathbf{x}(t_f)$. Substituting into (4.27) generates

$$\begin{aligned}\mathbf{x}(1) &= \mathbf{A}\mathbf{x}(0) - \Gamma\mathbf{Q}\Gamma^T\mathbf{A}^{(t_f-1)T}\mathbf{R}^{-1}\big(\mathbf{x}(t_f) - \mathbf{x}_d\big)\\ \mathbf{x}(2) &= \mathbf{A}\mathbf{x}(1) - \Gamma\mathbf{Q}\Gamma^T\mathbf{A}^{(t_f-2)T}\mathbf{R}^{-1}\big(\mathbf{x}(t_f) - \mathbf{x}_d\big)\\ &= \mathbf{A}^2\mathbf{x}(0) - \mathbf{A}\Gamma\mathbf{Q}\Gamma^T\mathbf{A}^{(t_f-1)T}\mathbf{R}^{-1}\big(\mathbf{x}(t_f) - \mathbf{x}_d\big)\\ &\quad - \Gamma\mathbf{Q}\Gamma^T\mathbf{A}^{(t_f-2)T}\mathbf{R}^{-1}\big(\mathbf{x}(t_f) - \mathbf{x}_d\big)\\ &\;\;\vdots\\ \mathbf{x}(t_f) &= \mathbf{A}^{t_f}\mathbf{x}(0) - \mathbf{A}^{(t_f-1)}\Gamma\mathbf{Q}\Gamma^T\mathbf{A}^{(t_f-1)T}\mathbf{R}^{-1}\big(\mathbf{x}(t_f) - \mathbf{x}_d\big)\\ &\quad - \mathbf{A}^{(t_f-2)}\Gamma\mathbf{Q}\Gamma^T\mathbf{A}^{(t_f-2)T}\mathbf{R}^{-1}\big(\mathbf{x}(t_f) - \mathbf{x}_d\big)\\ &\quad - \cdots - \Gamma\mathbf{Q}\Gamma^T\mathbf{R}^{-1}\big(\mathbf{x}(t_f) - \mathbf{x}_d\big)\,.\end{aligned}$$

The last equation permits us to bring the terms in $\mathbf{x}(t_f)$ over to the left-hand side and solve for $\mathbf{x}(t_f)$ in terms of $\mathbf{x}_d$ and $\mathbf{x}(0)$:

$$\begin{aligned}&\Big\{\mathbf{I} + \mathbf{A}^{(t_f-1)}\Gamma\mathbf{Q}\Gamma^T\mathbf{A}^{(t_f-1)T}\mathbf{R}^{-1}\\ &\quad + \mathbf{A}^{(t_f-2)}\Gamma\mathbf{Q}\Gamma^T\mathbf{A}^{(t_f-2)T}\mathbf{R}^{-1} + \cdots + \Gamma\mathbf{Q}\Gamma^T\mathbf{R}^{-1}\Big\}\,\mathbf{x}(t_f)\\ &= \mathbf{A}^{t_f}\mathbf{x}(0) + \Big\{\mathbf{A}^{(t_f-1)}\Gamma\mathbf{Q}\Gamma^T\mathbf{A}^{(t_f-1)T}\mathbf{R}^{-1}\\ &\qquad + \mathbf{A}^{(t_f-2)}\Gamma\mathbf{Q}\Gamma^T\mathbf{A}^{(t_f-2)T}\mathbf{R}^{-1} + \cdots + \Gamma\mathbf{Q}\Gamma^T\mathbf{R}^{-1}\Big\}\,\mathbf{x}_d\,.\end{aligned}$$

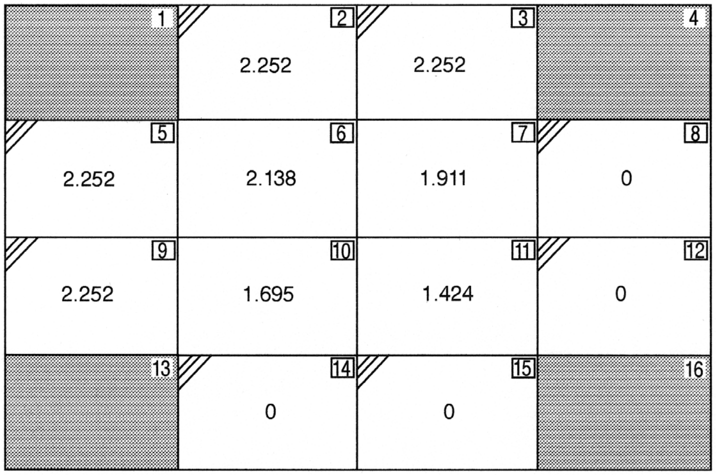


FIGURE 14. Terminal time concentrations (corresponding to the time histories in Figure 8), representing the terminal constraint (within error bars) of the control problem for transient tracers.

[?check all]

With $\mathbf{x}(t)$ now known, $\boldsymbol{\mu}(t)$ can be computed for all t from (4.28, 4.29). Then the control $\mathbf{u}(t)$ is also known from (4.25) and the state vector can be found from (4.27). The resulting solution for $\tilde{\mathbf{u}}(t)$ is in terms of the externally prescribed $\mathbf{x}_0$, $\mathbf{x}_d$.

Let $\tilde{\mathbf{x}}(t_f)$ be as shown in Figure 14. The initial conditions were taken as zero. Then Figure 15 shows the controls able to produce these terminal conditions within a tolerance governed by $\mathbf{R}$. The adjoint solution is depicted in Figure 15c.[28]

The canonical form for a terminal constraint problem usually used in the literature differs slightly; it is specified in terms of a given, nonzero, initial condition $\mathbf{x}(0)$, and the controls are determined so as to come close to a desired terminal state, which is zero. The solution to this so-called *deadbeat* control (driving the system to rest) problem can be used to solve the problem for an arbitrary desired terminal state.

The smoothing problem has been solved without computing the uncertainty, and is the great advantage of Lagrange multiplier/adjoint methods over the sequential estimators. Adjoint methods solve for the entire time domain at once; consequently, there is no weighted averaging of intermediate solutions and no need for the uncertainties. On the other hand, in view of everything that has come before in this book, the utility of solutions without uncertainty estimates must be questioned. In the context of Chapter 1, problems of arbitrary posedness are being solved. The various methods using objective functions, prior statistics, etc., whether

[28] A version of this methodology was used by Wunsch (1988a) and Mémery and Wunsch (1990) to discuss the evolution of observed tritium in the North Atlantic.

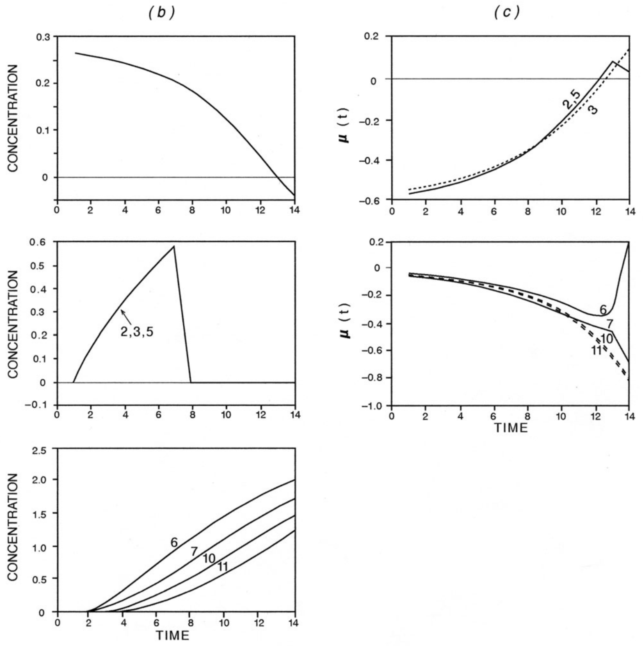


FIGURE 15. For the model in Fig. 14. the upper panel (b) displays the concentration rate of change—that is, du/dt — of the boundary concentrations (controls) as estimated through the Lagrange multiplier method. Middle panel shows du/dt for the "truth" and the lowest panel shows the time history of tracer concentrations in the interior boxes from the controls in the upper-most panel. These interior values can be compared to the "truth" in Fig. 8. The terminal constraint can be reached with the tolerances imposed from the initial conditions by a set of boundary controls quite different from the correct ones. The available terminal observations only loosely constrain the boundary conditions. (c) The upper panel shows the Lagrange multipliers (adjoint solution) for boundary boxes 2, 3, 5, and the lower panel for the interior boxes. As is typical of many control problems, the adjoint solution shows its greatest structure near the terminal time when comparatively large "forces" are exerted by the controls to push the system into the required terminal state.

in time-evolving or static situations, permit stable, useful estimates to be made under almost any circumstances, using almost any sort of available information. But the reader will by now appreciate that the use of such methods can produce structures in the solution, pleasing or otherwise, that may be present because they are required by (1) the observations, (2) the model, (3) the prior statistics, (4) some norm or smoothness demand on elements of the solution, or (5) all of the preceding in concert. A solution produced in ignorance of these differing sources of structure can hardly be thought very useful, and it is the uncertainty matrices that are usually the key to understanding. Consequently, we will later spend some time examining the problem of obtaining the missing covariances. In the meantime, one should note that the covariances of the filter/smoother will also describe the uncertainty of the Lagrange multiplier/adjoint solution, because they are the same solution to the same set of equations deriving from the same objective function.

There is one situation where a solution without uncertainty estimates is plainly useful—it is where one simply inquires, "Is there a solution at all?"—that is, when one wants to know if the observations actually contradict the model. In that situation, mere existence of an acceptable solution may be of greatest importance, suggesting, for example, that a model of adequate complexity is already available.

4.3. Representers/Method of Unit Solutions/Boundary Green Functions. The particular structure of Eqs. (4.15-4.19) permits several different methods of solution. The solution just given is an example. Let us generalize this problem by assuming observations at a set of arbitrary times (not just the terminal time),

$$\mathbf{y}\left(t\right)=\mathbf{E}\left(t\right)\mathbf{x}\left(t\right)+\mathbf{n}\left(t\right).$$

We will do this in two ways.

Take the objective function to be,

$$\begin{aligned}J=&\sum_{t=1}^{t_f}\left(\mathbf{E}\left(t\right)\mathbf{x}\left(t\right)-\mathbf{y}\left(t\right)\right)^T\mathbf{R}\left(t\right)^{-1}\left(\mathbf{E}\left(t\right)\mathbf{x}\left(t\right)-\mathbf{y}\left(t\right)\right)+\sum_{t=0}^{t_f-1}\mathbf{u}\left(t\right)^T\mathbf{Q}^{-1}\mathbf{u}\left(t\right)\\&-2\sum_{t=0}^{t_f-1}\boldsymbol{\mu}\left(t+1\right)^T\left[\mathbf{x}\left(t+1\right)-\mathbf{A}\mathbf{x}\left(t\right)-\mathbf{B}\mathbf{q}\left(t\right)-\boldsymbol{\Gamma}\left(t\right)\mathbf{u}\left(t\right)\right],\end{aligned}$$

so that the terminal state estimate is subsumed into the first term with $\mathbf{E}\left(t_f\right)=\mathbf{I}$, $\mathbf{R}\left(t_f\right)=\mathbf{P}\left(t_f\right)$. Let $\mathbf{x}_a\left(t\right)$ be the solution to the pure, unconstrained, forward problem,

$$\mathbf{x}_a\left(t+1\right)=\mathbf{A}\mathbf{x}_a\left(t\right)+\mathbf{B}\mathbf{q}\left(t\right),\ \mathbf{x}_a\left(0\right)=\mathbf{x}_0, \tag{4.31}$$

and redefine $\mathbf{x}\left(t\right)$ to be the difference, $\mathbf{x}\left(t\right)\rightarrow\mathbf{x}\left(t\right)-\mathbf{x}_a\left(t\right)$, that is the deviation from what can be regarded as the *a priori* solution. The purpose of this redefinition is to remove any inhomogeneous initial or boundary conditions from the problem—exploiting the system linearity. The normal equations are then,

$$\begin{aligned}\frac{1}{2}\frac{\partial J}{\partial\mathbf{u}\left(t\right)}&=\mathbf{Q}^{-1}\mathbf{u}\left(t\right)+\boldsymbol{\Gamma}^T\boldsymbol{\mu}\left(t+1\right)=0\\\frac{1}{2}\frac{\partial J}{\partial\mathbf{x}\left(t\right)}&=\mathbf{E}\left(t\right)^T\mathbf{R}\left(t\right)^{-1}\left[\mathbf{E}\left(t\right)\mathbf{x}\left(t\right)-\mathbf{y}\left(t\right)\right]+\mathbf{A}^T\boldsymbol{\mu}\left(t+1\right)-\boldsymbol{\mu}\left(t\right)=0\\\frac{1}{2}\frac{\partial J}{\partial\boldsymbol{\mu}\left(t\right)}&=\mathbf{x}\left(t+1\right)-\mathbf{A}\mathbf{x}\left(t\right)-\boldsymbol{\Gamma}\left(t\right)\mathbf{u}\left(t\right)=0,\ \mathbf{x}\left(0\right)=\mathbf{0}.\end{aligned}$$

Eliminating the control term in favor of $\boldsymbol{\mu}$, we have, as before,

$$\mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t) - \boldsymbol{\Gamma}\mathbf{Q}^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\mu}(t+1) \tag{4.32}$$

$$\boldsymbol{\mu}(t) = \mathbf{A}^T\boldsymbol{\mu}(t+1) + \mathbf{E}(t)^T\mathbf{R}(t)^{-1}[\mathbf{E}(t)\mathbf{x}(t) - \mathbf{y}(t)]. \tag{4.33}$$

The system is linear, so we can examine the solution for the inhomogeneous term in (4.33) at one time, $t = t_m$. The inhomogeneous term on the right-hand side of (4.33) is however, unknown until $\mathbf{x}(t_m)$ has been determined. So to proceed, let us first solve the different problem,

$$\mathbf{M}(t, t_m) = \mathbf{A}^T\mathbf{M}(t+1, t_m) + \mathbf{I}\boldsymbol{\delta}_{t,t_m},\ t \le t_m \tag{4.34}$$

$$\mathbf{M}(t, t_m) = 0,\ t > t_m, \tag{4.35}$$

where the second argument denotes the time of the observations (notice that $\mathbf{M}$ is a matrix). The inhomogeneous term is known here and has unit amplitude. There is a corresponding solution to (4.32) with these values of $\boldsymbol{\mu}$,

$$\mathbf{G}(t+1, t_m) = \mathbf{A}\mathbf{G}(t, t_m) - \boldsymbol{\Gamma}\mathbf{Q}^{-1}\boldsymbol{\Gamma}^T\mathbf{M}(t+1, t_m). \tag{4.36}$$

Both $\mathbf{G}, \mathbf{M}$ are computable independent of the actual data values. Now put,

$$\mathbf{m}(t, t_m) = \left\{\mathbf{E}(t_m)^T\mathbf{R}(t_m)^{-1}[\mathbf{E}(t_m)\mathbf{x}(t_m) - \mathbf{y}(t_m)]\right\}^T\mathbf{M}(t, t_m), \tag{4.37}$$

which is the solution to (4.33) once $\mathbf{x}(t_m)$ is known. Let

$$\mathbf{x}(t, t_m) = \left\{\mathbf{E}(t_m)^T\mathbf{R}(t_m)^{-1}[\mathbf{E}(t_m)\mathbf{x}(t_m, t_m) - \mathbf{y}(t_m)]\right\}^T\mathbf{G}(t, t_m), \tag{4.38}$$

substitute into (4.32) and solve,:

$$\begin{aligned}&\mathbf{x}(t_m, t_m) = \\ &-\left[\mathbf{I} - \mathbf{E}(t_m)^T\mathbf{R}(t_m)^{-1}\mathbf{E}(t_m)\right]^{-1}[\mathbf{E}(t_m)\mathbf{R}(t_m)\mathbf{y}(t_m)]^T\mathbf{G}(t_m, t_m).\end{aligned} \tag{4.39}$$

With $\mathbf{x}(t_m, t_m)$ known, Eq. (4.38) produces a fully determined $\mathbf{x}(t, t_m)$ in *representer* form. This solution is evidently just a variant of Eqs. (6.3.35-6.3.37).

