Inverse Problems, Inverse Methods, State Estimation, Data Assimilation, and All That: A Short Course in Five Lectures

January 2013. Lecture 1

Carl Wunsch
Earth and Planetary Sciences, Harvard University

January 24, 2013

The motivation for many people here is undoubtedly what the meteorologists call "data assimilation."
Example, Kalnay et al. (1996)

The NCEP/NCAR 40-Year Reanalysis Project

E. Kalnay,* M. Kanamitsu,* R. Kistler,* W. Collins,* D. Deaven,* L. Gandin,*
M. Iredell,* S. Saha,* G. White,* J. Woollen,* Y. Zhu,* M. Chelliah,* W. Ebisuzaki,*
W. Higgins, J. Janowiak,* K. C. Mo, C. Ropelewski,* J. Wang,*
A. Leetmaa,* R. Reynolds,* Roy Jenne,* and Dennis Joseph*

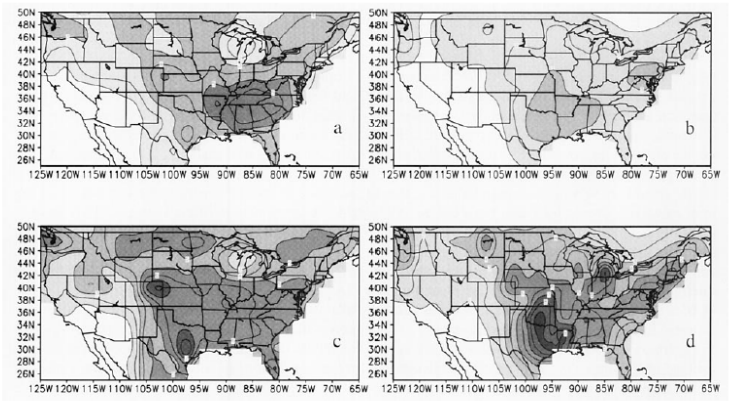


FIG. 8. Daily mean precipitation rates (mm day^{-1}) for May 1985–89 over the United States in (a) the NCEP/NCAR reanalysis and in (b) the observations. Standard deviation of the daily mean precipitation rates (mm day^{-1}) within May 1985–89 in (c) the NCEP/NCAR reanalysis and (d) the observations. Contour interval is 1 mm day^{-1}, greater than 1 mm day^{-1} is shaded.

Cited 9500+ times. What is this? How is it done? Citation index gives about 9000 papers on "data assimilation" (June 2012).

RMetS
Royal Meteorological Society

Review Article
The Twentieth Century Reanalysis Project

G. P. Compo,[a,b]* J. S. Whitaker,[b‡] P. D. Sardeshmukh,[a,b] N. Matsui,[a,b] R. J. Allan,[c†] X. Yin,[d] B. E. Gleason, Jr.,[e‡] R. S. Vose,[e‡] G. Rutledge,[e‡] P. Bessemoulin,[f] S. Brönnimann,[g,h] M. Brunet,[i,l] R. I. Crouthamel,[j] A. N. Grant,[g] P. Y. Groisman,[e,k] P. D. Jones,[l] M. C. Kruk,[d] A. C. Kruger,[m] G. J. Marshall,[n] M. Maugeri,[o] H. Y. Mok,[p] Ø. Nordli,[q] T. F. Ross,[r‡] R. M. Trigo,[s] X. L. Wang,[t] S. D. Woodruff,[b‡] and S. J. Worley[u]

[a]*University of Colorado, CIRES, Climate Diagnostics Center, Boulder, CO, USA*
[b]*NOAA Earth System Research Laboratory, Physical Sciences Division, Boulder, CO, USA*
[c]*ACRE Project, Hadley Centre, Met Office, Exeter, UK*
[d]*STG Inc., Asheville, NC, USA*
[e]*NOAA National Climatic Data Center, Asheville, NC, USA*
[f]*MétéoFrance, Toulouse, France*
[g]*ETH Zurich, Switzerland*
[h]*Oeschger Center for Climate Change Research, University of Bern, Switzerland*
[i]*Centre for Climate Change, Universitat Rovira i Virgili, Tarragona, Spain*
[j]*International Environmental Data Rescue Organization, Deale, MD, USA*
[k]*University Corporation for Atmospheric Research, Boulder, CO, USA*
[l]*Climatic Research Unit, University of East Anglia, Norwich, UK*
[m]*South African Weather Service, Pretoria, South Africa*
[n]*British Antarctic Survey, Cambridge, UK*
[o]*Dipartimento di Fisica, Università degli Studi di Milano, Milano, Italy*
[p]*Hong Kong Observatory, Hong Kong, China*
[q]*Norwegian Meteorological Institute, Oslo, Norway*
[r]*NOAA Climate Database Modernization Program, NCDC, Asheville, NC, USA*
[s]*Centro de Geofísica da Universidade de Lisboa, IDL, University of Lisbon, Portugal*
[t]*Environment Canada, Toronto, Ontario, Canada*
[u]*National Center for Atmospheric Research, Boulder, CO, USA*
*Correspondence to: Gilbert P. Compo, 325 Broadway R/PSD1, Boulder, CO USA 80305-3328.
E-mail: compo@colorado.edu, gilbert.p.compo@noaa.gov

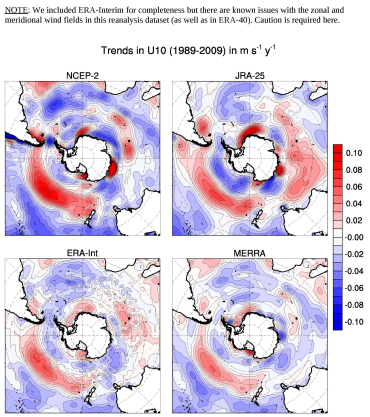
†The contribution of R. J. Allan was written in the course of his employment at the Met Office, UK, and is published with the permission of the Controller of HMSO and the Queen's Printer for Scotland.

‡The contributions of these authors were prepared as part of their official duties as US Federal Government employees.

The Twentieth Century Reanalysis (20CR) project is an international effort to produce a comprehensive global atmospheric circulation dataset spanning the twentieth century, assimilating only surface pressure reports and using observed monthly sea-surface temperature and sea-ice distributions as boundary conditions. It is chiefly motivated by a need to provide an observational dataset with quantified uncertainties for validations of climate model simulations of the twentieth century on all time-scales, with emphasis on the statistics of daily weather. It uses an Ensemble Kalman Filter data assimilation method with background 'first guess' fields supplied by an ensemble of forecasts from a global numerical weather prediction model. This directly yields a global analysis every 6 hours as the most likely state of the atmosphere, and also an uncertainty estimate of that analysis.

The 20CR dataset provides the first estimates of global tropospheric variability, and of the dataset's time-varying quality, from 1871 to the present at 6-hourly temporal and 2° spatial resolutions. Intercomparisons with independent radiosonde data indicate that the reanalyses are generally of high quality. The quality in the extratropical Northern Hemisphere throughout the century

Now a considerable industry...

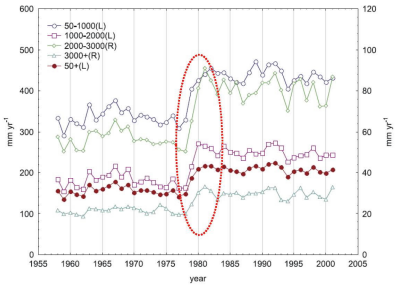


Calculated trends in southern hemisphere winds. Same data, very similar models, similar methods. Why are they so different? Do you believe one of them is best? Why?

Background

Mean annual Antarctic net precipitation (P-E) from ERA-40 reanalysis for various elevation areas.

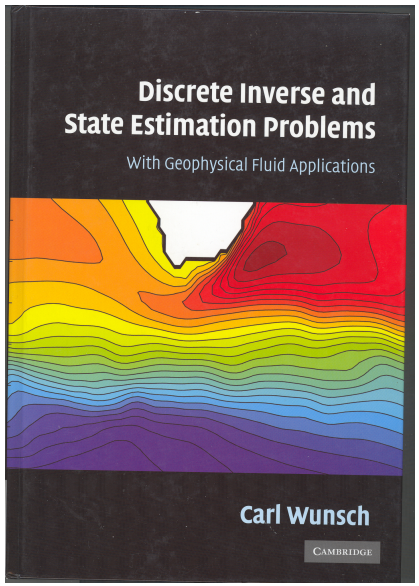
[Bromwich et al. 2007, adapted from Van de Berg et al. 2005]



- Spurious trends in the high latitudes resulting from changes in the observing system, especially the assimilation of satellite observations in the late 1970s.
- Jump in Antarctic P-E in 1978-79, particularly marked at high elevations.

Figure: From D. Bromwich. The "jump" with the new data type shows that (A) either something is very wrong with the procedure and/or, (B) the uncertainty of the estimates exeeds the magnitude of the jump.

This short course is intended to be a high speed skim through the first part of C. Wunsch, *Discrete Inverse and State Estimation Problems*, Cambridge, 2006. (DISEP. Will not cover the oceanography in the last chapters.)



Approach is based upon the inference that the basic ideas are really very simple, but in practice are obscured by a fog of jargon and unnecessary mathematics. I intend only to convey the basic ideas.

For those interested in a more mathematical approach, good starting points are R. L. Parker, *Geophysical Inverse Theory,* 1996, Princeton Un. Press, 377pp., which leads into the methods pioneered by George Backus and Freeman Gilbert and which dominate the solid earth literature.
J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations, 1971,* Springer-Verlag, 396 pp. The entire business is perhaps best looked upon as an application of methods borrowed from control theory.

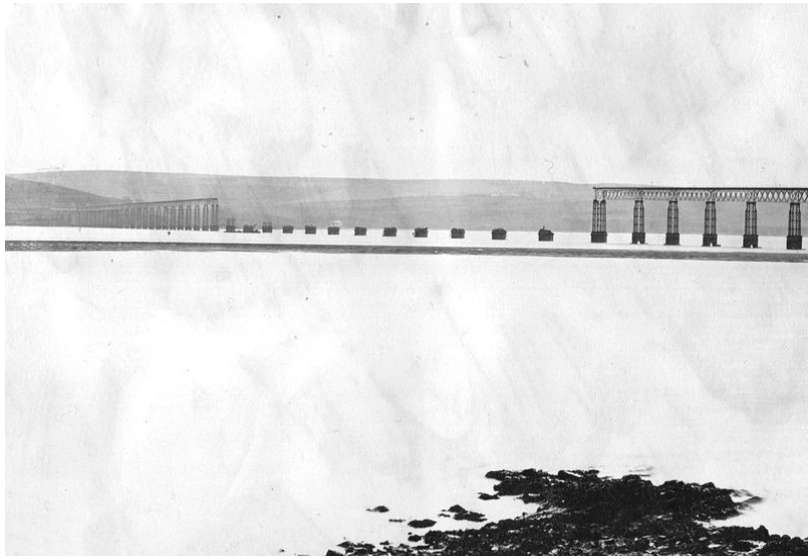
Consider the physics and construction of a bridge:

One can understand the basic principles in terms of levers, loads, stress-strain relations etc. and have an understanding of how any particular design works and why—enough to give confidence to cross on it. Actually building a real bridge is a major *engineering* problem involving such things as the detailed geology of the river bed and banks, the various strengths of different steel and concrete types, costs, volume of traffic,construction logistics etc. A great challenge.

Of course, someone can come up with a modified design that requires going back and working it through:

Sometimes unexpected things happen! (See Youtube for an animation – the Tacoma Narrows bridge, http://www.youtube.com/watch?v=j-zczJXSxnw.) Are in a parameter range that was not anticipated:

A bridge is a full-scale, analogue, model of the hypothetical equations and never exactly conforms to them (e.g., stress-strain relationships are always approximations).



The Tay Bridge. And the Millennium Bridge,....

The goal here is to make the basic concepts clear by discussing mainly very simple examples. There is a huge amount of numerical engineering involved in building an estimation system, including all of the details of computational efficiency, IO rates, parallelization, numerical stability, checkpointing, calibration, etc. Some very clever people have been involved in the engineering effort—it isn't easy. But you should be able to understand the basic principles without slogging through questions such as the relative efficiency of storage versus recomputation—which may be the key to practicality.

BASIC NOTIONS:

Everything is eventually put onto a computer. That means all real problems are discrete and of finite dimension. We work in finite dimensional vector spaces, not infinite dimensional Hilbert or Banach spaces.

Essentially all methodologies and problems are solutions, exact or approximate, to simultaneous algebraic equations, linear or nonlinear.

ASSUMED BASIC KNOWLEDGE:
Elementary linear algebra, including matrices, $\mathbf{A}$, vectors, $\mathbf{y}$, transposes, $\mathbf{A}^T$, $\mathbf{y}^T$, inverses, $\mathbf{A}^{-1}$, where it exists, and the eigenvalue, eigenvector problem,

$$\mathbf{B}\mathbf{d}_i = \lambda_i \mathbf{d}_i$$

A bit of intuition about vector spaces. Here, matrices are bold-face upper case, $\mathbf{B}$, and vectors are bold-face lower case, $\mathbf{d}$, and by default are column vectors. Differentiation rules such as

$$\frac{\partial \left(\mathbf{a}^T \mathbf{b}\right)}{\partial \mathbf{b}} = \mathbf{a}, \; \frac{\partial \left(\mathbf{a}^T \mathbf{b}\right)}{\partial \mathbf{a}} = \mathbf{b}, \; \frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}^T, \; \frac{\partial \mathbf{x}^T \mathbf{A}}{\partial \mathbf{x}} = \mathbf{A}$$

(the first two imply the second pair). The size (*norm*) of a matrix is assumed to be known in some useful definition.
The matrix inversion lemma. Several forms. One is,

$$\left[\mathbf{C} - \mathbf{B}^T \mathbf{A} \mathbf{B}^T\right]^{-1} = \left[\mathbf{I} - \mathbf{C}^{-1} \mathbf{B}^T \mathbf{A}^{-1} \mathbf{B}\right]^{-1} \mathbf{C}^{-1}$$

(see DISEP, P. 29).

Elementary statistics at the level of sample and theoretical means, variances, bias, etc.

Simple analytical solutions of conventional linear differential and partial differential equations $\left(\nabla^2 \phi = \rho\right)$.

Very basic methods of discretization (simple differences). Many interesting extensions are possible.

Am going to try to convince you that in practice it can all (including the "reanalyses") be understood from consideration of the finding of useful approximate solutions to simultaneous equations, with conventional least-squares as commonly the method of choice.
(An immense jargon surrounds the subject: forward problems, inverse problems, inverse methods, resolution, Kalman filters, Gauss-Markov Theorem, RTS smoothers, adjoints, 4DVAR, 3DVAR, Pontryagin Principle, singular vectors, EOFs, ensemble filters,...) Most of the complexity has nothing to do with concepts, and more with coping with efficiency issues. Less to it than meets the eye!

Some necessary terminology (jargon).
A *forward problem*:
Any of the conventional DE or PDE problems of physics or chemistry or mathematics. Example:
Solve

$$\frac{d^2y}{dr^2} = q(r), \quad y(0) = y_0, \; y(R) = y_1 \tag{1}$$

where $q(r), y_0, y_1$ are known (and no mathematical pathologies in $q(r)$). Many ways to solve this (a boundary value problem.)
Or,

$$\frac{d^2y}{dr^2} = q(r), \quad y(0) = y_0, \; y'(0) = y_0' \tag{2}$$

$q(r), y_0, y_1'$ are known (an initial value problem). Or,

$$\nabla^2\phi = \rho(x,y), \; \phi(x_b,y_b) = \phi_0(x_b,y_b), x_b, y_b \in \partial D$$

(the Dirichlet problem), or

$$\nabla^2\phi = \rho(x,y), \; \phi'(x_b,y_b) = \phi_0(x_b,y_b), x_b, y_b \in \partial D$$

the Neumann problem.