One can then sum the results from all observation times:

$$\tilde{\mathbf{x}}(t) = \sum_{t_m=1}^{t_f}\mathbf{x}(t, t_m). \tag{4.40}$$

and after adding $\mathbf{x}_a(t)$ to the result, the entire problem is solved.

The solutions $\mathbf{M}(t, t_m)$ are obviously the Green function for the adjoint model equation, and the $\mathbf{G}(t, t_m)$ are "representers."[29] *If the data distribution is spatially sparse, one need only compute the subsets of the columns of* $\mathbf{M}, \mathbf{G}$ *that correspond to the positions of the data.*

[29] Bennett (2002).

The representer emerged naturally from the Lagrange multiplier formulation. Let us rederive the solution without the use of Lagrange multipliers to demonstrate how the adjoint model appears in unconstrained l_2 norm problems (soft constraints) and to connect to the continuous time formulation. Introduce the model into the same objective function as above, except we do it by substitution for the control terms; let $\mathbf{\Gamma} = \mathbf{I}$, making it possible to solve for $\mathbf{u}(t)$ explicitly and producing the simplest results. The objective function then is,

$$J = \sum_{t=0}^{t_f} \left(\mathbf{E}(t)\,\mathbf{x}(t) - \mathbf{y}(t)\right)^T \mathbf{R}(t)^{-1} \left(\mathbf{E}(t)\,\mathbf{x}(t) - \mathbf{y}(t)\right) + \tag{4.41}$$

$$\sum_{t=0}^{t_f-1} \left[\mathbf{x}(t+1) - \mathbf{A}\mathbf{x}(t)\right]^T \mathbf{Q}(t)^{-1} \left[\mathbf{x}(t+1) - \mathbf{A}\mathbf{x}(t)\right], \tag{4.42}$$

again assume that $\mathbf{x}(t)$ is the anomaly relative to the known $\mathbf{x}_a(t)$.

The normal equations are:

$$\begin{aligned} \frac{1}{2}\frac{\partial J}{\partial \mathbf{x}(t)} &= \mathbf{E}(t)^T \mathbf{R}(t)^{-1} \left[\mathbf{E}(t)\,\mathbf{x}(t) - \mathbf{y}(t)\right] - \\ &\quad \mathbf{A}^T \mathbf{Q}(t)^{-1} \left[\mathbf{x}(t+1) - \mathbf{A}\mathbf{x}(t)\right] + \\ &\quad \mathbf{Q}(t)^{-1} \left[\mathbf{x}(t) - \mathbf{A}\mathbf{x}(t-1)\right] = 0 \end{aligned} \tag{4.43}$$

Define

$$\boldsymbol{\nu}(t+1) = -\mathbf{Q}(t)^{-1} \left[\mathbf{x}(t+1) - \mathbf{A}\mathbf{x}(t)\right] \tag{4.44}$$

so that the system (4.43) can be written as

$$\boldsymbol{\nu}(t) = \mathbf{A}^T \boldsymbol{\nu}(t+1) + \mathbf{E}(t)^T \mathbf{R}(t)^{-1} \left[\mathbf{E}(t)\,\mathbf{x}(t) - \mathbf{y}(t)\right] \tag{4.45}$$

which along with (4.44) is precisely the same system of equations (4.32, 4.33) that emerged from the adjoint/Lagrange multiplier approach, if we let $\boldsymbol{\mu} \rightarrow \boldsymbol{\nu},\ \mathbf{\Gamma} = \mathbf{I}$. Representers are again defined as the unit disturbance solution to the system. As a by-product, we see once again, that l_2-norm least-squares and the adjoint method are simply different algorithmic approaches to the same problem.[30]

4.4. The Initialization Problem. Another special case of wide interest is determination of the initial conditions, $\tilde{\mathbf{x}}(0)$, from later observations. Let us attempt it using several different methods.

For notational simplicity and without loss of generality, assume that the known controls vanish so that the model is

$$\mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t) + \Gamma\mathbf{u}(t)\,, \tag{4.46}$$

[30] The use of the adjoint to solve l_2-norm problems is discussed by Bryson and Ho (1975, Section 13.3), who relax the restriction of full controllability, $\mathbf{\Gamma} = \mathbf{I}$. Because of the connection to regulator/control problems, a variety of methods for solution is explored there.

that there is an existing estimate of the initial conditions, $\tilde{\mathbf{x}}_0(0)$, with estimated uncertainty $\mathbf{P}(0)$, and that there is a single terminal observation of the complete state,

$$\mathbf{y}(t_f) = \mathbf{E}\mathbf{x}(t_f) + \mathbf{n}(t_f)\,, \quad \mathbf{E} = \mathbf{I} \tag{4.47}$$

where the observational noise covariance is again $\mathbf{R}(t_f)$.

We now can solve this problem in four seemingly different but nonetheless identical ways:

(1) The terminal observations can be written explicitly in terms of the initial conditions as

$$\begin{aligned}\mathbf{y}(t_f) = \mathbf{A}^{t_f}\mathbf{x}(0) + \mathbf{A}^{t_f-1}\Gamma\mathbf{u}(0) + \mathbf{A}^{t_f-2}\Gamma\mathbf{u}(1) + \cdots \\ + \Gamma\mathbf{u}(t_f-1) + \mathbf{n}(t_f)\,,\end{aligned} \tag{4.48}$$

which is in canonical observation equation form,

$$\begin{aligned}\mathbf{y}(t_f) = \mathbf{E}_p\mathbf{x}(0) + \mathbf{n}_p(t_f)\,, \quad \mathbf{E}_p = \mathbf{A}^{t_f}\,, \\ \mathbf{n}_p = \mathbf{A}^{t_f-1}\Gamma\mathbf{u}(0) + \cdots + \Gamma\mathbf{u}(t_f-1) + \mathbf{n}(t_f)\,,\end{aligned}$$

and where the covariance of this combined error is

$$\mathbf{R}_p \equiv \left\langle \mathbf{n}_p\mathbf{n}_p^T \right\rangle = \mathbf{A}^{t_f-1}\Gamma\mathbf{Q}\Gamma^T\mathbf{A}^{(t_f-1)T} + \cdots + \Gamma\mathbf{Q}\Gamma^T + \mathbf{R}(t_f)\,. \tag{4.49}$$

Then the least-squares recursive solution leads to

$$\tilde{\mathbf{x}}(0) = \tilde{\mathbf{x}}_0(0) + \mathbf{P}(0)\mathbf{E}_p^T\left(\mathbf{E}_p\mathbf{P}(0)\mathbf{E}_p^T + \mathbf{R}_p\right)^{-1}\left[\mathbf{y}(t_f) - \mathbf{E}_p\tilde{\mathbf{x}}_0(0)\right]\,, \tag{4.50}$$

and the uncertainty estimate follows immediately.

(2) A second method (which the reader should confirm produces the same answer) is to run the Kalman filter forward to t_f and then run the smoother backward to $t = 0$. There is more computation here, but a byproduct is an estimate of the intermediate values of the state vectors, of the controls, and their uncertainty.

(3) Write the model in backward form,

$$\mathbf{x}(t) = \mathbf{A}^{-1}\mathbf{x}(t+1) - \mathbf{A}^{-1}\Gamma\mathbf{u}\,, \tag{4.51}$$

and use the Kalman filter on this model, with time running backward. The observation equation (4.47) provides the initial estimate of $\mathbf{x}(t_f)$, and its error covariance becomes the initial estimate covariance $\mathbf{P}(t_f)$. At $t = 0$, the original estimate of $\tilde{\mathbf{x}}_0(0)$ is treated as an observation, with uncertainty $\mathbf{P}(0)$ taking the place of the usual $\mathbf{R}$. The reader should again confirm that the answer is the same as in (1).

(4) The problem has already been solved using the Lagrange multiplier/adjoint formalism.

5. The Steady-State Filter and Adjoint

For linear models, the Lagrange multiplier/adjoint method and the filter/smoother algorithms produce identical solutions. In both cases, the computation of the uncertainty remains an issue–in the former case because it is not part of the solution, and in the latter because it may overwhelm the computation. However, if the uncertainty is computed for the sequential estimator solutions, it must also represent the uncertainty derived from the Lagrange multiplier/adjoint principle. In the interests of gaining insight into both methods, and of ultimately finding uncertainty estimates, consider the covariance propagation equation for the Kalman filter:

$$\mathbf{P}(t+1,-) = \mathbf{A}(t)\mathbf{P}(t)\mathbf{A}(t)^T + \Gamma(t)\mathbf{Q}(t)\Gamma^T \tag{5.1}$$

$$\begin{aligned}\mathbf{P}(t+1) = {} & \mathbf{P}(t+1,-) \\ & - \mathbf{P}(t+1,-)\mathbf{E}(t+1)^T\big[\mathbf{E}(t+1)\mathbf{P}(t+1,-)\mathbf{E}(t+1)^T \\ & + \mathbf{R}(t+1)\big]^{-1}\mathbf{E}(t+1)\mathbf{P}(t+1,-)\end{aligned} \tag{5.2}$$

where we have substituted for $\mathbf{K}(t+1)$.

As we have seen, the computation of the covariances often dominates the Kalman filter (and smoother) and sometimes leads to the conclusion that the procedures are impractical. But as with all linear least-square—like estimation problems, the state vector uncertainty does not depend upon the actual data values, only upon the prior error covariances. Thus, the filter and smoother uncertainties (and the filter and smoother gains) can be computed in advance of the actual application to data, and stored. The computation can be done by stepping through the recursion in (5.1) starting from $t = 0$

Furthermore, it was pointed out that in Kalman filter problems, the covariances and Kalman gain approach a steady state, in which $\mathbf{P}(t)$, $\mathbf{P}(t,-)$, $\mathbf{K}(t)$ do not depend upon time after awhile. Physically, the growth in error from the propagation Equation (5.1) is just balanced by the reduction in uncertainty from the incoming data stream. [This simple description supposes the data come in at every time step; often the data appear only intermittently (but periodically), and the steady-state solution is periodic—errors displaying a characteristic saw-tooth structure between observation times.[31]]

[31] See the examples in Ghil et al. (1981) or Figure 5e.

If one can find these steady-state values, then the necessity to update the covariances and gain matrix disappears, and the computational load is much reduced, potentially by many orders of magnitude. The steady-state equation is often known as the *algebraic Riccati equation.*[32].

A steady-state solution to the Riccati equation would correspond not only to a determination of the steady-state filter and smoother covariances but also to the steady-state solution of the Lagrange multiplier/adjoint equations—a so-called steady-state control. Generalizations to the steady-state problem exist; an important one is the possibility of a periodic steady state.[33]

Before seeking a steady-state solution, one must determine whether one exists. That no such solution will exist in general is readily seen by considering, for example, a physical system in which certain components (elements of the flow) are not readily observed. If these components are initialized with partially erroneous values, then if they are unstable, they will grow without bound, and there will be no limiting steady-state value for the uncertainty, which will also have to grow without bound. Alternatively, suppose there are elements of the state vector whose values cannot be modified by the available control variables. Then no observations of the state vector produce information about the control variables; if the control vector uncertainty is described by $\mathbf{Q}$, then this uncertainty will accumulate from one time step to another, growing without bound with the number of time steps.

6. Controllability and Observability

In addition to determining whether there exists a steady-state solution either to the Riccati equation, there are many reasons for examining in some detail the existence of many of the matrix operations that have been employed routinely throughout. Matrix inverses occur throughout the developments above, and the issue of whether they actually exist has been ignored. Ultimately, however, one must face up to questions of whether some of the computations (e.g., the determination of deadbeat controls, or the finding of initial conditions from later observations, or the steady-state solution to the Riccati equation) are actually possible. The questions are intimately connected to some very useful structural descriptions of models and data that we will now examine briefly.

Consider the question of whether controls can be found to drive a system from a given initial state $\mathbf{x}(0)$ to an arbitrary $\mathbf{x}(t_f)$. If the answer is "yes," the system is said to be *controllable.* To find an answer, consider for simplicity,[34] a model with $\mathbf{B} = \mathbf{0}$ and with the control, u, a scalar.

[32] See Reid (1972) for a discussion of the history of the Riccati equation in general; it is intimately related to Bessel's equation and has been studied in scalar form since the 18th century. Bittanti et al. (1991a) discuss many different aspects of the matrix form.