These problems are classical in part because they are known to be "well-posed", meaning that they have unique, well behaved solutions in which e.g., small changes in a function q, or in the boundary values lead to small changes in the solution. (The Neumann problem is in fact ill-posed.) Commonly differentiability etc. is assured. Textbooks will tell you that you should never try to solve an "ill-posed" problem.

Trouble is, well-posed problems almost never exist in practice for anyone using observations. Consider one-more, the mass-spring oscillator (linear pendulum),

$$m\frac{d^2y}{dt^2} + r\frac{dy}{dt} + ky = q(t),\ y(0) = 1. \tag{3}$$

A necessary initial condition is missing and it's ill-posed. But Eq. (3) expresses a large amount of information. Should one abandon it? Or, suppose instead,

$$y(0) = 1,\ y(10) = 2,\ y(50) = -6$$

Now there is too much information, so it is once-again ill-posed. (Extra information could be redundant, or contradictory.)
Or suppose

$$y(0) = 1 \pm 0.3,\ y(10) = 2 \pm 1.$$

Solution cannot be unique, so it is again ill-posed. Or suppose,
$q(t) = q_0(t) \pm \Delta q(t)$

What is an inverse problem?

Defined relative to conventional forward problems. Consider again

$$\frac{d^2 y}{dr^2} = q(r)$$

but now suppose that y *is known* and the problem is to find $q(r)$. The solution to this inverse problem is trivial: just differentiate twice and one is done. It's an inverse problem that can be solved by purely conventional means.

So what is an **inverse method***?*

For our purposes, define it as any method that can be used to deal with ill-posed problems, including the ones above, that is able to (1) tell us how (un)certain the results are; and (2) which of our "data" really mattered; (3) whether some elements of the solution remain completely unknown (a special case of (1)). So an inverse *problem* might *not* formally involve an inverse *method* for solution.

(The most famous inverse problem is probably "Can one hear the shape of a drum?" posed by Mark Kač (1966):

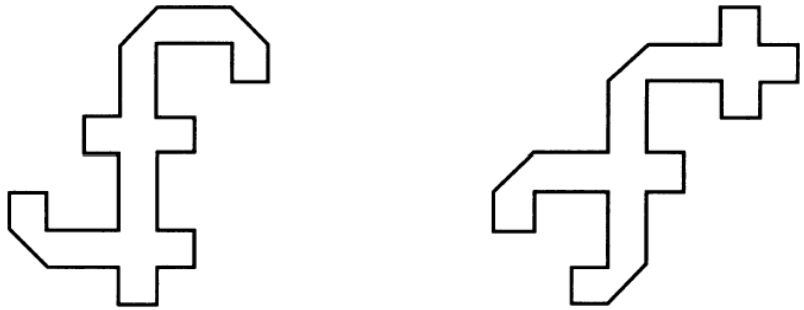


Figure: Gordon et al. (1992)

Forward problem: Given the boundary, and membrane properties, compute the eigen frequencies from

$$\nabla^2 \phi - \frac{1}{c^2} \frac{\partial^2 \phi}{\partial t^2} = 0, \qquad (**)$$

Inverse problem: Given the eigen-frequencies (the spectrum) of all solutions to (**) with $\phi = 0$ on the boundary, can one determine the shape of the boundary? Was answered in the negative a few years ago, Gordon et al. (1992). A real, and important, analogy is the determination of the interior structure of the Earth from the measured frequencies of its free oscillations.

Postulate:
All real problems can be rendered into discrete, finite, form with accuracy adequate for all practical purposes (might take a very large computer, but always a finite dimensional one). **Challenge: find me a counter-example.**
So, Eq. (1) can be written

$$\begin{aligned} y(n\Delta r + \Delta r) - 2y(n\Delta r) + y((n-1)\Delta r) &= \Delta r^2 q(n\Delta r) \\ y(0\Delta r) &= y_0, y(N\Delta r) = y_1 \end{aligned}$$

(Δr may have to be extremely small, but it is never zero.) r might be a spatial coordinate, in which case this represents a "static" problem. If r is time, it represents a time-evolution.

Writing it out,

$$
\begin{aligned}
y(0) &= y_0 \\
y(1\Delta r) &= y_1 \\
y(1\Delta r) - 2y(0\Delta r) + 0 &= \Delta r^2 q(0\Delta r) \\
y(2\Delta r) - 2y(1\Delta r) + y(0\Delta r) &= \Delta r^2 q(1\Delta r) \\
y(3\Delta r) - 2y(\Delta r) + y(0\Delta r) &= \Delta r^2 q(2\Delta r) \\
y(4\Delta r) - 2y(3\Delta r) + y(1\Delta r) &= \Delta r^2 q(3\Delta r) \\
&\cdot \\
&\cdot \\
y(N\Delta r) - 2y((N-1)\Delta r) + y((N-2)\Delta r) &= \Delta r^2 q((N-1)\Delta r)
\end{aligned}
$$

A set of N equations in N unknowns and which can in turn be rewritten,

$$\mathbf{Ex} = \mathbf{q}$$

where $\mathbf{x} = [y(0), y(1\Delta r), ..., y(N\Delta r)]^T$. Can solve it as

$$\mathbf{x} = \mathbf{E}^{-1}\mathbf{q}$$

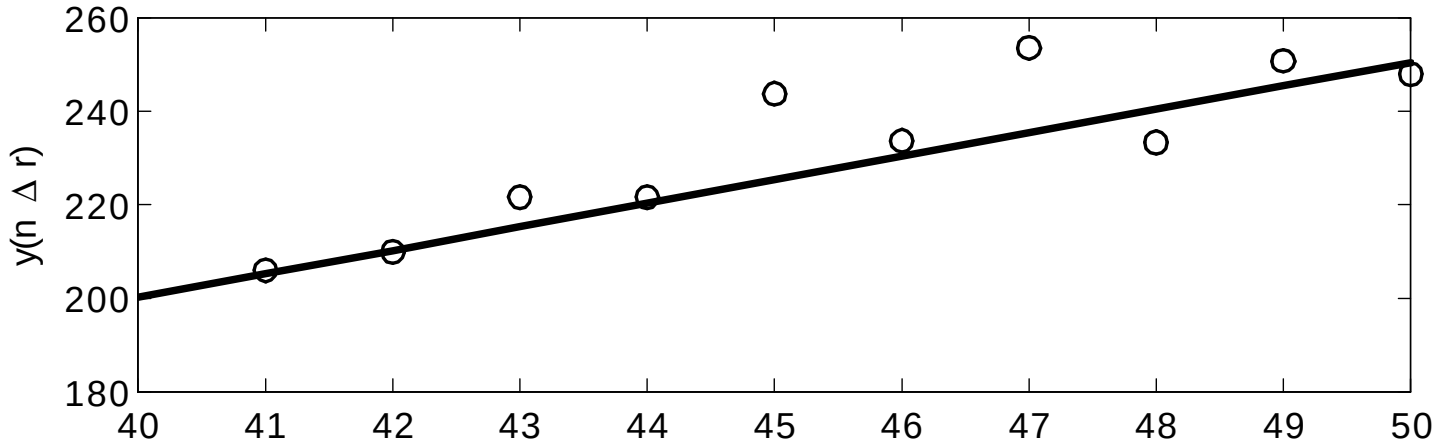
(the well-behaved nature of this system of equations is one approach to proving well-posedness). If one considers instead the initial value problem (2), the equations are identical, except the second one is replaced by,

$$y(1\Delta r) - y(0\Delta r) = y_1'$$

with the same solution but with an matrix $\mathbf{E}_1$. Unlikely, however, to solve it that way, as one can, with complete initial conditions as now, can *step* it forward in Δr. *The end result is identical to having inverted the matrix.* Keep that in mind! If $y(1\Delta r)$ is replaced by an end-condition, $y(N\Delta r)$, time-stepping is no longer possible, but matrix inversion is.

So now consider one of the ill-posed problems, with an "extra" value. That now means there is one more row in the matrix $\mathbf{E}$ than there are columns. It is conventionally labelled "overdetermined". It's clear that the system can be "inconsistent" or "consistent". Mathematically, a consistent solution exists only if all equations, including the extra one, are exactly satisfied. Would be nice to understand how to decide whether that is possible.
Alternatively, suppose the system is ill-posed because one of the boundary or initial conditions is missing. Then there is one fewer equation than unknowns and the problem is conventionally "underdetermined" and something many people find off-putting. Yet it might still have no solutions at all. We repeat the remark above: that there is *in practice no such thing as an over-determined problem—if any kind of observation is involved*.

This direction leads us to consider ordinary least-squares as one practices it in beginning science courses. Many of the conceptual issues can be understood from the most elementary problem of fitting a straight line to data:



We can write the problem as above:

$$\frac{d^2y}{dt^2} = 0, \; y\left(41\Delta r\right) = y_{41} \pm \varepsilon_{41}, \; y_{42} \pm \varepsilon_{42}, ...,$$

($r = 41, 42, ...$,is totally arbitrary). Could discretize the equation as above,

$$y\left(r + \Delta r\right) - 2y\left(r\right) + y\left(r - \Delta r\right) = 0 \tag{4}$$

and is a set of simultaneous equations, although with errors in some of them (the ones involving the observations, and maybe the model isn't perfect either).

Alternatively, we can reformulate it as

$$y = a + bt$$

which reduces the number of unknowns to 2 instead of all of $y(41), y(42)$,etc. (Am setting $\Delta r = 1$.)
Thus

$$\begin{aligned} a + 41b &= y(41) \\ a + 42b &= y(42) \\ &.. \\ a + (40 + N)b &= y(40 + N) \end{aligned}$$

or

$$\left\{ \begin{array}{cc} 1 & 41 \\ 1 & 42 \\ . & . \\ . & . \\ 1 & 40 + N \end{array} \right\} \left[\begin{array}{c} a \\ b \end{array} \right] = \left[\begin{array}{c} y(41) \\ y(42) \\ . \\ . \\ y(40 + N) \end{array} \right]$$

or

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

but which from the graph we know is contradictory. No straight line will produce a solution in the mathematical sense unless $N = 2$. Should really write it,

$$\mathbf{Ex} \approx \mathbf{y} \tag{5}$$

But equations are much easier to deal with than constructs like, (5), so convert it,

$$\mathbf{E}\mathbf{x}+\boldsymbol{\varepsilon}=\mathbf{y}$$

where $\boldsymbol{\varepsilon}$ represents the noise. It's still just a set of simultaneous equations except now we *could* write it as,

$$\left\{\begin{array}{cccccc} 1 & 41 & 1 & 0 & . & 0 \\ 1 & 42 & 0 & 1 & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & 40+N & 0 & 0 & 0 & 1 \end{array}\right\} \left[\begin{array}{c} a \\ b \\ \varepsilon(41)) \\ \varepsilon(42) \\ . \\ \varepsilon(40+N) \end{array}\right] = \left[\begin{array}{c} y(41) \\ y(42) \\ . \\ . \\ y(40+N) \end{array}\right]$$

which is once again

$$\mathbf{E}_1\mathbf{x}=\mathbf{y}$$

except now there are still N equations, but $N+2$ unknowns. Or use the finite form to the same end.
One is taught in school to solve this problem by minimizing $\sum \varepsilon_i^2 = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}$ (**why?**).

Using matrix-vector notation, write

$$J = \varepsilon^T \varepsilon = (\mathbf{y} - \mathbf{E}\mathbf{x})^T (\mathbf{y} - \mathbf{E}\mathbf{x})$$

and minimize it with respect to $\mathbf{x} = [a, b]^T$. Exercising matrix calculus:

$$\begin{aligned} \frac{\partial J}{\partial \mathbf{x}} &= \frac{\partial \left\{ \mathbf{y}^T\mathbf{y} - \mathbf{x}^T\mathbf{E}^T\mathbf{y} - \mathbf{y}^T\mathbf{E}\mathbf{x} + \mathbf{x}^T\mathbf{E}^T\mathbf{E}\mathbf{x} \right\}}{\partial \mathbf{x}} \\ &= -\mathbf{E}^T\mathbf{y} - \mathbf{E}^T\mathbf{y} + 2\mathbf{E}^T\mathbf{E}\mathbf{x} = \mathbf{0} \end{aligned}$$

or,

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

We can then substitute, and find,

$$\tilde{\varepsilon} = \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}} = \mathbf{y} - \mathbf{E}\left(\mathbf{E}^T\mathbf{E}\right)^{-1} \mathbf{E}^T\mathbf{y}$$

Not so obvious why this should vanish if there's no noise. Notice that we now know a, b plus N values of ε, that is, $N + 2$ values, although we only had N equations. So why is it called "overdetermined"?

The straight-line problem suggests looking at the nature of extremely small, but provocative problems. Suppose have one relation,

$$x_1 + 3x_2 - 2x_3 = 7$$

It's clearly an ill-posed problem, with an infinite number of solutions, but if one were interested in this particular linear combination, the inability to determine $x_1, .., x_3$ separately may not matter. Suppose it were actually a measurement, so that *always,* it is best written,

$$x_1 + 3x_2 - 2x_3 + \varepsilon_1 = 7$$

and now there's a fourth unknown. Maybe it is known that $\langle \varepsilon_1 \rangle = 0$ (the mean), and that $\left\langle \varepsilon_1^2 \right\rangle = .01$. What can one then say about the combination?
Suppose there are two measurements,

$$\begin{aligned} x_1 &= 3 \\ x_1 &= 1 \end{aligned}$$

Both can't be true, and so this is a nonsensical statement. It is always sensible to rewrite it as

$$\begin{aligned} x_1 + \varepsilon_1 &= 3 \\ x_1 + \varepsilon_2 &= 1 \end{aligned}$$

Suppose it were known that $\left\langle \varepsilon_1^2 \right\rangle = 1, \left\langle \varepsilon_2^2 \right\rangle = 10$. What is best statement you can make about x_1? Does it make sense to minimize $\varepsilon_1^2 + \varepsilon_2^2$?
Suppose

$$x_1 + 3x_2 - 2x_3 = 7$$

exactly. There again exists an infinite number of solutions to this relationship (can you find them?) Suppose you are faced with

$$\begin{aligned} x_1 + 3x_2 - 2x_3 &= 7 \\ x_1 + 3x_2 - 2x_3 &= 7\,(1+\delta) \end{aligned}$$

where δ is a very small number. Now there are *no solutions* in any mathematical sense. Both underdetermined and *inconsistent.*

Or,

$$\begin{aligned} x_1 + 3x_2 - 2x_3 &= 7 \\ x_1 + 3x_2 - 2x_3 &= 10 \end{aligned}$$

which has *no* solutions, but

$$\begin{aligned} x_1 + 3x_2 - 2x_3 &= 7 \\ x_1 + 3(1+\delta)x_2 - 2x_3 &= 10 \end{aligned}$$

again has an infinite number no matter how small is δ.
All this suggests that the mathematical notion of a "solution" is not particularly useful in this context.