[33] Discussed by Bittanti et al. (1991b)

[34] (following Franklin et al., 1990)

Then the model time steps can be written

$$
\begin{aligned}
\mathbf{x}(1) &= \mathbf{A}\mathbf{x}(0) + \Gamma u(0) \\
\mathbf{x}(2) &= \mathbf{A}\mathbf{x}(1) + \Gamma u(1) \\
&= \mathbf{A}^2\mathbf{x}(0) + \mathbf{A}\Gamma u(0) + \Gamma u(1) \\
&\vdots \\
\mathbf{x}(t_f) &= \mathbf{A}^{t_f}\mathbf{x}(0) + \sum_{j=0}^{t_f-1} \mathbf{A}^{t_f-1-j}\Gamma u(j) \qquad (6.1)\\
&= \mathbf{A}^{t_f}\mathbf{x}(0) + \left[\Gamma \;\; \mathbf{A}\Gamma \cdots \mathbf{A}^{t_f-1}\Gamma\right] \begin{bmatrix} u(t_f-1) \\ \vdots \\ u(0) \end{bmatrix} .
\end{aligned}
$$

To determine $u(t)$, we must be able to solve the system

$$
\left[\Gamma \;\; \mathbf{A}\Gamma \cdots \mathbf{A}^{t_f-1}\Gamma\right] \begin{bmatrix} u(t_f-1) \\ \vdots \\ u(0) \end{bmatrix} = \mathbf{x}(t_f) - \mathbf{A}^{t_f}\mathbf{x}(0)\,, \qquad (6.2)
$$

or

$$
\mathbf{C}\mathbf{u} = \mathbf{x}(t_f) - \mathbf{A}^{t_f}\mathbf{x}(0)
$$

for $u(t)$. The state vector dimension is N; therefore the dimension of $\mathbf{C}$ is N by the number of columns, t_f [a special case—$u(t)$ being scalar means that Γ is $N \times 1$]. Therefore, Equation (6.2) has no (ordinary) solution if t_f is less than N. If $t_f = N$ and $\mathbf{C}$ is nonsingular—that is, of rank N—there is a unique solution, and the system is controllable. If the dimensions of $\mathbf{C}$ are nonsquare, one could have a discussion, familiar from Chapter 2, of solutions for $u(t)$ with nullspaces present. If $t_f < N$, there is a nullspace of the desired output, and the system would not be controllable. If $t_f > N$, then there will still be a nullspace of the desired output, unless the rank is N, when $t_f = N$, and the system is controllable. The test can therefore be restricted to this last case.

This concept of controllability can be described in a number of interesting and useful ways[35] and generalized to vector controls and time-dependent models. To the extent that a model is found to be uncontrollable, it shows that some elements of the state vector are not connected to the controls, and one might ask why this is so and whether the model cannot then be usefully simplified.

[35] (e.g., Franklin et al., 1990; Stengel, 1986)

The concept of *observability* is connected to the question of whether given N perfect observations, it is possible to infer all of the initial conditions. Suppose that the same model 6.1 is used, and that we have (for simplicity only) a scalar observation sequence

$$y(t) = \mathbf{E}(t)\mathbf{x}(t) + n(t)\,, \quad 0 \le t \le t_f\,. \tag{6.3}$$

Can we find $\mathbf{x}(0)$? The sequence of observations can be written, with $u(t) \equiv 0$, as

$$\begin{aligned} y(1) &= \mathbf{E}(1)\mathbf{x}(1) = \mathbf{E}(1)\mathbf{A}\mathbf{x}(0) \\ &\vdots \\ y(t_f) &= \mathbf{E}(t_f)\mathbf{A}^{t_f}\mathbf{x}(0)\,, \end{aligned}$$

which is,

$$\begin{aligned} \mathbf{O}\mathbf{x}(0) &= \begin{bmatrix} y(1) \\ \vdots \\ y(t_f) \end{bmatrix} \\ \mathbf{O} &= \left\{ \begin{matrix} \mathbf{E}(1)\mathbf{A} \\ \vdots \\ \mathbf{E}(t_f)\mathbf{A}^{t_f} \end{matrix} \right\}. \end{aligned} \tag{6.4}$$

If the *observability matrix* $\mathbf{O}$ is square—that is, $t_f = N$ and is full rank—there is a unique solution for $\mathbf{x}(0)$, and the system is said to be observable. Should it fail to be observable, it suggests that at least some of the initial conditions are not determinable by an observation sequence and are irrelevant. Determining why that should be would surely shed light on the model one was using. As with controllability, the test (6.4) can be rewritten in a number of ways, and the concept can be extended to more complicated systems. The concepts of *stabilizability*, *reachability*, *reconstructability*, and *detectability* are closely related.[36], and there is a close connection between observability and controllability and the existence of a steady-state solution for the algebraic Riccati equations.

In practice, with real data and models, one must distinguish between mathematical observability and controllability and practical limitations imposed by the realities of observational systems. It is characteristic of fluids that changes occurring in some region at a particular time are ultimately communicated to all locations, no matter how remote, at later times, although the delay may be considerable, and the magnitudes of the signal may be much reduced by dissipation and geometrical spreading. Nonetheless, one anticipates that there is almost no possible element of a fluid flow, no matter how distant from a particular observation, that is not in principle observable.

[36] Goodwin and Sin (1984) or Stengel (1986)

7. Nonlinear Models

Fluid flows are nonlinear by nature, and one must address the data/model combination problem where the model is nonlinear. (There are also, as noted above, instances in which the data are nonlinear combinations of the state vector elements.) Nonetheless, the focus here on linear models is hardly wasted effort. As with more conventional systems, there are not many general methods for solving nonlinear estimation or control problems; rather, as with forward modeling, each situation has to be analyzed as a special case. Much insight is derived from a thorough understanding of the linear case, and indeed it is difficult to imagine tackling any nonlinear situation without a thorough grasp of the linear one. Not unexpectedly, the most general approaches to nonlinear estimation/control are based upon linearizations.

A complicating factor in the use of nonlinear models is that the objective functions need no longer have unique minima. There can be many nearby, or distant, minima, and the one chosen by the usual algorithms may depend upon exactly where one starts in the parameter space and how the search for the minimum is conducted. Indeed, the structure of the cost function may come to resemble a chaotic function, filled with hills, plateaus, and valleys into which one may stumble, never to get out again.[37] The combinatorial methods described in Chapter 3 are a partial solution.

7.1. The Linearized and Extended Kalman Filter. If one employs a nonlinear model,

$$\mathbf{x}(t+1) = \mathbf{L}\big(\mathbf{x}(t),\, \mathbf{Bq}(t),\, \Gamma(t)\mathbf{u}(t)\big)\,, \tag{7.1}$$

then reference to the Kalman filter recursion shows that the forecast step can be taken as before,

$$\tilde{\mathbf{x}}(t+1,-) = \mathbf{L}\big(\tilde{\mathbf{x}}(t),\, \mathbf{Bq}(t),\, 0\big)\,, \tag{7.2}$$

but it is far from clear how to propagate the uncertainty from $\mathbf{P}(t)$ to $\mathbf{P}(t+1,-)$, the previous derivation being based upon the assumption that the error propagates linearly, independent of the true value of $\mathbf{x}(t)$ (or equivalently, that if the initial error is Gaussian in character, then so is the propagated error). A number of approaches exist to finding approximate solutions to this problem, but they can no longer be regarded as strictly optimal, representing different linearizations.

Suppose that we write

$$\mathbf{x}(t) = \mathbf{x}_o(t) + \Delta\mathbf{x}(t)\,, \qquad \mathbf{q} = \mathbf{q}_0(t) + \Delta\mathbf{q}(t)\,, \tag{7.3}$$

[37] Miller, Ghil, and Gauthiez (1994) discuss some of the practical difficulties.

$$\begin{aligned}\mathbf{L}\big(\mathbf{x}(t),\,\mathbf{B}\mathbf{q}(t),\,\Gamma\mathbf{u}(t),\,t\big)\approx\mathbf{L}_o\big(&\mathbf{x}_o(t),\,\mathbf{q}_o(t),0,t\big)\\&+\,\mathbf{L}_x\left(\mathbf{x}_o(t),0\right)^T\Delta\mathbf{x}(t)\\&+\mathbf{L}_q\left(\mathbf{x}_o(t),\,\mathbf{q}_o(t)\right)^T\Delta\mathbf{q}(t)\\&+\,\mathbf{L}_u\left(\mathbf{x}_o(t),\,\mathbf{q}_o(t)\right)^T\mathbf{u}(t)\end{aligned}$$

where

$$\begin{aligned}\mathbf{L}_x\left(\mathbf{x}_o(t),\,\mathbf{q}_o(t)\right)&=\frac{\partial\mathbf{L}}{\partial\mathbf{x}(t)}\,,\\\mathbf{L}_q\left(\mathbf{x}_o(t),\,\mathbf{q}_o(t)\right)&=\frac{\partial\mathbf{L}}{\partial\mathbf{q}(t)}\,,\\\mathbf{L}_u\left(\mathbf{x}_o(t),\,\mathbf{q}_o(t)\right)&=\frac{\partial\mathbf{L}}{\partial\mathbf{u}(t)}\,.\end{aligned}\tag{7.4}$$

Then

$$\mathbf{x}_o(t+1)=\mathbf{L}_o\big(\mathbf{x}_o(t),\,\mathbf{B}\mathbf{q}_o(t),\,0,\,t\big)\tag{7.5}$$

defines a nominal solution, or trajectory, $\mathbf{x}_o(t)$. The model is assumed to be differentiable in this manner; all discrete models are so differentiable, numerically, barring a division by zero somewhere. Note that discrete models are, by definition, discontinuous, and discrete differentiation automatically accomodates that.

We have an equation for the solution perturbation:

$$\Delta\mathbf{x}(t+1)=\mathbf{L}_x\left(\mathbf{x}_o(t),\,\mathbf{q}_o(t)\right)^T\Delta\mathbf{x}(t)+\mathbf{L}_q^T\Delta\mathbf{q}(t)+\mathbf{L}_u^T\mathbf{u}(t)\,,\tag{7.6}$$

which is linear and of the form already used for the Kalman filter, but with redefinitions of the governing matrices. The full solution would be the sum of the nominal solution $\mathbf{x}_o(t)$ and the perturbation $\Delta\mathbf{x}(t)$. This form of estimate is sometimes known as the *linearized Kalman filter*, or the *neighboring optimal estimator*. Its usage depends upon the existence of a nominal solution, differentiability of the model, and the presumption that the controls $\Delta\mathbf{q}$, $\mathbf{u}$ do not drive the system too far from the nominal trajectory.

The so-called *extended Kalman filter* is nearly identical except that the linearization is taken not about a nominal solution but about the most recent estimate $\tilde{\mathbf{x}}(t)$; that is, the partial derivatives in (7.4) are evaluated using not $\mathbf{x}_o(t)$ but $\tilde{\mathbf{x}}(t)$. This latter form is more prone to instabilities, but if the system drifts very far from the nominal trajectory, it could well be more accurate than the linearized filter. The references go into these questions in great detail. Problems owing to multiple minima in the cost function[38] can always be overcome by having enough observations to keep the estimates close to the true state. Linearized smoothing algorithms can be developed in analogous ways, and the inability to track strong model nonlinearities is much less serious with a smoother than with a filter. The usual posterior checks of model and data

[38] (e.g., Miller et al., 1994)

residuals are also a very powerful precaution against a model failing to track the true state adequately.

7.2. Parameter Estimation (Adaptive Estimation). An important application of models used with data is to the estimation of empirical parameters used to describe the circulation. An example would be an attempt to improve estimates of eddy coefficients in the model by successively correcting them as more data appeared. Even with a linear dynamical model, such estimation generates a nonlinear estimation problem.

Suppose that the system is indeed linear and that the model contains a vector of parameters whose nominal values $_0$ we wish to improve upon while also estimating the state vector. Write the model as

$$\mathbf{x}(t+1) = \mathbf{A}\big(\mathbf{p}(t)\big)\mathbf{x}(t) + \mathbf{B}\mathbf{q}(t) + \Gamma\mathbf{u}(t) \tag{7.7}$$

where the time dependence in the parameters refers to the changing estimate of their value rather than a true physical time dependence. A general approach to solving this problem is to augment the state vector. That is,

$$\mathbf{x}_A(t) = \begin{bmatrix} \mathbf{x}(t) \\ \mathbf{p}(t) \end{bmatrix} . \tag{7.8}$$

Then write a model for this augmented state as

$$\mathbf{x}_A(t+1) = \mathbf{L}_A\big(\mathbf{x}_A(t),\, \mathbf{q}(t),\, \mathbf{u}(t)\big) \tag{7.9}$$

where

$$\mathbf{L}_A = \begin{Bmatrix} \mathbf{A}\big(\mathbf{p}(t)\big) & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{Bmatrix} \mathbf{x}_A + \mathbf{B}\mathbf{q}(t) + \Gamma\mathbf{u}(t)\,. \tag{7.10}$$

The observation equation is augmented simply as

$$\begin{aligned} \mathbf{y}_A(t) &= \mathbf{E}_A(t)\mathbf{x}_A(t) + \mathbf{n}_A(t)\,, \\ \mathbf{E}_A(t) &= \{\mathbf{E}(t) \quad \mathbf{0}\}\,, \quad \mathbf{n}_A(t) = \mathbf{n}(t)\,, \end{aligned}$$

assuming that there are no direct measurements of the parameters. The evolution equation for the parameters can be made more complex than indicated here. A solution can be found by using the linearized Kalman filter, for example, linearizing about the nominal parameter values. Parameter estimation is a very large subject.[39]

A major point of concern in estimation procedures based upon Gauss-Markov type methods lies in specification of the various covariance matrices, especially those describing the model error [here lumped into $\mathbf{Q}(t)$]. The reader will probably have concluded that there is, however, nothing precluding deduction of the covariance matrices from the model and observations, given that adequate numbers of observations are available. For example, it is straightforward to show

[39] (e.g., Anderson & Moore, 1979; Goodwin & Sin, 1984; Haykin, 1986)

that if a Kalman filter is operating properly, then the so-called innovation, $\mathbf{y}(t) - \mathbf{E}\tilde{\mathbf{x}}(t,-)$, should be uncorrelated with all previous measurements:

$$\langle \mathbf{y}(t')\big(\mathbf{y}(t) - \mathbf{E}\tilde{\mathbf{x}}(t,-)\big)\rangle = 0\,, \quad t' < t \tag{7.11}$$

[recall Ch. 2, [3.7.16]]. To the extent that (7.11) is not satisfied, the covariances need to be modified, and algorithms can be formulated for driving the system toward this condition. The possibilities for such procedures have been known for a long time and have an extended literature under the title *adaptive estimation*.[40]

7.3. Nonlinear Adjoint Equations; Searching for Solutions. Consider now a nonlinear model in the context of the Lagrange multipliers approach. Let the model be nonlinear so that a typical objective function is,

$$\begin{aligned} J &= \big(\mathbf{x}(0) - \tilde{\mathbf{x}}(0)\big)^T \mathbf{P}(0)^{-1}\big(\mathbf{x}(0) - \tilde{\mathbf{x}}(0)\big) \\ &+ \sum_{t=1}^{t_f} \big(\mathbf{E}(t)\mathbf{x}(t) - \mathbf{y}(t)\big)^T \mathbf{R}(t)^{-1}\big(\mathbf{E}(t)\mathbf{x}(t) - \mathbf{y}(t)\big) \\ &+ \sum_{t=0}^{t_f-1} \mathbf{u}(t)^T \mathbf{Q}(t)^{-1}\mathbf{u}(t) \\ &- 2\sum_{t=1}^{t_f} \boldsymbol{\mu}(t)^T \Big(\mathbf{x}(t) - \mathbf{L}\big(\mathbf{x}(t-1),\, \mathbf{B}\mathbf{q}(t-1),\, \Gamma\mathbf{u}(t-1)\big)\Big)\,. \end{aligned} \tag{7.12}$$

The normal equations are:

$$\frac{1}{2}\frac{\partial J}{\partial \mathbf{u}(t)} = \mathbf{Q}^{-1}\mathbf{u}(t) + \left(\frac{\partial \mathbf{L}}{\partial \mathbf{u}\,(t)}\right)^T \Gamma^T \boldsymbol{\mu}(t+1) = 0\,, \quad 0 \le t \le t_f - 1 \tag{7.13}$$

$$\frac{1}{2}\frac{\partial J}{\partial \boldsymbol{\mu}(t)} = \mathbf{x}(t) - \mathbf{L}\big(\mathbf{x}(t-1),\, \mathbf{B}\mathbf{q}(t-1),\, \Gamma\mathbf{u}(t-1)\big) = 0\,, \quad 1 \le t \le t_f \tag{7.14}$$

$$\frac{1}{2}\frac{\partial J}{\partial \mathbf{x}(0)} = \mathbf{P}(0)^{-1}\big(\mathbf{x}(0) - \tilde{\mathbf{x}}(0)\big) + \left(\frac{\partial \mathbf{L}}{\partial \mathbf{x}(0)}\right)^T \boldsymbol{\mu}(1) = 0\,, \tag{7.15}$$

$$\frac{1}{2}\frac{\partial J}{\partial \mathbf{x}(t)} = \mathbf{E}(t)^T \mathbf{R}(t)^{-1}\big(\mathbf{E}(t)\mathbf{x}(t) - \mathbf{y}(t)\big) - \boldsymbol{\mu}(t) \tag{7.16}$$

$$+ \left(\frac{\partial \mathbf{L}}{\partial \mathbf{x}(t)}\right)^T \boldsymbol{\mu}(t+1) = 0\,, \quad 1 \le t \le t_f - 1 \tag{7.17}$$

$$\frac{1}{2}\frac{\partial J}{\partial \mathbf{x}(t_f)} = \mathbf{E}(t_f)^T \mathbf{R}(t_f)^{-1}\big(\mathbf{E}(t_f)\mathbf{x}(t_f) - \mathbf{y}(t_f)\big) - \boldsymbol{\mu}(t_f) = 0\,. \tag{7.18}$$

[40] Among textbooks that discuss this subject are those of Haykin (1986), Goodwin and Sin (1984), and Ljung (1987).

These are nonlinear because of the nonlinear model (7.15)—although the adjoint model (7.16) remains linear in $\boldsymbol{\mu}(t)$—and the linear methods described above are no longer directly useful. As in the linearized Kalman filter, the operators,

$$\left(\frac{\partial \mathbf{L}\left(\mathbf{x}\left(t\right),\mathbf{Bq}\left(t\right),\mathbf{\Gamma u}\left(t\right),t\right)}{\partial \mathbf{u}\left(t\right)}\right)\ ,\ \left(\frac{\partial \mathbf{L}\left(\mathbf{x}\left(t\right),\mathbf{Bq}\left(t\right),\mathbf{\Gamma u}\left(t\right),t\right)}{\partial \mathbf{x}(t)}\right), \tag{7.19}$$

appearing in the above equations (or their transposes), are the derivatives of the model with respect to the control and statevectors. Assuming they exist, they represent a linearization of the model about the state and control vectors. They are sometimes called the “tangent linear model”. Their transposes are, in this context, the adjoint model. There is some ambiguity about the terminology: the form of (7.19) or the transposes are definable independent of the form of J. Otherwise, Eq. (7.16) and its boundary condition (7.18) depend upon the actual observations and the details of J; we will call this pair the “adjoint evolution” equation to distinguish it from the adjoint model.

One might anticipate that if the nonlinearity is not too great, perturbation methods might work. This notion leads to what is usually called *neighboring optimal control*[41]. Where the nonlinearity is too great, the approach to solution is an iterative one. Consider what one is trying to do. At the optimum, if we can find it, J will reach a stationary value in which the terms multiplying the $\boldsymbol{\mu}(t+1)$ will vanish identically. Essentially, one uses *search* methods that are able to find a solution (there may well be multiple such solutions, each corresponding to a local minimum of J).

There are many known ways to seek approximate solutions to a set of simultaneous equations, linear or nonlinear, using various search procedures. Most such methods are based upon what are usually denoted as *Newton* or *quasi-Newton* methods, or variations on steepest descent. The most popular approach to tackling the set (7.13)–(7.18) has been a form of conjugate gradient or modified steepest descent algorithm. The iteration cycles are commonly carried out by making a first estimate of the initial conditions and the boundary conditions–for example, setting $\mathbf{u} = \mathbf{0}$. One integrates (7.14) forward in time to produce a first guess for $\mathbf{x}(t)$. A first guess set of Lagrange multipliers is obtained by integrating (7.16) backward in time. Normally, (7.13) is not then satisfied, but because the values obtained provide information on the gradient of the objective function with respect to the controls, one knows the sign of the changes to make in the controls to reduce J. Perturbing the original guess for $\mathbf{u}(t)$ in this manner, one does another forward integration of the model and backward integration of the adjoint. In practice, the perturbations are determined by the conjugate gradient or other algorithm, continuing the iterations until convergence is obtained.

[41] (e.g., see Stengel, Chapter 5)

In this type of approximate solution, the adjoint solution, $\tilde{\boldsymbol{\mu}}(t)$, is really playing two distinct roles. On the one hand, it is a mathematical device to impose the model constraints; on the other, it is being used as a numerical convenience for determining the direction and step size to best reduce the objective function. The two roles are obviously intimately related, but as we have seen for the linear models, the first role is the primary one. The problem of possibly falling into the wrong minimum of the objective function remains here, too.

In practice, the AD tools already alluded to are often used to find the various model derivatives.

7.4. Sensitivity Analysis. Our employment so far of the adjoint model and the adjoint evolution equation, has been in the context of minimizing an objective function—and to some degree, the adjoint has been nothing but a numerical convenience for algorithms which find minima. As we have seen repeatedly however, Lagrange multipliers have a straightforward interpretation as the sensitivity of an objective function J, to perturbations in problem parameters. This use of the multipliers can be developed independently of the state estimation problem.

In many cases, one cares primarily about some quantity, $H\left(\tilde{\mathbf{x}}\left(t_f\right)\right)$, e.g. the heat flux or surface elevation, in ocean models, as given by the statevector at the end time $\mathbf{x}\left(t_f\right)$ of a model computation. Suppose[42] one seeks the sensitivity of that quantity to perturbations in the initial conditions (any other control variable could be considered), $\mathbf{x}_0$. Let $\mathbf{L}$ continue to be the operator defining the time-stepping of the model. Define $\Psi_n = \mathbf{L}\left(\mathbf{x}_n, \mathbf{q}\left(n\right), t = n\Delta t\right)$. Then,

$$H = H\left(\boldsymbol{\Psi}_{t_f}\left[\boldsymbol{\Psi}_{t_f-1}\left[...\boldsymbol{\Psi}_1\left[\mathbf{x}_0\right]\right]\right]\right),$$

that is, the function of the final state of interest is a nested set of operators working on the control vector $\mathbf{x}_0$. Then the derivative of H with respect to $\mathbf{x}_0$ is obtained from the chainrule,

$$\frac{\partial H}{\partial \mathbf{x}_0} = H'\left(\boldsymbol{\Psi}'_{t_f}\left[\boldsymbol{\Psi}'_{t_f-1}\left[...\boldsymbol{\Psi}'_1\left[\mathbf{x}_0\right]\right]\right]\right) \tag{7.20}$$

where the prime denotes the derivative with respect to the argument of the operator $\mathbf{L}$ evaluated at that time,

$$\frac{\partial \boldsymbol{\Psi}_n\left(\mathbf{p}\right)}{\partial \mathbf{p}}.$$

Notice that these derivatives, are the Jacobians (matrices) of dimension $N \times N$ at each time-step, and are the same derivatives that appear in the operators in (7.19). The nested operator (7.20) can be written as a matrix product,

$$\frac{\partial H}{\partial \mathbf{x}_0} = \nabla h^T \frac{\partial \boldsymbol{\Psi}_{t_f}\left(\mathbf{p}\right)}{\partial \mathbf{p}} \frac{\partial \boldsymbol{\Psi}_{t_f-1}\left(\mathbf{p}\right)}{\partial \mathbf{p}} ... \frac{\partial \boldsymbol{\Psi}_1\left(\mathbf{p}\right)}{\partial \mathbf{p}}. \tag{7.21}$$

∇h is the vector of derivatives of function H (the gradient) and so (7.21) is a column vector of dimension $N \times 1$. $\mathbf{p}$ represents the statevector at the prior timestep for each $\boldsymbol{\Psi}_n$. The adjoint

[42] We follow here, primarily, Marotzke et al. (1999).

compilers described above compute $\partial H/\partial \mathbf{x}_0$ in what is called the "forward mode", producing an operator which runs from right to left, multiplying t_f-$N\times N$ matrices starting with $\partial \mathbf{\Psi}_1\left(\mathbf{p}\right)/\partial \mathbf{p}$.

If however, Eq. (7.21) is transposed,

$$\left(\frac{\partial H}{\partial \mathbf{x}_0}\right)^T = \left(\frac{\partial \mathbf{\Psi}_1\left(\mathbf{p}\right)}{\partial \mathbf{p}}\right)^T \left(\frac{\partial \mathbf{\Psi}_2\left(\mathbf{p}\right)}{\partial \mathbf{p}}\right)^T ... \left(\frac{\partial \mathbf{\Psi}_{t_f}\left(\mathbf{p}\right)}{\partial \mathbf{p}}\right)^T \nabla h, \tag{7.22}$$

the first multiplication on the right involves multiplying the column vector ∇h by an $N \times N$ matrix, thus producing another $N\times 1$ vector. More generally, the set of products in (7.22), again taken from right to left, involves only multiplying a vector by a matrix, rather than a matrix by a matrix as in (7.21), with a potentially enormous computational saving. Such evaluation is the so-called reverse mode adjoint calculation (the transposes generate the required adjoint operators) and has only recently has become available in the automatic differential compilers. In comparing the computation in the forward and reverse modes, one must be aware that there is a storage penalty in (7.22) not incurred in (7.21).[43] In practice, the various operators $\partial \mathbf{\Psi}_i\left(\mathbf{p}\right)/\partial \mathbf{p}$ are not obtained explicitly, but are computed. Going much beyond these simple statements takes us too far into the technical details.

7.5. Adaptive Methods. All of the inverse problems we have discussed were reduced ultimately to finding the minimum of an objective function, either in unconstrained form or constrained by exact relationships (e.g., models). Once the model is formulated, the objective function agreed on, and the data obtained in appropriate form (often the most difficult step), the formal solution is reduced to finding the constrained or unconstrained minimum. *Optimization theory* is a very large, very sophisticated subject directed at finding such minima, and the methods we have described here—sequential estimation and adjoint/Pontryagin principles—are only two of a number of possibilities.

As we have seen, some of the methods stop at the point of finding a minimum and do not readily produce an estimate of the uncertainty of the solution. One can distinguish inverse methods from optimization methods by the requirement of the former for the requisite uncertainty estimates. Nonetheless, as noted before in some problems, mere knowledge that there is at least one solution may be of intense interest, irrespective of whether it is unique or whether its stability to perturbations in the data or model is well understood.

The reader interested in optimization methods generally is referred to the literature on that subject.[44] The oceanographic general circulation problem falls into the category of extremely large, nonlinear optimization, a category which tends to preclude the general use of many methods that are attractive for problems of more modest size.

[43] Restrepo et al. (1995) discuss some of the considerations.

[44] e.g., Gill et al., 1981; Luenberger, 1984; Scales, 1985

In the context of the more conventional methods, the continued exploration of ways to reduce the computational load without significantly degrading either the proximity to the true minimum or the information content (the uncertainties of the results) is a very high priority. Several approaches are known. We have already described the use of steady-state filters and smoothers where they exist. Textbooks discuss a variety of possibilities for simplifying various elements of the solutions. In addition to the steady-state assumption, methods include: (1) "state reduction"—attempting to remove from the model (and thus from the uncertainty calculation) elements of the state vector that are either of no interest or comparatively unchanging[45]; (2) "reduced-order observers",[46] in which some components of the model are so well observed that they do not need to be calculated; (3) proving or assuming that the uncertainty matrices (or the corresponding information matrices) are block diagonal or banded, permitting use of a variety of sparse algorithms. This list is not exhaustive.

8. Forward Models

The focus we have had on the solution of inverse problems has perhaps given the impression that there is some fundamental distinction between forward and inverse modeling. The point was made at the beginning of this book that inverse *methods* are important in solving forward as well as inverse *problems*. Almost all the inverse problems discussed here involved the use of an objective function, and such objective functions do not normally appear in forward modeling. The presence or absence of such functions thus might be considered a fundamental difference between the problem types.

But numerical models do not produce universal, uniformly accurate solutions to the fluid equations. Any modeler makes a series of decisions about which aspects of the flow are most important for accurate depiction—the energy flux, the large-scale velocities, the nonlinear cascades, etc.—and which cannot normally be achieved simultaneously with equal fidelity. It is rare that these goals are written explicitly, but they could be, and the modeler could choose the grid and differencing scheme, etc., to minimize a specific objective function. The use of such explicit objective functions would prove beneficial because it would quantify the purpose of the model.

One can also consider the solution of ill-posed forward problems. In view of the discussion throughout this book, the remedy is straightforward: One must introduce an explicit objective function of the now-familiar type, involving state vectors, observations, control, etc., and this approach is precisely that recommended by them. If a Lagrange multiplier method is adopted, then the relations Ch. 2 Eqs. (3.5.22,3.5.23) show that an over- or underspecified forward model produces a complementary under- or overspecified adjoint model, and it is difficult to sustain a

[45] Gelb, 1974

[46] Luenberger, 1964; O'Reilly, 1983

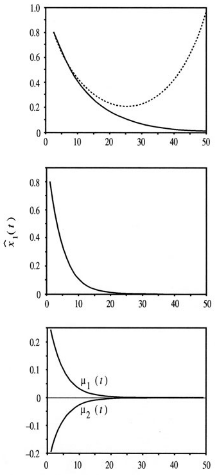


FIGURE 16. (Top) Solutions $x_1(t)$ to the discretized version of Eq. (8.1) for two slightly different initial conditions. The solid line shows the solution when the initial conditions are carefully chosen to avoid exciting the exponentially growing mode ($\mathbf{x}(1) = [0.800, 1.00]^T$). The dashed line shows the eventual emergence of the growing mode out of the background after sufficient time has passed, when the initial conditions were modified slightly to $\mathbf{x}(1) = [0.805, 1.00]^T$. (Middle) Solution to Eq. (8.1) when the growing mode is excited by the initial conditions, but a terminal constraint is provided—one which simultaneously overspecifies the system and stabilizes it through the objective function. (Lower) The two elements of the adjoint solution (Lagrange multipliers), $\boldsymbol{\mu}(t)$, used to constrain the objective function. Note the largest values are at the origin: the solution is most sensitive to specification of the initial conditions.

claim that modeling in the forward direction is fundamentally distinct from that in the inverse sense.