Inverse Problems, Inverse Methods, State Estimation, Data Assimilation, and All That

EPS Harvard. Lecture 2. January 2013

Carl Wunsch
Harvard University

January 24, 2013

The straight-line problem suggests looking at the nature of extremely small, but provocative problems. Suppose have one relation,

$$x_1 + 3x_2 - 2x_3 = 7$$

It's clearly an ill-posed problem, with an infinite number of solutions, but if one were interested in this particular linear combination, the inability to determine $x_1, .., x_3$ separately may not matter. Suppose it were actually a measurement, so that *always,* it is best written,

$$x_1 + 3x_2 - 2x_3 + \varepsilon_1 = 7$$

and now there's a fourth unknown. Maybe it is known that $\langle \varepsilon_1 \rangle = 0$ (the mean), and that $\left\langle \varepsilon_1^2 \right\rangle = .01$. What can one then say about the combination?
Suppose there are two measurements,

$$\begin{aligned} x_1 &= 3 \\ x_1 &= 1 \end{aligned}$$

Both can't be true, and so this is a mathematically nonsensical statement. Sensibly, rewrite it as

$$\begin{aligned} x_1 + \varepsilon_1 &= 3 \\ x_1 + \varepsilon_2 &= 1 \end{aligned}$$

Suppose it were known that $\left\langle \varepsilon_1^2 \right\rangle = 1, \left\langle \varepsilon_2^2 \right\rangle = 10$. What is best statement you can make about x_1? Does it make sense to minimize $\varepsilon_1^2 + \varepsilon_2^2$?
Suppose

$$x_1 + 3x_2 - 2x_3 = 7$$

exactly. There again exists an infinite number of solutions to this relationship (can you find them?) Suppose you are faced with

$$\begin{aligned} x_1 + 3x_2 - 2x_3 &= 7 \\ x_1 + 3x_2 - 2x_3 &= 7\,(1+\delta) \end{aligned}$$

where δ is a very small number. Now there are *no solutions* in any strict mathematical sense. Both underdetermined and *inconsistent.*

Or,

$$\begin{aligned} x_1 + 3x_2 - 2x_3 &= 7 \\ x_1 + 3x_2 - 2x_3 &= 10 \end{aligned}$$

which has *no* solutions, but

$$\begin{aligned} x_1 + 3x_2 - 2x_3 &= 7 \\ x_1 + 3(1+\delta)x_2 - 2x_3 &= 10 \end{aligned}$$

again has an infinite number no matter how small is δ.
All this suggests that the mathematical notion of a "solution" is not particularly useful in this context.

Another issue: suppose you have reason to believe that the straight line is $\tilde{y} = \tilde{a} + \tilde{b}t$, where $\tilde{a} = a \pm \Delta a$, $\tilde{b} = b \pm \Delta b$ (where the Δa etc. are to be thought of as a standard error). But now you are given a single new measurement, $y(t_{new}) = y_{new} \pm \Delta y$. What is the most sensible thing to do with this measurement? And how do you do it? (You might have a problem with 10^8 parameters such as $a, b, c, \ldots$ and a single measurement. Are you stopped?)

The most general tool for linear problems is probably the singular value decomposition (SVD). Almost magical.
Start with two, simple, standard ideas.
(1) An $N-$dimensional vector $\mathbf{q}$ is completely known if its *projection.* $\mathbf{q}^T \mathbf{f}_j$ onto N independent vectors, $\mathbf{f}_j$ are known. Particularly simple if the $\mathbf{f}_i^T \mathbf{f}_j = \delta_{ij}$ are a complete *orthonormal* set or basis

$$\mathbf{q} = \sum_{j=1}^{N} \alpha_i \mathbf{f}_i = \left(\mathbf{f}_1^T \mathbf{q}\right) \mathbf{f}_1 + \left(\mathbf{f}_2^T \mathbf{q}\right) \mathbf{f}_2 + ... + \left(\mathbf{f}_N^T \mathbf{q}\right) \mathbf{f}_N$$

Suppose you know only K of the coefficients; then there will be an error:

$$\mathbf{q} = \sum_{j=1}^{K} \alpha_i' \mathbf{f}_i + \boldsymbol{\varepsilon}$$

Easy to show that to make the squared error magnitude, $\boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}$, as small as possible, one can do no better than choosing, again, $\alpha_i' = \alpha_i$. That is, the best representation would be,

$$\begin{aligned} \tilde{\mathbf{q}} &= \sum_{j=1}^{K} \alpha_i \mathbf{f}_i = \left(\mathbf{f}_1^T \mathbf{q}\right) \mathbf{f}_1 + \left(\mathbf{f}_2^T \mathbf{q}\right) \mathbf{f}_2 + ... + \left(\mathbf{f}_K^T \mathbf{q}\right) \mathbf{f}_K . \\ \mathbf{q} &= \tilde{\mathbf{q}} + \boldsymbol{\varepsilon} \end{aligned}$$

(return to this later)

Make an $N \times N$ matrix of the $\mathbf{f}_i$,

$$\mathbf{F} = \{\mathbf{f}_1 | \mathbf{f}_2 | ... | \mathbf{f}_N\}$$

then

$$\mathbf{q} = \mathbf{F}\left(\mathbf{q}^T \mathbf{F}\right)$$

If truncate to the first P columns,

$$\tilde{\mathbf{q}} = \mathbf{F}_P \left(\mathbf{q}^T \mathbf{F}_P\right)$$

Aside: If $\mathbf{g}_i$ are a set of N-vectors such that

$$\mathbf{q} = \sum_{j=1}^{N} \beta_j \mathbf{g}_i$$

exactly, for any $N-$dimensional $\mathbf{q}$, then the $\mathbf{g}_i$ are a basis, albeit the coefficients β_j are found trivially only if the $\mathbf{g}_j$ are also orthogonal or orthonormal. Have to solve the simultaneous equations

$$\mathbf{q}^T \mathbf{g}_i = \sum_{j=1}^{N} \beta_j \left(\mathbf{g}_j^T \mathbf{g}_i \right)$$

(2) Let $\mathbf{A}$ be a *square*, *symmetric* matrix, $\mathbf{A} = \mathbf{A}^T$ of dimension, N. Then there is a theorem that the "eigenvalue, eigenvector" problem,

$$\mathbf{A}\mathbf{f}_i = \lambda_i \mathbf{f}_i$$

always has a solution such that the λ_i are all real, and the $\mathbf{f}_i$ *are a complete orthonormal set.*
Make

$$\mathbf{F} = \{\mathbf{f}_1 | \mathbf{f}_2 | ... | \mathbf{f}_N\}$$

$N \times N$ from the column vectors, and from the λ_i,

$$\mathbf{\Lambda} = \mathbf{diag}\,(\lambda_i)\,,$$

so that

$$\mathbf{AF} = \mathbf{\Lambda F}, \quad \mathbf{F}^T\mathbf{AF} = \mathbf{\Lambda}$$

and hence,

$$\mathbf{A} = \mathbf{F\Lambda F}^T. \tag{1}$$

We used, $\mathbf{F}^T\mathbf{F} = \mathbf{I} = \mathbf{F}\mathbf{F}^T$ (an *orthogonal matrix)*. So $\mathbf{F}^{-1} = \mathbf{F}^T$ and a theorem proves that for a square matrix, a left inverse is also a right inverse.

$$\mathbf{A} = \mathbf{F}\mathbf{\Lambda}\mathbf{F}^T. \tag{2}$$

If some of the $\lambda_j = 0$, one has *exactly,*

$$\mathbf{A} = \mathbf{F}_K \mathbf{\Lambda}_K \mathbf{F}_K^T$$

where K is the number of non-zero λ_j, and,

$$\begin{aligned} \mathbf{F}_K &= \{\mathbf{f}_1|\mathbf{f}_2|...|\mathbf{f}_K\}, \quad N \times K \\ \mathbf{\Lambda}_K &= \begin{Bmatrix} \lambda_1 & 0 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 & 0 \\ 0 & 0 & . & 0 & 0 \\ 0 & 0 & 0 & . & 0 \\ 0 & 0 & 0 & 0 & \lambda_K \end{Bmatrix} \end{aligned}$$

(Mathematically, zero really means *zero.* In the observed world, a decision is required as to what zero actually means.)