Example: Consider the ordinary differential equation

$$\frac{d^2\xi(t)}{dt^2} - k^2\xi(t) = 0\,. \tag{8.1}$$

Formulated as an initial value problem, it is properly posed with Cauchy conditions $\xi(0) = \xi_0$, $\xi'(0) = \xi'_0$. The solution is

$$\xi(t) = A\exp(kt) + B\exp(-kt)\,, \tag{8.2}$$

with A, B determined by the initial conditions. If we add another condition—for example, at the end of the interval of interest, $\xi(t_f) = \xi_{t_f}$—the problem is ill-posed because it is now overspecified. To analyze and solve such a problem using the methods of this book, discretize it as

$$\xi(t+1) - (2+k^2)\xi(t) + \xi(t-1) = 0\,, \tag{8.3}$$

taking $\Delta t = 1$, with corresponding redefinition of k^2. A canonical form is,

$$\mathbf{x}\left(t+1\right) = \mathbf{A}\mathbf{x}\left(t\right), \quad \mathbf{x}\left(t\right) = \left[\xi\left(t\right), \xi\left(t-1\right)\right], \quad \mathbf{A} = \begin{Bmatrix} 2+k^2 & -1 \\ 1 & 0 \end{Bmatrix}$$

These equations are easily solved by a backward sweep of the adjoint model to obtain $\boldsymbol{\mu}(1)$, which produces $\hat{\mathbf{x}}(1)$ in terms of $\mathbf{x}(t_f) - \mathbf{x}_d(t_f)$. A forward sweep of the model, to t_f, produces the numerical value of $\hat{\mathbf{x}}(t_f)$; the backward sweep of the adjoint model gives the corresponding numerical value of $\hat{\mathbf{x}}(1)$, and a final forward sweep of the model completes the solution. The subproblem forward and backward sweeps are always well posed. This recipe was run for

$$k^2 = 0.05, \quad \Delta t = 1, \quad \mathbf{x}_d(1) = [0.805,\ 1.0]^T, \quad \mathbf{P}(1) = 10^{-2}\mathbf{I}\,,$$

$$x(t_f = 50) = 1.427 \times 10^{-5}, \quad \mathbf{P}(50) = \begin{Bmatrix} 10^{-4} & 0 \\ 0 & 10^4 \end{Bmatrix}\,,$$

with results in Figure 16b,c. [The large subelement uncertainty in $\mathbf{P}(50)$, corresponding to scalar element $x(49)$, is present because we sought only to specify scalar element $x(50)$, in $\mathbf{x}(50)$.] If $\mathbf{x}_d(1) = [0.800,\ 1.0]^T$ instead of the value actually used, the unstable mode would not have been excited. Notice that the original ill-posedness in both overspecification and instability of the initial value problem have been dealt with. For a full GCM, the technical details are much more intricate, but the principle is not in doubt.

We are ending as we began—this result can be thought of as the solution to a forward problem, albeit ill-posed, or as the solution to a more or less conventional inverse one. The distinction between forward and inverse problems has nearly vanished.

Any forward model that is driven by observed conditions—for example, of the windstress or the buoyancy flux—is ill-posed in the sense that there can again be no unique solution, only a most probable one, smoothest one, etc. As with an inverse solution, forward calculations no more produce unique solutions in these circumstances than do inverse ones.

Adjoint model code generation via automatic differentiation and its application to ocean / sea ice state estimation

Patrick Heimbach and Dimitris Menemenlis

Massachusetts Intitute of Technology and JPL

Some people involved:

- **MITgcm model development:**
  A. Adcroft, J.M. Campin, C. Hill, J. Marshall
- **ECCO environment development:**
  A. Köhl, D. Stammer, C. Wunsch
- **Sea ice model development:**
  D. Menemenlis, J. Zhang
- **Adjoint model aspects**
  R. Giering, D. Ferreira, G. Gebbie, M. Losch

Table of contents (preliminary)

- References
- Examples of adjoint applications
- The estimation / optimization / optimal control problem
- Some definitions & some algebra
- The adjoint method
- Automatic differentiation (AD)
- A simple example: 3-box model of the ’THC’
- Reverse order flow and integration
- Input/output and active file handling
- Checking the automatically derived adjoint gradient
- Extension to vector-valued objective / ’cost’ functions
- More AD challenges: scalability, parameterization schemes
- The ECCO estimation problem: control variables, cost function, problem size
- Sea ice estimation: model description, adjoint sensitivity, work-in-progress
- Conclusions & outlook (scientific & code development)

Some references I

- **Some 'classics' on the adjoint method (subjective & incomplete):**
  - O. Talagrand and P. Courtier: Variational assimilation of meteorological observations with the adjoint vorticity equation. I: Theory. *Q. J. R. Meteorol. Soc.*, **113**, pp. 1311–1328, 1987.
  - P. Courtier and O. Talagrand: Variational assimilation of meteorological observations with the adjoint vorticity equation. II: Numerical results. *Q. J. R. Meteorol. Soc.*, **113**, pp. 1329–1347, 1987.
  - W.C. Thacker and R.B. Long: Fitting dynamics to data. *J. Geophys. Res.*, **93, C2**, pp. 1,227–1,240, 1988.
  - W.C. Thacker: The role of the Hessian matrix in fitting models to measurements. *J. Geophys. Res.*, **94, C5**, pp. 6,177–6,196, 1989.
  - E. Tziperman and W.C. Thacker: An optimal control / adjoint equation approach to studying the ocean general circulation. *J. Phys. Oceanogr.*, **19**, pp. 1471–1485, 1989.
  - E. Tziperman, W.C. Thacker, R.B. Long and S. Hwang: Oceanic data analysis using a general circulation model. I: Simulations. *J. Phys. Oceanogr.*, **22**, 1434–1457, 1992.
  - J. Marotzke and C. Wunsch: Finding the steady state of a general circulation model through data assimilation: Application to the North Atlantic Ocean. *J. Geophys. Res.*, **98, C11**, pp. 20,149–20,167, 1993.
  - O. Talagrand: Assimilation of observations, an introduction. *J. Meteorol. Soc. Japan*, **75**, pp. 191–209, 1997.
  - R. Errico: What is an adjoint model? *BAMS*, **78**, pp. 2577–2591, 1997.
  - C. Wunsch: The ocean circulation inverse problem. *CUP*, 1996.

Some references II

- **Automatic/algorithmic differentiation (AD):**
  - Giering & Kaminski: *Recipes for Adjoint Code Construction*.
    ACM Transactions on Mathematical Software, vol. 24, no. 4, 1998
  - *MITgcm manual*, chapter 5, available at `http://mitgcm.org/sealion`
  - Heimbach, Hill & Giering: *An efficient exact adjoint of the MIT general circulation model, generated via automatic differentiation*.
    submitted to 'Future Generation Computer Systems', 2002.
  - A. Griewank: *Evaluating Derivatives. Principles and Techniques of Algorithmic Differentiation.* Frontiers in Applied Mathematics, vol. 19, 369 pp., SIAM, Philadelphia, 2000.
  - A collection of resources on AD,
    `http://www.autodiff.org/` ,
    including an ongoing NSF-ITR project
    Adjoint Compiler Technology & Standards (ACTS)
    `http://www-unix.mcs.anl.gov/∼naumann/ACTS/`
  - A collection of publications using TAMC/TAF:
    `http://www.fastopt.de/references/all.html`

Some references III

- **Adjoint model applications to sensitivity studies:**

Atmosphere:

- M. Hall: Application of adjoint sensitivity theory to an atmospheric general circulation model. *J. Atmos. Sci.*, **43**, pp. 2644–2651, 1986.
- R. Errico and T. Vukicevic: Sensitivity analysis using an adjoint of the PSU/NCAR mesoscale model. *Mon. Wea. Rev.*, **12**, pp. 1644–1660, 1992.
- S. Corti and T. Palmer: Sensitivity analysis of atmospheric low-frequency variability. *Q. J. R. Meteorol. Soc.*, **123**, 2425–2447, 1997.

Ocean:

- J. Marotzke, R. Giering, K.Q. Zhang, D. Stammer, C. Hill and T. Lee: Construction of the adjoint MIT ocean general circulation model and application to Atlantic heat transport variability. *J. Geophys. Res.*, **104, C12**, pp. 29,529–29,547, 1999.
- V. Bugnion: Driving the ocean's overturning: an adjoint sensitivity study. Ph.D. thesis, MIT/EAPS, Cambridge (MA), USA, June 2001.
- M. Junge and T. Haine: Mechanisms of North Atlantic Wintertime Sea Surface Temperature Anomalies. *J. Clim.*, **14**, pp. 4560–4572, 2001.
- E. Galanti, E. Tziperman, M. Harrison, A. Rosati, R. Giering and Z. Sirkes: The equatorial thermocline outcropping - A seasonal control on the tropical Pacific ocean-atmosphere instability. *J. Clim.*, **15**, pp. 2721–2739, 2002.

Some references IV

- **Adjoint model applications to ocean state estimation:**
  - J. Sheinbaum and D. Anderson: Variational assimilation of XBT data. Part 1. *J. Phys. Oceanogr.*, **20**, 672–688, 1990.
  - J. Schröter, U. Seiler, and M. Wenzel: Variational assimilation of Geosat data into an eddy-resolving model of the Gulf stream extension area. *J. Phys. Oceanogr.*, **23**, pp. 925–953, 1993.
  - B. Luong, J. Blum and J. Verron: A variational method for the resolution of a data assimilation problem in oceanography. *Inverse Problems*, **14**, 979–997, 1998.
  - Martin Losch, René Riedler and Jens Schröter: Estimating a mean ocean state from hydrography and sea-surface height data with a nonlinear inverse section model: twin experiments with a synthetic dataset. *J. Phys. Oceanogr.*, **32**, pp. 2096-2112, 2002.
  - Jens Schröter, Martin Losch, and Bernadette Sloyan: Impact of the Gravity field and steady-state Ocean Circulation Explorer (GOCE) mission on ocean circulation estimates. Volume and heat transports across hydrographic sections. *J. Geophys. Res.*, **107, C2**, 2002.
  - D. Stammer, C. Wunsch, R. Giering, C. Eckert, P. Heimbach, J. Marotzke, A. Adcroft, C.N. Hill and J. Marshall: The global ocean circulation and transports during 1992 – 1997, estimated from ocean observations and a general circulation model. *J. Geophys. Res.*, **107, C9**, pp. 3118, 2002.
  - D. Stammer, C. Wunsch, R. Giering, C. Eckert, P. Heimbach, J. Marotzke, A. Adcroft, C.N. Hill and J. Marshall: Volume, heat and freshwater transports of the global ocean circulation 1993 – 2000, estimated from a general circulation model constrained by WOCE data. *J. Geophys. Res.*, **108, C1**, pp. 3007, 2003.

Some references V

- **Data assimilation in highly nonlinear systems:**

  Atmosphere:

  - M. Tanguay, P. Bartello and P. Gauthier: Four-dimensional data assimilation with a wide range of scales. *Tellus*, **47A**, pp. 974–997, 1995.
  - C. Pires, R. Vautard, O. Talagrand: On extending the limits of variational assimilation in nonlinear chaotic systems. *Tellus*, **48A**, pp. 96–121, 1996
  - R. Miller, E. Carter and S. Blue: Data assimilation into nonlinear stochastic models. *Tellus*, **51A**, pp. 167–194, 1999.

  Ocean:

  - D. Lea and M. Allen and T. Haine: Sensitivity analysis of the climate of a chaotic system. *Tellus*, **52A**, pp. 523–532, 2000.
  - D. Lea, M. Allen, T. Haine, and J. Hansen: Sensitivity analysis of the climate of a chaotic ocean circulation model. *Q. J. R. Meteorol. Soc.*, **128**, pp. 2587-2605, 2002.
  - A. Köhl and J. Willebrand: An adjoint method for the assimilation of statistical characteristics into eddy-resolving ocean models. *Tellus*, **54A**, pp. 406–425, 2002.
  - A. Köhl and J. Willebrand: Variational assimilation of SSH variablity from TOPEX/POSEIDON and ERS-1 into an eddy-permitting model of the North Atlantic. *J. Geophys. Res.*, **108, C3**, pp. 3092, 2003.
  - G. Gebbie et. al., in preparation, 2003.

The context of this state estimation system

- The ongoing MITgcm model development, `http://mitgcm.org/sealion/`
- The ongoing *Estimating the Circulation and Climate of the Ocean* (ECCO) project of global WOCE data–model synthesis, `http://www.ecco-group.org`
- The newly developed sea ice model by *Menemenlis & Zhang*
- A wealth of little-explored high-latitude ocean and sea ice data

Adjoint applications (I): Ocean State Estimation

- Given:
  - a set of (possibly different types of) observations
  - a numerical model & set of initial / boundary conditions
- Question: (estimation / optimal control problem)
  Find "*optimal*" model trajectory consistent with available observations within given prior errors
- Iterative approach:
  Minimize least square function $\mathcal{J}(\vec{u})$ measuring model vs. data misfit
  $\longrightarrow$ seek $\boxed{\vec{\nabla}_u \mathcal{J}(\vec{u})}$ to infer update $\Delta\vec{u}$ from variation of controls $\vec{u}$

$$\vec{u}^{n+1} = \vec{u}^{n} + \Delta\vec{u}$$

- Results:
  - optimal/consistent ocean state estimate
  - corrected initial/boundary value estimates

Adjoint applications (II): Sensitivity analysis

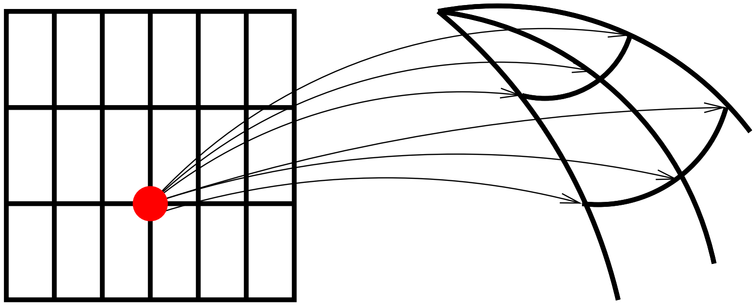
▶ Finite difference approach:

- Take a "guessed" anomaly (SST) and determine its impact on model output (MOC)
- Perturb each input element ($\mathrm{SST(i,j)}$) to determine its impact on output (MOC).