K is called the *rank*, and the first K eigenvectors are the *range*. The last $(N-K)$ are the *nullspace*, as $\mathbf{A}\mathbf{f}_j = 0$. Define the nullspace matrix,

$$\mathbf{F}_{null} = \{\mathbf{f}_{K+1}|\mathbf{f}_{K+2}|...|\mathbf{f}_N\}, \quad N \times (N-K)$$

Because the $\mathbf{f}$ are an orthonormal basis, an arbitrary $N-$dimensional vector $\mathbf{q}$ can be *exactly* represented as

$$\mathbf{q} = \sum_{j=1}^{N} \left(\mathbf{f}_j^T \mathbf{q}\right) \mathbf{f}_j$$

The set of coefficients from the *projections* $\mathbf{f}_j^T \mathbf{q}$ carry the essential information. Note that should some of them be missing e.g., $\mathbf{f}_j^T \mathbf{q}$, $j = P+1, ..., N$ are unknown, it follows that

$$\tilde{\mathbf{q}} = \sum_{j=1}^{P} \left(\mathbf{f}_j^T \mathbf{q}\right) \mathbf{f}_j$$

is the best representation of $\mathbf{q}$ that can be made, in the sense that $\|\tilde{\mathbf{q}} - \mathbf{q}\|$ cannot be made any smaller by any other choice of coefficients. $\|.\|$ denotes the *norm*, here just the square root of the sum of squares (other norms exist).

Note $\mathbf{F}_K^T \mathbf{F}_K = \mathbf{I}$, and $\mathbf{F}_K \mathbf{F}_K^T \neq \mathbf{I}$. ($\mathbf{F}_K$ is not square.) Note that

$$\mathbf{v} = \mathbf{F}\mathbf{F}^T \mathbf{v} = \mathbf{I}\mathbf{v} = \mathbf{v}$$

$$\tilde{\mathbf{v}} = \mathbf{F}_K \left(\mathbf{F}_K^T \mathbf{v} \right) = \left(\mathbf{F}_K \mathbf{F}_K^T \right) \mathbf{v} \neq \mathbf{I}\mathbf{v}$$

$\mathbf{F}_K \mathbf{F}_K^T$ is the *resolution matrix.* Its significance is that a truncated expansion is a weighted linear combination of the elements of the true vector.

Consider solving the simultaneous equations,

$$\mathbf{E}\mathbf{x} = \mathbf{q},$$

using the completeness of the orthonomormal $\mathbf{f}_j$. Expand $\mathbf{x} = \sum_j \alpha_j \mathbf{f}_j$, $\mathbf{q} = \sum_{j=1}^{N} \left(\mathbf{f}_j^T \mathbf{q} \right) \mathbf{f}_j$ and substitute:

$$\sum_{j=1}^{N} \alpha_j \mathbf{E}\mathbf{f}_j = \sum_{j=1}^{N} \alpha_j \lambda_j \mathbf{f}_j = \sum_{j=1}^{N} \left(\mathbf{f}_j^T \mathbf{q} \right) \mathbf{f}_j$$

By orthogonality,

$$\alpha_j = \frac{\left(\mathbf{f}_j^T \mathbf{q} \right)}{\lambda_j}$$

and hence,

$$\mathbf{x} = \sum_{j=1}^{N} \frac{\left(\mathbf{f}_j^T \mathbf{q} \right)}{\lambda_j} \mathbf{f}_j$$

Unless, one or more of the λ_j vanishes.

Alternatively,

$$\mathbf{A}\mathbf{x} = \mathbf{F}\mathbf{\Lambda}\mathbf{F}^T\mathbf{x} = \mathbf{q},$$

hence,

$$\mathbf{\Lambda}\mathbf{F}^T\mathbf{x} = \mathbf{F}^T\mathbf{q},$$

or, $\mathbf{U}^T\mathbf{x} = \mathbf{\Lambda}^{-1}\mathbf{F}^T\mathbf{q}$, which are the expansion coefficients. Hence, $\mathbf{x} = \mathbf{F}\mathbf{\Lambda}^{-1}\mathbf{F}^T\mathbf{q}$, as long as no diagonal element of $\mathbf{\Lambda}$ vanishes.
Suppose there are only K non-zero λ_j. Then, $N - K$ of the α_j cannot be determined, and the best we could do is to write,

$$\mathbf{x} = \sum_{j=1}^{K} \frac{\left(\mathbf{f}_j^T\mathbf{q}\right)}{\lambda_j}\mathbf{f}_j + \sum_{j=K+1}^{N} \alpha_j \mathbf{f}_j.$$

The equations give us no information about α_j for $j = K+1, .., N$. These are the *nullspace*–and we do know their structures.

Then we can write e.g.,

$$\begin{aligned}
\mathbf{q} &= \left(\mathbf{F}\mathbf{F}^T\mathbf{q}\right) = \sum_j \mathbf{f}_j \left(\mathbf{f}_j^T \mathbf{q}\right), \\
\mathbf{x} &= \left(\mathbf{F}\boldsymbol{\Lambda}^{-1}\mathbf{F}^T\right)\mathbf{q} = \mathbf{F}\mathbf{F}^T\mathbf{x} = \sum_j \mathbf{f}_j \left(\mathbf{f}_j^T \mathbf{x}\right), \quad K = N \\
\tilde{\mathbf{x}} &= \mathbf{F}_K \left(\boldsymbol{\Lambda}_K^{-1}\mathbf{F}_K^T\mathbf{q}\right) + \mathbf{F}_{null}\boldsymbol{\alpha}_{null}, \quad K < N
\end{aligned}$$

If $\boldsymbol{\alpha}_{null}$ are set to zero (there is no reason to give them any finite value), one has an estimated

$$\tilde{\mathbf{x}} = \mathbf{F}_K \mathbf{F}_K^T \mathbf{x}.$$

which one might call "the" SVD solution—a linear combination of the elements of the correct one. When $K < N$, one can generate an infinite number of solutions from arbitrary choices of $\boldsymbol{\alpha}_{null}$.

Suppose, in the special case of a square symmetric problem, $\mathbf{Ax} = \mathbf{q}$, that it is "rank-deficient" $(K < N)$,

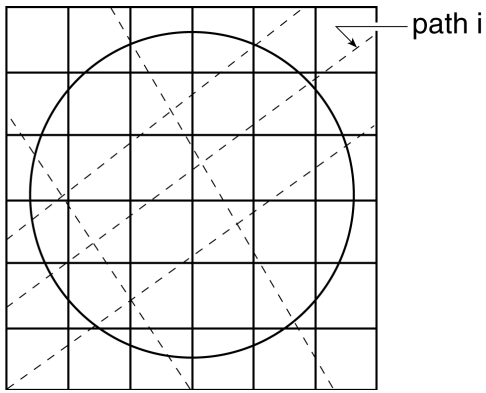
$$\tilde{\mathbf{x}} = \mathbf{F}_K \left(\mathbf{\Lambda}_K^{-1} \mathbf{F}_K^T \mathbf{q} \right) + \mathbf{F}_{null} \boldsymbol{\alpha}_{null}, \quad K < N$$

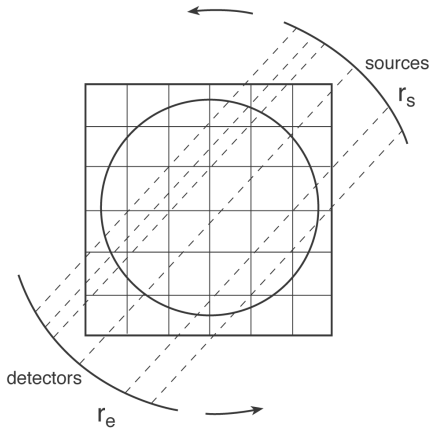
so that,

$$\begin{aligned} \mathbf{A}\tilde{\mathbf{x}} &= \left(\mathbf{F}_K \mathbf{\Lambda}_K \mathbf{F}_K^T \right) \mathbf{F}_K \left(\mathbf{\Lambda}_K^{-1} \mathbf{F}_K^T \mathbf{q} \right) \\ &= \mathbf{F}_K \mathbf{F}_K^T \mathbf{q} \neq \mathbf{q} \end{aligned}$$

The difference, $\mathbf{q} - \mathbf{F}_K \mathbf{F}_K^T \mathbf{q} = \boldsymbol{\varepsilon}$, is the "residual" and is easily seen to consist of the projection of $\mathbf{q}$ onto the null-space vectors.

A physical example of a simple inverse problem, often useful for thought experiments, is that of generic tomography:





If integrals are of travel time (could be absorption or any equivalent property), each integral is

$$\tau_p = \int_{source}^{receiver} \frac{ds}{c}.$$

Suppose seek the anomaly of soundspeed Δc_{ij} in box ij, $c_{ij} = c_0 + \Delta c_{ij}$

$$\begin{aligned}
\tau_p &= \int_{source}^{receiver} \frac{ds}{c_0 + \Delta c_{ij}} = \int_{source}^{receiver} \frac{ds}{c_0 \left(1 + \Delta c_{ij}/c_0\right)} \\
&= \int_{source}^{receiver} \frac{ds}{c_0} \left(1 - \Delta c_{ij}/c_0 + \left(\Delta c_{ij}/c_0\right)^2 + ...\right) \\
\tau_p &= \int_{source}^{receiver} \frac{ds}{c_0} - \int_{source}^{receiver} \frac{\Delta c_{ij} ds}{c_0^2} + ... \\
\Delta\tau_p &= \tau_p - \int_{source}^{receiver} \frac{ds}{c_0} \approx - \int_{source}^{receiver} \frac{\Delta c_{ij} ds}{c_0^2} \\
\Delta\tau_p &= \sum_{path\ p(ij)} \frac{\Delta c_{ij} \Delta s_{ij}}{c_0^2} + \varepsilon_p
\end{aligned}$$

Data represent integrals of some property (e.g., absorption or travel time) along each ray through the intersected boxes. Depending upon the number of rays and the number of boxes, there may be many fewer equations than (formal) unknowns, or many more. (A mathematician would note that in the continuous case, one can invoke the Radon transform. We don't need it.)

A tracer problem:

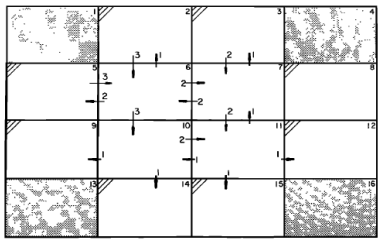


Fig. 1. Simple box model used to illustrate control form of transient tracer problem. The four corner boxes (stippled) are not involved in the calculation; boxes with partial shading are regions in which formal boundary conditions are required. The flow fields J_{ij} used in the computations are displayed with the numerical values attached. Concentrations in boundary boxes from which there is no flow into the interior are physically irrelevant ("unobservable"), e.g., box 9. Concentrations in boxes 8, 12, 14, and 15 are specified as zero. Active interior boxes are 6, 7, 10, and 11.

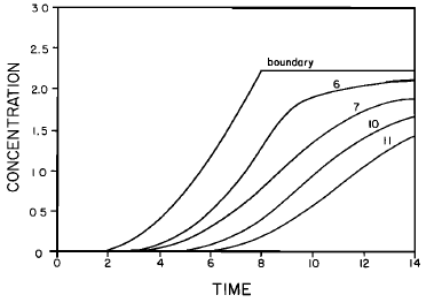


Fig. 3a. Concentration time histories for a 14 time step forward integration. Boundary boxes were driven by the concentration rates of change shown in Figure 2 resulting in the concentrations marked "boundary." Time histories for interior boxes are also depicted as calculated from (4).

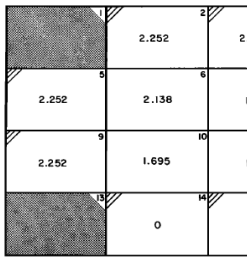


Fig. 3b. Concentrations at termi

$$\frac{C_q\left(t+\Delta t\right)-C_q\left(t\right)}{\Delta t}=\sum_{nbrs\ of\ q} J_{pq}C_p-\sum_{nbrs\ of\ q} J_{qp}C_q-\lambda C_q+\varepsilon_q$$

Forward problem: given boundary values, and initial conditions $C_q\left(0\right)$, compute $C_q\left(t\right)$. Many inverse problems: given $C_q\left(t\right)$ what are the J? What were the boundary or initial conditions? Etc.

Confirm the operation of the resolution matrix—example for an arbitrary vector $\mathbf{v}$: Let $\mathbf{v} = [3, -1]^T$. Let $\mathbf{f}_1 = 1/\sqrt{2}[1, -1]^T$, $\mathbf{f}_2 = 1/\sqrt{2}[1, 1]^T$ a complete, orthonormal pair. So

$$\mathbf{F} = \frac{1}{\sqrt{2}} \left\{ \begin{array}{cc} 1 & 1 \\ -1 & 1 \end{array} \right\}$$

(confirm $\mathbf{F}\mathbf{F}^T = \mathbf{F}^T\mathbf{F} = \mathbf{I}$). Then,

$$\mathbf{F}_1 = \frac{1}{\sqrt{2}} \left\{ \begin{array}{c} 1 \\ -1 \end{array} \right\}$$

$$\begin{aligned} \mathbf{v} &= \underbrace{\frac{1}{\sqrt{2}} \left[\begin{array}{c} 1 \\ -1 \end{array} \right]}_{\mathbf{f}_1} \left(4/\sqrt{2} \right) + \underbrace{\frac{1}{\sqrt{2}} \left[\begin{array}{c} 1 \\ 1 \end{array} \right]}_{\mathbf{f}_2} \left(2/\sqrt{2} \right), \\ \tilde{\mathbf{v}} &= \underbrace{\frac{1}{\sqrt{2}} \left[\begin{array}{c} 1 \\ -1 \end{array} \right]}_{\mathbf{f}_1} \left(4/\sqrt{2} \right) \stackrel{?}{=} \underbrace{\left\{ \frac{1}{\sqrt{2}} \left[\begin{array}{c} 1 \\ -1 \end{array} \right] \frac{1}{\sqrt{2}} \left[\begin{array}{cc} 1 & -1 \end{array} \right] \right\}}_{\mathbf{F}_1\mathbf{F}_1^T} \underbrace{\left[\begin{array}{c} 3 \\ -1 \end{array} \right]}_{\mathbf{v}} \end{aligned}$$

(It checks.)

Most problems, like the tomographic one, are neither symmetric nor square, $M \neq N$. Suppose now that one has an arbitrary $M \times N$ matrix, $\mathbf{E}$, perhaps being used for a set of equations,

$$\mathbf{Ex} + \boldsymbol{\varepsilon} = \mathbf{y}$$

Now the eigenvalue/eigenvector theorem for square, symmetric, matrices cannot be used. So let's construct a square symmetric matrix out of block sub-matrices:

$$\left\{ \begin{array}{cc} \mathbf{0} & \mathbf{E}^T \\ \mathbf{E} & \mathbf{0} \end{array} \right\} \equiv \mathbf{B}.$$

where the non-square zero matrices have the right dimensions to make this work ($\mathbf{B}$ is $(M+N) \times (M+N)$). Then $\mathbf{B} = \mathbf{B}^T$. Applying the eigenvalue/eigenvector theorem,

$$\mathbf{B}\mathbf{f}_i = \lambda_i \mathbf{f}_i, \quad i = 1, ..., M+N, \quad \mathbf{f}_i^T \mathbf{f}_j = \delta_{ij}$$