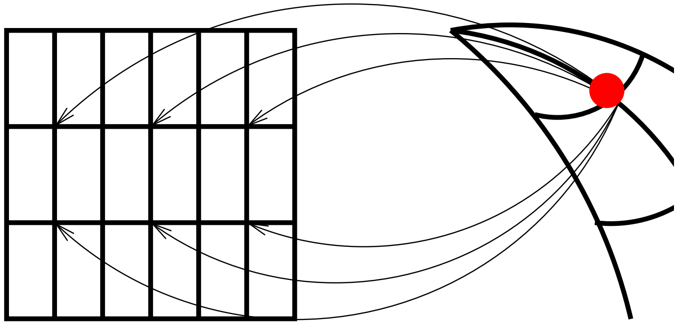
▶ Reverse/adjoint approach:

- Calculates "full" sensitivity field
  $$\frac{\partial\,\mathrm{MOC}}{\partial\,\mathrm{SST}(x,y,t)}$$
- Approach:
  Let $\mathcal{J} = \mathrm{MOC}$, $\vec{u} = \mathrm{SST(i,j)}$

$$\longrightarrow \boxed{\vec{\nabla}_u \mathcal{J}(\vec{u})} = \frac{\partial\,\mathrm{MOC}}{\partial\,\mathrm{SST}(x,y,t)}$$



finite difference approach



adjoint approach

Adjoint applications (III): SVD / Optimal Perturbations

- For dynamical system $\vec{v}(t) = M\vec{v}(0)$, with $\vec{v}_0 = \vec{u}$, find initial conditions $\vec{u}$, such that model state $\vec{v}(t)$ maximizes a chosen norm $\langle\, . \ , \ . \,\rangle_{v(t)} = X_t$ with norm X_0 unity at time $t = 0$:

$$\max_{\vec{u}} \Big\{ (M\vec{u})^T \cdot X_t \cdot M\vec{u} - \lambda \big[\vec{u}^T \cdot X_0 \cdot \vec{u} - 1 \big] \Big\}$$

Leads to generalized eigenvalue problem:

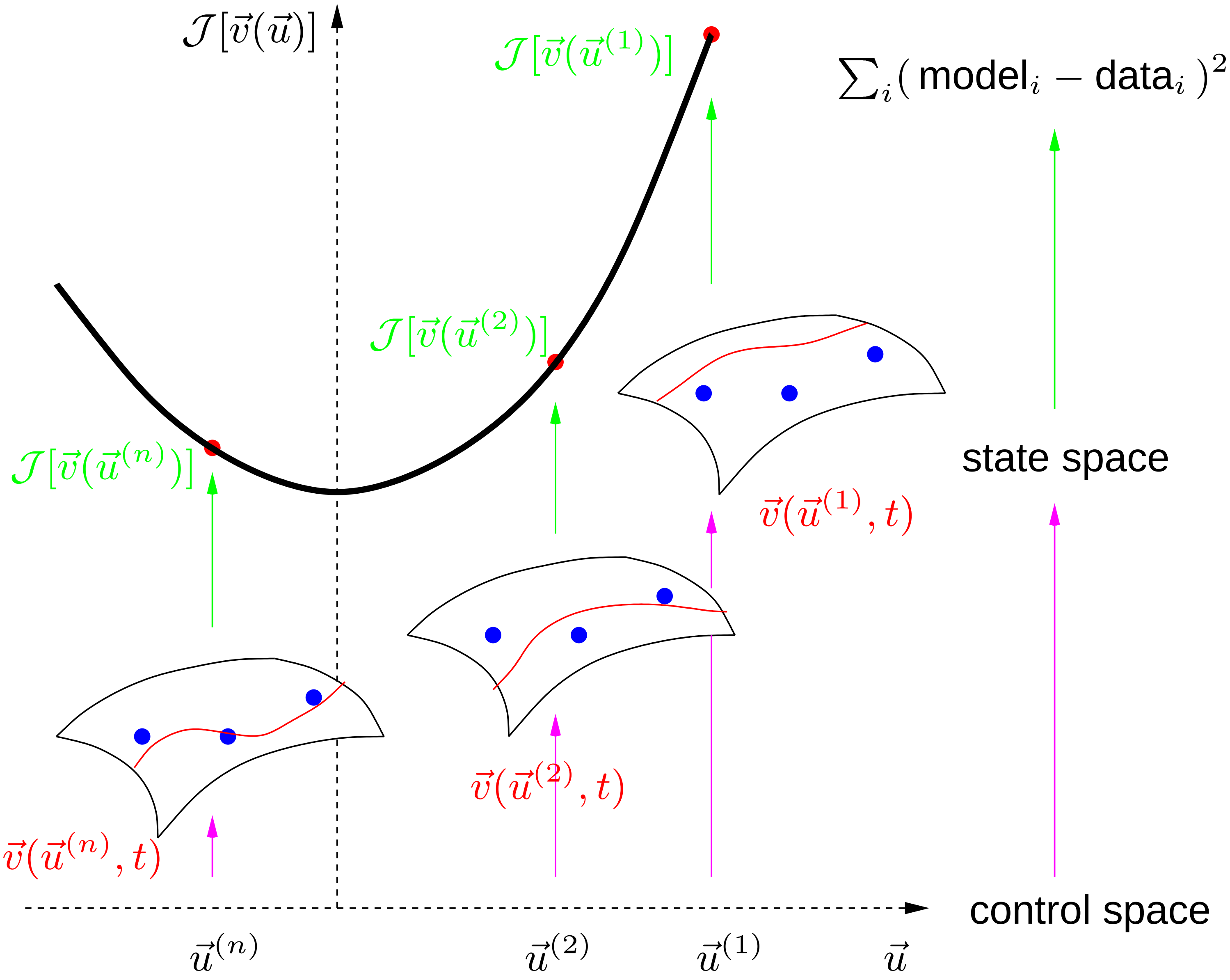
$$M^T \cdot X_t \cdot M \cdot \delta\vec{u} = \lambda X_0 \cdot \delta\vec{u}$$

Thus, in terms of the tangent linear and adjoint operator, this reads

$$\mathtt{ADM} \cdot \mathtt{X_t} \cdot \mathtt{TLM} \cdot \delta\mathtt{u} = \lambda\ \mathtt{X_0} \cdot \delta\mathtt{u}$$

$\longrightarrow$ iterative solution for optimum $\delta\vec{u}_0$

Iterative optimization via gradient descent


$\mathcal{J}[\vec{v}(\vec{u})]$
$\mathcal{J}[\vec{v}(\vec{u}^{(1)})]$
$\sum_i (\text{ model}_i - \text{data}_i\ )^2$
$\mathcal{J}[\vec{v}(\vec{u}^{(2)})]$
$\mathcal{J}[\vec{v}(\vec{u}^{(n)})]$
state space
$\vec{v}(\vec{u}^{(1)}, t)$
$\vec{v}(\vec{u}^{(2)}, t)$
$\vec{v}(\vec{u}^{(n)}, t)$
control space
$\vec{u}^{(n)}$
$\vec{u}^{(2)}$
$\vec{u}^{(1)}$
$\vec{u}$

Some definitions

$\vec{u}$	independent / control variables (e.g. initial T, S, surface forcing, background diffusivites)
$\vec{v}$	model state (at time t): T, S, U, V, W
$\mathcal{M}$	(nonlinear) model operator $\mathcal{M}(\vec{u}, t) \; = \; \vec{v}(t)$
$\mathcal{J}$	objective / cost function (e.g. least-square misfit)

$\delta\vec{u}$	perturbation of independent /control variable
$\delta\vec{v}$	perturbed model state due to control perturbation
M	tangent linear operator of model operator, $M \; = \; (\, \partial\mathcal{M}_i \, / \, \partial u_j \,)$
$\delta\mathcal{J}$	cost function variation as result of control perturbation

Some algebra (i)

model $\mathcal{M}$: control space → model state space

tangent linear M: δ(control space) → δ(model state space)

Consider cost function $\mathcal{J}$

$$\begin{tikzcd}
\mathcal{J} \, : & U \arrow[r] & V \arrow[r] & \mathbb{R} \\
& \vec{u} \arrow[r, mapsto] & \vec{v} = \mathcal{M}(\vec{u}) \arrow[r, mapsto] & \mathcal{J}(\vec{u}) = \mathcal{J}(\mathcal{M}(\vec{u}))
\end{tikzcd}$$

$$\begin{tikzcd}
TLM \, : & \delta\vec{u} \arrow[r, mapsto] & \delta\vec{v} = M \cdot \delta\vec{u} \arrow[r, mapsto] & \delta\mathcal{J} = \vec{\nabla}_u \mathcal{J}^T \cdot \delta\vec{u}
\end{tikzcd}$$

with tangent linear model

$$M \; = \; \left(\frac{\partial \mathcal{M}_i}{\partial u_j} \right) \Big|_{\vec{u}_0}$$

Minimize cost $\mathcal{J}$: Seek control variables $\vec{u}$ such that gradient vanishes:

$$\vec{\nabla}_u \mathcal{J}(\vec{u}) \; = \; 0$$

Some algebra (ii)

Expansion of $\mathcal{J}$ in terms of $\delta\vec{u}$ or $\delta\vec{v}$:

$$
\begin{aligned}
\mathcal{J} \;=\; \mathcal{J}_0 + \delta\mathcal{J} \;&=\; \mathcal{J}|_{\vec{u}_0} + \left\langle \nabla_u \mathcal{J}^T \,,\, \delta\vec{u} \right\rangle + O(\delta\vec{u}^2) \\
&=\; \mathcal{J}|_{\vec{v}_0} + \left\langle \nabla_v \mathcal{J}^T \,,\, \delta\vec{v} \right\rangle + O(\delta\vec{v}^2)
\end{aligned}
$$

Then, evaluate $\delta\mathcal{J}$:

$$
\delta\mathcal{J} \;=\; \left\langle \nabla_v \mathcal{J}^T \,,\, \delta\vec{v} \right\rangle \;=\; \left\langle \nabla_v \mathcal{J}^T \,,\, M\,\delta\vec{u} \right\rangle \;=\; \left\langle M^T\,\nabla_v \mathcal{J}^T \,,\, \delta\vec{u} \right\rangle
$$

The full gradient can be inferred via the **adjoint model (ADM)**

$$
\begin{aligned}
\nabla_u \mathcal{J}^T|_{\vec{u}} &= M^T \cdot \nabla_v \mathcal{J}^T|_{\vec{v}} \\
&= M^T \cdot \delta\vec{v}^* \\
&= \delta\vec{u}^*
\end{aligned}
$$

with M^T the adjoint (transpose) of the tangent linear operator M

Some algebra (iii)

Application of the chain rule:

$$\begin{aligned}
\vec{v} &= \mathcal{M}(\vec{u}) &= \mathcal{M}_\Lambda(\mathcal{M}_{\Lambda-1}(\ldots(\mathcal{M}_\lambda(\ldots(\mathcal{M}_1(\mathcal{M}_0(\vec{u})))))))\\
\delta\vec{v} &= M\cdot\delta\vec{u} &= M_\Lambda\cdot M_{\Lambda-1}\cdot\ldots\cdot M_\lambda\cdot\ldots\cdot M_1\cdot M_0\cdot\delta\vec{u}
\end{aligned}$$

Reveals the reverse nature of the adjoint calculation:

$$\begin{aligned}
\delta\mathcal{J} &= \langle\ \nabla_v\mathcal{J}^T\ ,\ \delta\vec{v}\ \rangle\\
&= \langle\ \nabla_v\mathcal{J}^T\ ,\ M_\Lambda\cdot\ldots\cdot M_0\cdot\delta\vec{u}\ \rangle\\
&= \langle\ M_0^T\cdot\ldots\cdot M_\Lambda^T\cdot\nabla_v\mathcal{J}^T\ ,\ \delta\vec{u}\ \rangle\\
&= \langle\ \nabla_u\mathcal{J}^T\ ,\ \delta\vec{u}\ \rangle
\end{aligned}$$

The Adjoint method

Need $\vec{\nabla}_u \mathcal{J}|_{u_0}$ of $\mathcal{J}(\vec{u}_0) \in I\!R^1$ w.r.t. control variable $\vec{u} \in I\!R^m$

$$\begin{array}{rcccccc}
\hline
\mathcal{J} & : & \vec{u} & \mapsto & \vec{v} = \mathcal{M}(\vec{u}) & \mapsto & \mathcal{J}(\mathcal{M}(\vec{u})) \\
TLM & : & \delta\vec{u} & \mapsto & \delta\vec{v} = M \cdot \delta\vec{u} & \mapsto & \delta\mathcal{J} = \vec{\nabla}_u \mathcal{J} \cdot \delta\vec{u} \\
ADM & : & \delta\vec{u}^* = \vec{\nabla}_u \mathcal{J}^T & \mapsfrom & \delta\vec{v}^* & \mapsfrom & \delta\mathcal{J} \\
\hline
\end{array}$$

- $\vec{v} = \mathcal{M}(\vec{u})$ nonlinear model
- $M \;,\; M^T$ tangent linear (TLM) / adjoint (ADM)
- $\delta\vec{u} \;,\; \delta\vec{u}^*$ perturbation / dual (or sensitivity)

$$\boxed{\vec{\nabla}_u \mathcal{J}^T|_{\vec{u}} \quad = \quad M^T|_{\vec{v}} \cdot \vec{\nabla}_v \mathcal{J}|_{\vec{v}}}$$

$$\begin{array}{llrlll}
TLM : & m \,(\sim n_x n_y n_z\,) & \text{integrations} & @ & 1\cdot & \text{(\#forward)} \\
ADM : & 1 & \text{integration} & @ & \gamma\cdot & \text{(\#forward)}
\end{array}$$

Automatic differentiation (AD)

- Model code

$$\vec{v} \, = \, \mathcal{M}_\Lambda \left(\mathcal{M}_{\Lambda-1} \left(\ldots\ldots \left(\mathcal{M}_0 \left(\vec{u} \right) \right) \right) \right)$$

- Adjoint code

$$\delta\vec{u}^* \, = \, M_0^T \cdot M_1^T \cdot \ldots\ldots \cdot M_\Lambda^T \cdot \delta\vec{v}^*$$

- Automatic differentiation:

each line of code is elementary operator $\mathcal{M}_\lambda$

$\longrightarrow$ rules for differentiating elementary operations

$\longrightarrow$ yield elementary Jacobians M_λ

$\longrightarrow$ composition of M_λ's according to chain rule

yield full tangent linear / adjoint model

- TAMC/TAF source-to-source tool (Giering & Kaminski, 1998)

$$\left.\begin{array}{l} \bullet \text{ model } \mathcal{M} \\ \bullet \text{ independent } \vec{u} \\ \bullet \text{ dependent } \mathcal{J} \end{array}\right\} \xrightarrow{\text{TAMC/TAF}} \left\{\begin{array}{l} \text{TLM } M, \text{ or} \\ \text{ADM } M^T, \text{ or} \\ \text{gradient } \delta\vec{u}^* = \vec{\nabla}_u \mathcal{J} \end{array}\right.$$

Example: Eli's 3-box model of THC

```
DO t = 1, nTimeSteps
```

- calculate density

$$\rho = -\alpha T + \beta S$$

- calculate thermohaline transport

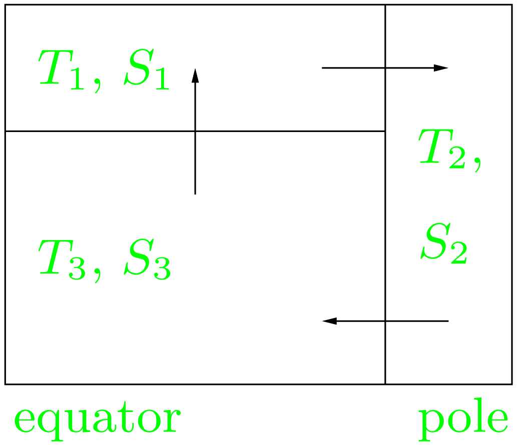
$$U = U(\rho(T,S))$$

- calculate tracer advection

$$\frac{d}{dt}Tr = f(Tr, U)$$

calculate timestepping and update tracer fields $Tr = \{T, S\}$

```
END DO
```



Focus on advection equation for T_3 (i)

$$\begin{aligned} \frac{dT_3}{dt} &= U(T_3 - T_2), \quad \text{for } U \geq 0 \\ \texttt{diffT3} &= \texttt{u} * (\texttt{T3} - \texttt{T2}) \end{aligned}$$

derivative $\delta\texttt{diffT3}$:

$$\delta\texttt{diffT3} = \frac{\partial\texttt{diffT3}}{\partial\texttt{U}}\delta\texttt{U} + \frac{\partial\texttt{diffT3}}{\partial\texttt{T}_2}\delta\texttt{T}_2 + \frac{\partial\texttt{diffT3}}{\partial\texttt{T}_3}\delta\texttt{T}_3$$

in matrix form:

$$\begin{pmatrix} \delta\texttt{diffT3} \\ \delta\texttt{T}_3 \\ \delta\texttt{T}_2 \\ \delta\texttt{U} \end{pmatrix}^{\lambda} = \begin{pmatrix} 0 & -\texttt{U} & \texttt{U} & \texttt{T3} - \texttt{T1} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \delta\texttt{diffT3} \\ \delta\texttt{T}_3 \\ \delta\texttt{T}_2 \\ \delta\texttt{U} \end{pmatrix}^{\lambda - 1}$$

Focus on advection equation for T_3 (ii)

Transposed relationship yields:

$$\begin{pmatrix} \delta^*\mathtt{diffT3} \\ \delta^*\mathtt{T}_3 \\ \delta^*\mathtt{T}_2 \\ \delta^*\mathtt{U} \end{pmatrix}^{\lambda-1} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ -\mathtt{U} & 1 & 0 & 0 \\ \mathtt{U} & 0 & 1 & 0 \\ \mathtt{T3}-\mathtt{T1} & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \delta^*\mathtt{diffT3} \\ \delta^*\mathtt{T}_3 \\ \delta^*\mathtt{T}_2 \\ \delta^*\mathtt{U} \end{pmatrix}^{\lambda}$$

and thus adjoint code:

```
adT3     = adT3        - u*addiffT3
adT2     = adT2        + u*addiffT3
adU      = adu   + (T3-T2)*addiffT3
addiffT3 = 0
```

Note: state $\mathtt{T2}, \mathtt{T3}, \mathtt{U}$ are required to evaluate derivative at each time step, in reverse order!

⟶ *TANGENT* linearity

Reverse order integration (i)

```
DO istep = 1, nTimeSteps
  • call density($\rho$)
  • call transport($U$)
  • call timestep($T, S$)
  • call update($T, S$)
END DO
```

```
DO istep = nTimeSteps, 1, -1
C recompute required variables
  • DO iaux = 1, istep
      – call density($\rho$)
      – call transport($U$)
      – call timestep($T, S$)
      – call update($T, S$)
    END DO
C perform adjoint timestep
  • call adupdate($T, S$)
  • call adtimestep($T, S$)
  • call adtransport($U$)
  • call addensity($\rho$)
END DO
```

Reverse order integration (ii)

- *Adjoint = transpose of TLM*

→ evaluated in reverse order

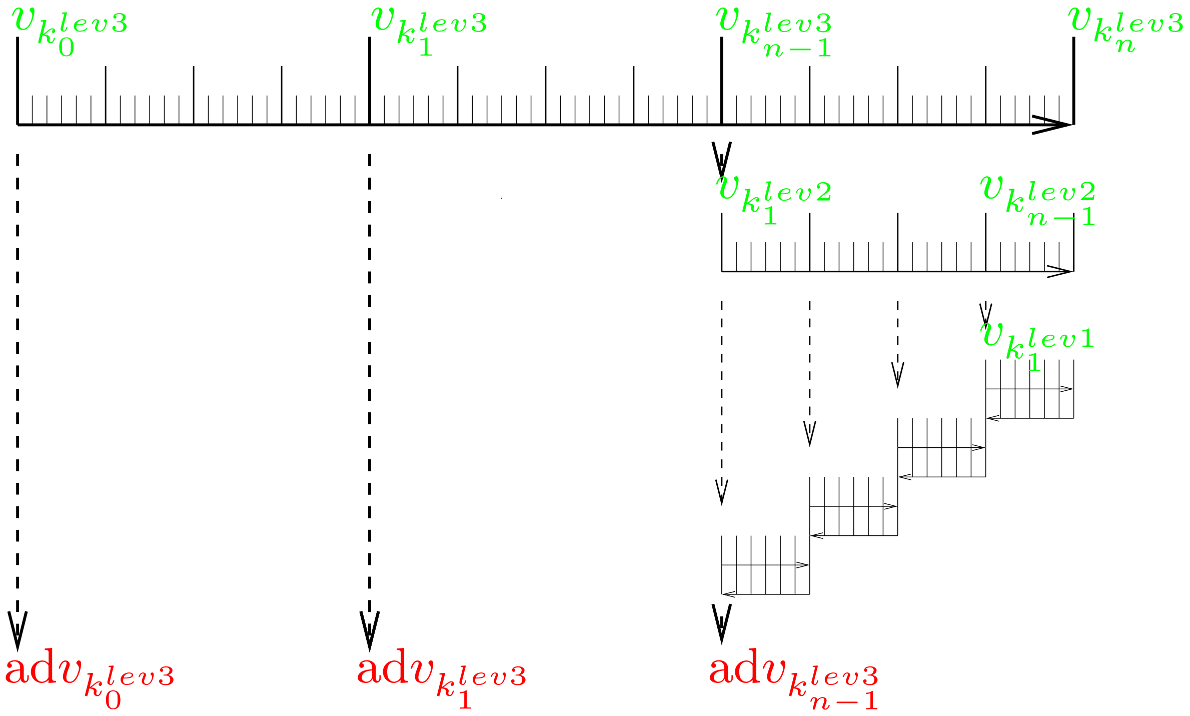
→ model state at every time step required in reverse

→ all state stored or recomputed

- *Solution: Checkpointing*

(e.g. Griewank, 1992)

balances storing vs. recomputation



Reverse order integration (iii)

```
DO iOuter = 1, nOuter
   • CADJ STORE $T, S \longrightarrow$ disk
   • DO iInner = 1, nInner
        – call density($\rho$)
        – call transport($U$)
        – call timestep($T, S$)
        – call update($T, S$)
     END DO
END DO
```

```
DO iOuter = nOuter, 1, -1
   • CADJ RESTORE $T, S \longleftarrow$ disk
   • DO iInner = 1, nInner
        – call density($\rho$)
        – call transport($U$)
        – CADJ STORE $T, S, U \longrightarrow$ common
        – call timestep($T, S$)
        – call update($T, S$)
     END DO
   • DO iInner = nInner, 1, -1
        – call adupdate($\mathrm{ad}T, \mathrm{ad}S$)
        – call adtimestep($\mathrm{ad}T, \mathrm{ad}S$)
        – CADJ RESTORE $T, S, U \longleftarrow$ common
        – call adtransport($\mathrm{ad}U$)
        – call addensity($\mathrm{ad}\rho$)
     END DO
END DO
```

Reverse order integration (iv)

- e.g. 3-level checkpointing:

$$n_{TimeSteps} = n_1 \cdot n_2 \cdot n_3$$

$\rightarrow$ **Storing:** reduced from $n_1 \cdot n_2 \cdot n_3$ to

- disk: $n_2 + n_3$,
- memory: n_1

$\rightarrow$ **CPU:** $3 \cdot \mathrm{forward} + 1 \cdot \mathrm{adjoint} \ \approx \ 5.5 \cdot \mathrm{forward}$

- Closely related to **adjoint dump & restart** problem.
  Available queue sizes at HPC Centres may be limited

- Insertion of store directive requires detailed knowledge of code *and* AD tool behaviour
  $\longrightarrow$ not straightforward ("semi-automatic" differentiation only)

Input/output – active file handling

I/O of active variables is common and needs to be accounted for in derivative

READ assigning a value to a variable

WRITE referencing a variable

code	hypothetical code	adjoint hypothetical code	adjoint code
`OPEN(8)`		`ADXD = 0.`	`OPEN(9)`
⋮	⋮	⋮	⋮
`WRITE(8) X`	`XD = X`	`ADXD = ADXD + ADZ`	`WRITE(9) ADZ`
		`ADZ = 0.`	`ADZ = 0.`
⋮	⋮	⋮	⋮
`READ(8) Z`	`Z = XD`	`ADX = ADX + ADXD`	`READ(9) ADXD`
		`ADXD = 0.`	`ADX = ADX + ADXD`
			`ADXD = 0.`
⋮	⋮	⋮	⋮
`CLOSE(8)`			`CLOSE(9)`

(from *Giering & Kaminski (1998)*)

Test / assure correctness of adjoint-derived gradient

Procedure to check that automatically derived gradient G_i^{ad} is correct; consider perturbation of i-th control vector element u_i and $\Delta u_i = \delta_{ij}$

finite difference vs. adjoint	tangent linear vs. adjoint
$G_i^{fd} = \frac{\mathcal{J}(u_i+\epsilon)-\mathcal{J}(u_i-\epsilon)}{2\epsilon}$	$G_i^{tl} = \vec{\nabla}_u \mathcal{J} \cdot \Delta\vec{u} = \left(\vec{\nabla}_u \mathcal{J}\right)_i$
$R_i^{fd} = 1 - \frac{G_i^{fd}}{G_i^{ad}}$	$R_i^{tl} = 1 - \frac{G_i^{tl}}{G_i^{ad}}$

→ can test ’correctness’ of adjoint gradient G_i^{ad}

→ can test ’time horizon’ of linearity assumption

Extension to vector-valued “cost”

For $\vec{\mathcal{J}}(\mathcal{M}(\vec{u})) \in \mathbb{R}^m$ previous expression for gradient is generalized to

$$M^T \cdot d_v \vec{\mathcal{J}}^T \cdot \delta \vec{\mathcal{J}} \; = \; d_u \vec{\mathcal{J}}^T \cdot \delta \vec{\mathcal{J}}$$

Thus, with $\delta \vec{u} \in \mathbb{R}^n$ and $\delta \vec{\mathcal{J}} \in \mathbb{R}^m$,

$$\left.\begin{array}{l} m \text{ adjoint} \\ n \text{ tangent linear} \end{array}\right\} \text{ runs for each perturbation } \left\{\begin{array}{rcll} (\delta \vec{\mathcal{J}})_j &=& \delta_{ij}, & j = 1, \ldots, m \\ (\delta \vec{u})_j &=& \delta_{ij}, & j = 1, \ldots, n \end{array}\right.$$

to recover full derivative $d_u \vec{\mathcal{J}}^T$.

$$\longrightarrow \quad \left.\begin{array}{c} \text{ADM} \\ \text{TLM} \end{array}\right\} \text{ is preferable, for } \left\{\begin{array}{rcl} n &>& m \\ n &<& m \end{array}\right.$$

Examples

- Least-square model-vs.-data misfit: $\mathcal{J} = \langle \mathcal{H}(\vec{v}) - \vec{d}\,, \mathcal{H}(\vec{v}) - \vec{d} \rangle$

  e.g. for state estimation / data assimilation

  with $\vec{d}$: data vector,

  $\mathcal{H}$: projector of model state space onto data space

$$\nabla_v \mathcal{J}^T = 2\ H \cdot \Big(\mathcal{H}(\vec{v}) - \vec{d} \Big)$$

$$\Big(\nabla_v \mathcal{J}^T \Big)_j = 2 \left\{ \sum_k \frac{\partial \mathcal{H}_k}{\partial v_j} \left(\mathcal{H}_k(\vec{v}) - d_k \right) \right\}$$

- Final state: $\vec{\mathcal{J}} = \vec{v}$

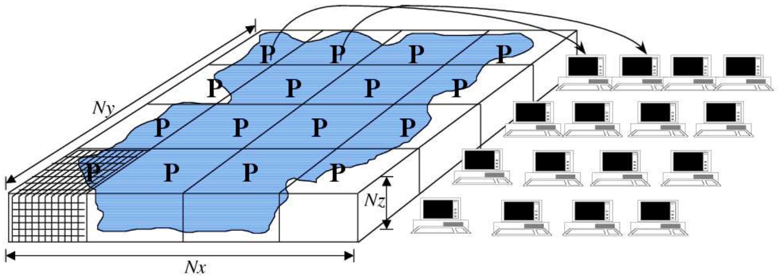
  e.g. for SVD calculation, source-sink estimation

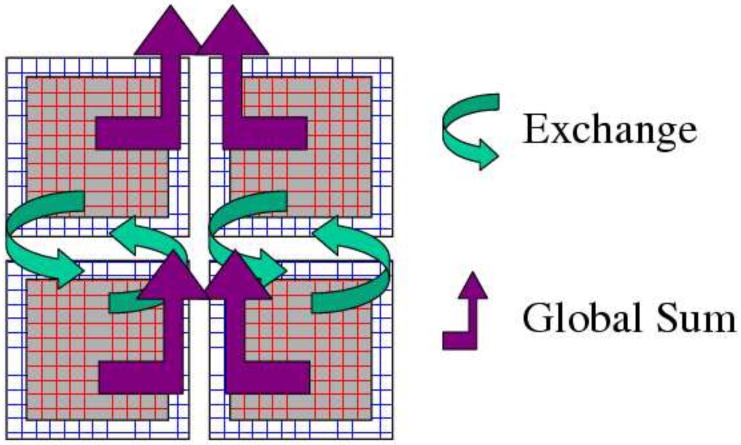
$$d_v \vec{\mathcal{J}}^T = \mathrm{Id}$$

$$d_u \vec{\mathcal{J}}^T = M^T \cdot d_v \vec{\mathcal{J}}^T = M^T$$

AD challenges: (II) Scalability

- domain decomposition (tiles) & overlaps (halos)
- split into extensive on-processor and global phase





Global communication/arithmetic op.'s supported by MITgcm's intermediate layer (WRAPPER) which need hand-written adjoint forms

	operation/primitive	forward		reverse
•	communication (MPI,...):	send	⟷	receive
•	arithmetic (global sum,...):	gather	⟷	scatter
•	active parallel I/O:	read	⟷	write

AD challenges: (III) Parameterization schemes

- Parameterization schemes required for
  - turbulence closure for nonlinear Navier-Stokes equations
  - subgrid-scale processes
- Main issues:
  - nonlinear expressions
    (momentum advection, nonlinear equation of state, ...)
  - State-dependent conditional statements (thresholds and jumps)
  - small numbers `(SQRT(.), 1/x, ...)`; non-differentiable points
  - combinations thereof
- e.g.: many forms of mixing processes
  (shear instability, convection, diffusion, ...)
- Require significant attention to obtain stable adjoint solution!

N.B.: Currently, common method to circumvene these problems is to exclude exact adjoint of parameterization schemes

Example: Limits on boundary layer depth in KPP

- for neutral stratification: H_{bl} should be smaller than both h_e and L
- generally: H_{bl} should be larger than some minimal value

```
   CADJ STORE hbl, bfsfc TO TAPE
   do i = 1, Nx * Ny
     if (bfsfc(i).gt.0.0) then
         hekman  =  cekman * ustar(i) / max(abs(Coriol(i)), eps)        $h_e = 0.7u^*/f$
         hmonob  =  cmonob * ustar(i) * ustar(i) * ustar(i)             $L = {u^*}^3/(\kappa B_f)$
 &                   / vonk / bfsfc(i)
         hlimit  =                                                      $h_{bl}$ limit for:
 &                   stable(i) * min(hekman, hmonob)                    – stable case
 &                   + (stable(i) - 1.) * zgrid(Nr)                     – unstable case

         hbl(i)  =  min(hbl(i), hlimit)                                 apply upper limit
     end if
     hbl(i)  =  max(hbl(i), minKPPhbl)                                  apply lower limit
   end do
```

Example: Limits on boundary layer depth in KPP (cont’d)

```
CADJ RESTORE hbl_1   FROM TAPE
CADJ RESTORE bfsfc   FROM TAPE

do i = 1, Nx*Ny
  adhbl(i) = adhbl(i)*(0.5+sign(0.5d0,hbl_2(i)-minkpphbl))
  if (bfsfc(i) .gt. 0.) then
!      recompute hekman, hmonob, hlimit
    adhlimit = adhlimit+adhbl(i)*(0.5-sign(0.5d0,hlimit-hbl_1(i)))
    adhbl(i) = adhbl(i)*(0.5+sign(0.5d0,hlimit-hbl_1(i)))
    adhekman = adhekman+adhlimit*stable(i)*(0.5+sign(0.5d0,hmonob-hekman))
    adhmonob = adhmonob+adhlimit*stable(i)*(0.5-sign(0.5d0,hmonob-hekman))
    adstable(i) = adstable(i)+adhlimit*(zgrid(nr)+min(hekman,hmonob))
    adhlimit = 0.d0
    ...
  endif
end do
```

ECCO state estimation: Control variables

- Initial values (temperature, salinity, passive tracer)
- time-dependent surface forcing
  (either air-sea fluxes or atmospheric fields plus bulk formulae)
  - heat flux (or surface air temperature)
  - freshwater flux (or atmos. humidity)
  - zonal/meridional windstresses (or surface wind speeds)
- time-dependent open-boundary values
- background mixing coefficient
- Eliassen-Palm fluxes
- bottom topography (in progress)
- ...

ECCO state estimation: observational elements

$$
\begin{array}{rcll}
\mathcal{J} & = & (\bar{\eta} - \bar{\eta}_{TP})^t \, \mathbf{W_{geoid}} \, (\bar{\eta} - \bar{\eta}_{TP}) & \text{TOPEX absolute SSH} \\
 & & + \, (\eta - \eta'_{TP})^t \, \mathbf{W_{TP}} \, (\eta - \eta'_{TP}) & \text{TOPEX SSH anomalies} \\
 & & + \, (\eta - \eta'_{ERS})^t \, \mathbf{W_{ERS}} \, (\eta - \eta'_{ERS}) & \text{ERS SSH anomalies} \\
 & & + \, (\bar{T}_{surf} - \bar{T}_{Reyn})^t \, \mathbf{W_{SST}} \, (\bar{T}_{surf} - \bar{T}_{Reyn}) & \text{Reynolds SST} \\
 & & + \, (\bar{T} - \bar{T}_{Lev})^t \, \mathbf{W_{T_{Lev}}} \, (\bar{T} - \bar{T}_{Lev}) & \text{Levitus clim.} \\
 & & + \, (\bar{S} - \bar{S}_{Lev})^t \, \mathbf{W_{S_{Lev}}} \, (\bar{S} - \bar{S}_{Lev}) & \text{Levitus clim.} \\
 & & + \, (\tau_x - \tau_{x,NCEP})^t \, \mathbf{W}_{\tau_\mathbf{x}} \, (\tau_x - \tau_{x,NCEP}) & \text{zonal wind stress} \\
 & & + \, (\tau_y - \tau_{y,NCEP})^t \, \mathbf{W}_{\tau_\mathbf{y}} \, (\tau_y - \tau_{y,NCEP}) & \text{merid. wind stress} \\
 & & + \, (H_Q - H_{Q,NCEP})^t \, \mathbf{W_{H_Q}} \, (H_Q - H_{Q,NCEP}) & \text{NCEP heat flux} \\
 & & + \, (H_F - H_{F,NCEP})^t \, \mathbf{W_{H_F}} \, (H_F - H_{F,NCEP}) & \text{NCEP freshwater flux}
\end{array}
$$

Currently added:

- Jason-1 altimetry (sea surface height)
- WOCE hydrography, XBT, TAO buoys
- PALACE/ARGO tracer profiles and drift velocities
- surface drifter velocities
- NSCAT/QuickScat surface wind stress fields
- TRMM/TMI tropical surface temperature fields

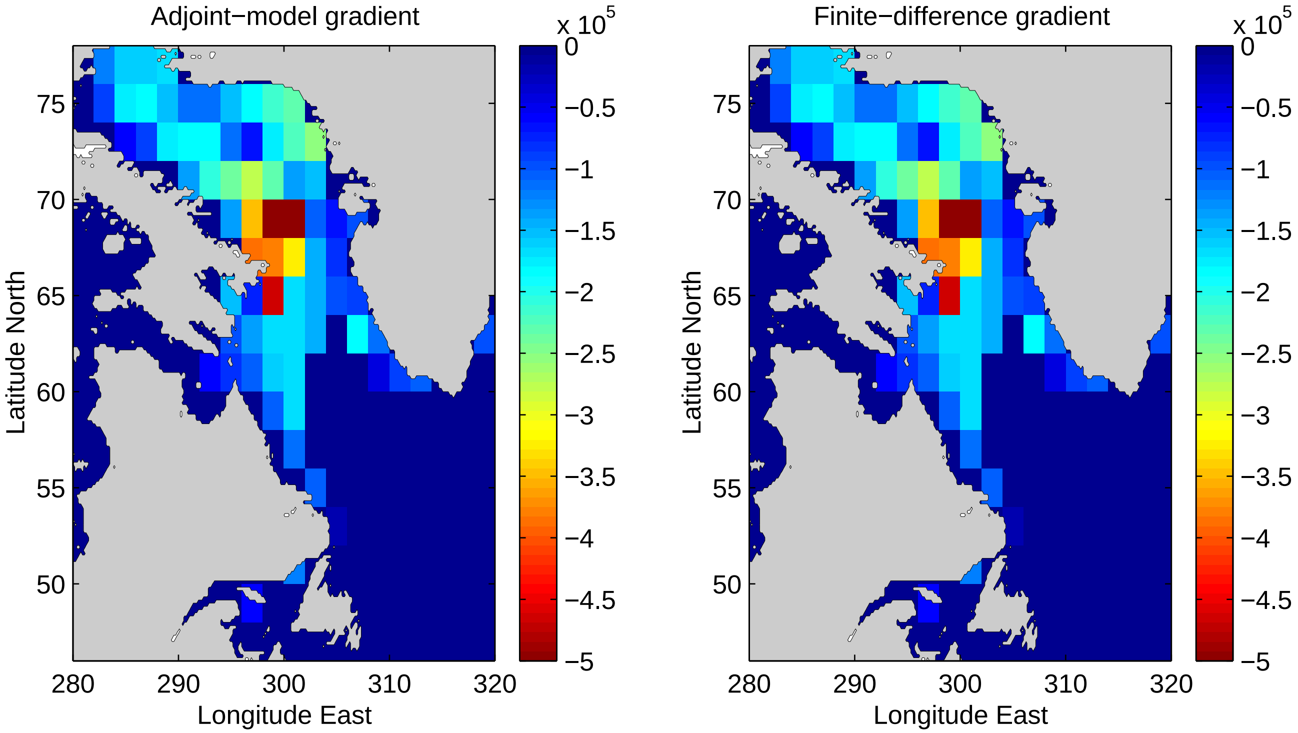
ECCO state estimation: problem size

- Dimensionality:
  - grid @ 1°×1° resolution: $n_x \cdot n_y \cdot n_z$ = 360 · 160 · 23 $1,324,800$
  - model state: 17 3D + 2 2D fields $\sim 2 \cdot 10^7$
  - timesteps: 10 years @ 1-hour time step $87,600$
  - control vector $\sim 1 \cdot 10^8$
    - initial temperature (T), salinity (S)
    - time-dependent surface forcing (every 2 days)
  - cost function: observational elements: $\sim 1 \cdot 10^8$
- Computational size:
  - 60 processors (15 nodes) @ 512MB per proc.
  - I/O: 10 GB input, 35GB output
  - time: 59 hours per iteration @ 60 processors
- What we would ideally want:
  - 1/10°×1/10° resol., 1000 years, full model error covariance ...

Sea-ice model description

- Sea-ice model based on *Hibler (1979,1980)* has been added to MIT/ECCO adjoint-model estimation infrastructure.
- Snow is simulated as per *Zhang et al. (1998)*.
- Dynamics are governed by viscous-plastic rheology and solved with numerical method of *Zhang and Rothrock (2000)*.
- Relatively simple 2-category sea-ice model is chosen to to simplify adjoint-model development and to reduce computational cost.
- References:
  - W. D. Hibler, III, 1979. A dynamic thermodynamic sea ice model. J. Phys. Oceanogr., 9:815.
  - W. D. Hibler, III, 1980. Modeling a variable thickness sea ice cover. Mon. Wea. Rev., 1:1943.
  - J. Zhang, W. D. Hibler, III, M. Steele, and D. A. Rothrock, 1998. Arctic ice-ocean modeling with and without climate restoring. J. Phys. Oceanogr., 28:191.
  - J. Zhang and D. A. Rothrock, 2000. Modeling arctic sea ice with an efficient plastic solution. J. Geophys. Res., 105:3325.

Coupled ocean-sea-ice model: adjoint sensitivity experiment



- Preliminary test of the coupled model (sea-ice thermodynamics only).
- Sensitivity of sea-ice volume in the Labrador Sea to surface atmospheric temperature perturbations over a 4 hour integration period in units of m^3/$^\circ$C.
- Left panel: adjoint-derived gradient,
  right panel: perturbation of surface atmospheric temperature at each location.
- The small difference between the two panels, less than one part in 10^5, demonstrates the accuracy of the adjoint-model solution.

Coupled ocean-sea-ice model: work in progress

- The coupled ocean-sea-ice adjoint model provides accurate results in the small test domain only for up to 10-day integrations.

  For longer integrations the forward-model gradient is ill-defined. Therefore it cannot be computed using the adjoint method.

  Work is underway to simplify the sea-ice adjoint model in order to permit longer integrations.

- Accuracy of dynamic solver needs to be increased and computational cost decreased for adjoint-model computations.

- Projection operators for sea ice data need to be written, and corresponding a priori errors determined.

Other issues

- impact of *nonlinearities, discontinuities* need further analyses
  What are limits of applicability for high-resolution, long-term integrations?
- cost regularization, gradient preconditioning, and model error analysis
  $\longrightarrow$ compute 2nd derivative (Hessian)
- *bulk formulae:* flux controls vs. atmospheric state controls
- *atmospheric setup:*
  - dynamic setup (Held-Suarez like) is adjointed
  - cubed-sphere needs update of adjoint components of WRAPPER
  - adjoint of physics packages will require (quite) some work
  - coupling
- *ESMF / PRISMA:* adjoint correspondents of coupler primitives

Conclusions & Outlook – Science aspects

- Global ocean state estimation via adjoint method is feasible,
  - yields a dynamical consistent model trajectory and parameter estimates
  - uses data in an ’optimal’ way
  - yields a quantitative measure of model vs. data misfit
- State estimation at the eddy-resolving scale remains subject to research/discussion
- Incorporation of different data sets can reveil new insights into model vs. data incompatibility/inconsistency (and its causes)
- Careful analysis of ’adjoint state’ yields complementary info on model behaviour (assertions on which can be tested)
- Inclusion of model error remains untackled problem
- Sea ice state estimation has just started; despite remaining problems with parameterizations, first results look promising; is expected to greatly enhance current state estimation system.

Conclusions & Outlook – Adjoint code generation

- *Exact scalable adjoint* code generation is feasible
- AD tools are indispensable for *evolving code* environment
- Nevertheless, *challenges remain* despite AD
  efficient adjoint code generation is **semi-automatic** so far
- Code development should have AD “in mind”
- Libraries/couplers need derivative forms
- Room for AD tool improvement has led to
  NSF Information Technology Research (ITR) project:
  Adjoint Compiler Technology & Standard (ACTS)

  common platform that should facilitate AD tool
  development/improvement by larger community