Write it out:

$$\left\{ \begin{array}{cc} \mathbf{0} & \mathbf{E}^T \\ \mathbf{E} & \mathbf{0} \end{array} \right\} \begin{bmatrix} f_{i1} \\ f_{i2} \\ . \\ f_{iN} \\ f_{i,N+1} \\ . \\ f_{i,M+N} \end{bmatrix} = \lambda \begin{bmatrix} f_{i1} \\ f_{i2} \\ . \\ f_{N} \\ f_{i,N+1} \\ . \\ f_{i,M+N} \end{bmatrix}$$

or

$$\mathbf{E}^T \begin{bmatrix} f_{i,N+1} \\ . \\ f_{i,M+N} \end{bmatrix} = \lambda_i \begin{bmatrix} f_{i1} \\ . \\ f_{iN} \end{bmatrix},$$

$$\mathbf{E} \begin{bmatrix} f_{i1} \\ . \\ f_{IN} \end{bmatrix} = \lambda_i \begin{bmatrix} f_{i,N+1} \\ . \\ f_{i,M+N} \end{bmatrix}$$

Or, as a short-hand,

$$\begin{aligned} \mathbf{E}^T \mathbf{u}_i &= \lambda_i \mathbf{v}_i, \\ \mathbf{E}\mathbf{v}_i &= \lambda_i \mathbf{u}_i \end{aligned} \tag{3a}$$

with the obvious substitutions. Left multiply the first of these by $\mathbf{E}$:

$$\mathbf{E}\mathbf{E}^T \mathbf{u}_i = \lambda_i \mathbf{E}\mathbf{v}_i = \lambda_i^2 \mathbf{u}_i$$

and for the second, left multiplying by $\mathbf{E}^T$ produces,

$$\mathbf{E}^T \mathbf{E}\mathbf{v}_i = \lambda_i^2 \mathbf{v}_i.$$

But both $\mathbf{E}^T\mathbf{E}$ and $\mathbf{E}\mathbf{E}^T$ are square symmetric matrices. Thus the $\mathbf{u}_i$, $\mathbf{v}_i$ are *separately* complete orthonormal bases (slightly amazing). Make matrices of the $\mathbf{u}_i$, $\mathbf{v}_i$ and λ_i

$$\mathbf{U} = \{\mathbf{u}_1|\mathbf{u}_2|...|\mathbf{u}_M\}, \quad \mathbf{V} = \{\mathbf{v}_1|\mathbf{v}_2|...|\mathbf{v}_N\}, \quad \mathbf{\Lambda} = \mathrm{diag}\,(\lambda_i)$$

where diag is generalized to an $M \times N$ matrix. The equations can be collected as,

$$\mathbf{E}\mathbf{V} = \mathbf{U}\mathbf{\Lambda}, \qquad \mathbf{E}^T\mathbf{U} = \mathbf{V}\mathbf{\Lambda}^T, \tag{4}$$

$$\mathbf{E}^T\mathbf{E}\mathbf{V} = \mathbf{V}\mathbf{\Lambda}^T\mathbf{\Lambda}, \qquad \mathbf{E}\mathbf{E}^T\mathbf{U} = \mathbf{U}\mathbf{\Lambda}\mathbf{\Lambda}^T. \tag{5}$$

Left multiply the first by $\mathbf{U}^T$ and right multiply it by $\mathbf{V}^T$, and invoking Eq. (**??**),

$$\mathbf{U}^T \mathbf{E} \mathbf{V} = \mathbf{\Lambda} . \tag{6}$$

So $\mathbf{U}$, $\mathbf{V}$ diagonalize $\mathbf{E}$ (with "diagonal" having an obvious extended meaning for a rectangular matrix)

$$\mathbf{E} = \mathbf{U} \mathbf{\Lambda} \mathbf{V}^T . \tag{7}$$

This last equation represents a product, called the "singular value decomposition" (SVD), of an arbitrary matrix, consisting of two orthogonal matrices, $\mathbf{U}$, $\mathbf{V}$, of different dimension, (M, N respectively) and a usually non-square diagonal matrix, $\mathbf{\Lambda}$, dimension $M \times N$.

There is one further step to take. Notice that for a rectangular $\mathbf{\Lambda}$, as in the examples above, one or more rows or columns must be all zero, depending upon the shape of the matrix. In addition, if any of the $\lambda_i = 0$, $i < \min(M, N)$, the corresponding rows or columns of $\mathbf{\Lambda}$ 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 $\mathbf{U}_K$, and is $M \times K$ and $\mathbf{V}$ becomes $\mathbf{V}_k$, $N \times K$, and reducing $\mathbf{\Lambda}$ to a $K \times K$ square matrix, then the representation (**??**) remains *exact*, in the form,

$$\mathbf{E} = \mathbf{U}_K \mathbf{\Lambda}_K \mathbf{V}_K^T = \lambda_1 \mathbf{u}_1 \mathbf{v}_1^T + \lambda_2 \mathbf{u}_2 \mathbf{v}_2^T + ... + \lambda_K \mathbf{u}_K \mathbf{v}_K^T \ , \tag{8}$$

What are usually called *empirical orthogonal functions (EOFs)*, or sometimes *principal components,* or Karhunen-Loève vectors (and other labels) are just the $\mathbf{u}_i$ or the $\mathbf{v}_i$ (depending upon which dimension is space and which is time, if any). Sometimes rows and columns are weighted in various ways before computing.

Because of the mixed dimensions in Eqs. (**??**), the only way there is consistency is if the number of non-zero λ_i is no more than the minimum of (M, N) (called K – –the rank). If one accounts for the zero λ_i, one has, exactly,

$$\mathbf{E} = \mathbf{U}_K \mathbf{\Lambda}_K \mathbf{V}_K^T \tag{9}$$

where the columns corresponding to indices greater than K are omitted from $\mathbf{U}_K, \mathbf{V}_K$ and $\mathbf{\Lambda}_K$ is now *square* of dimension K with all non-zeros on the diagonal. (If some of the λ_i are very small, one might omit them from Eq. (**??**) with little error. This becomes the basis for the use of reduced numbers of EOFs and is a result of the *Eckart-Young-Mirsky Theorem.)*

Suppose now we have a set of equations,

$$\mathbf{E}\mathbf{x}+\boldsymbol{\varepsilon}=\mathbf{y}$$

There are two spaces of different dimensions, M, N present. $\boldsymbol{\varepsilon}$, $\mathbf{y}$ are in an $M-$dimensional space and $\mathbf{x}$ is in an $N-$dimensional one. Suppose

$$\begin{aligned}\mathbf{x} &= \sum_{j=1}^{N} \alpha_i \mathbf{v}_i, \quad \mathbf{y} = \sum_{i=1}^{M} \mathbf{u}_i \left(\mathbf{u}_i^T \mathbf{y}\right) \\ \boldsymbol{\varepsilon} &= \sum_{j=1}^{M} \beta_i \mathbf{u}_i\end{aligned}$$

(the α_i, β_i are at the moment unknown) Thus,

$$\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^T\mathbf{x} + \mathbf{U}\left(\mathbf{U}^T\boldsymbol{\varepsilon}\right) = \mathbf{y}$$

Left multiply by $\mathbf{U}^T$:

$$\boldsymbol{\Lambda}\mathbf{V}^T\mathbf{x} + \mathbf{U}^T\,\boldsymbol{\varepsilon} = \mathbf{U}^T\mathbf{y},$$

or,

$$\begin{aligned} \lambda_i \mathbf{u}_i^T \mathbf{x} + \mathbf{u}_i^T \boldsymbol{\varepsilon} &= \mathbf{u}_i^T \mathbf{y}, \quad i = 1, .., K \\ 0 \mathbf{u}_i^T \mathbf{x} + \mathbf{u}_i^T \boldsymbol{\varepsilon} &= \mathbf{u}_i^T \mathbf{y}, \quad i = K+1, .., M \end{aligned}$$

If we want $\boldsymbol{\varepsilon}$ to be as small as possible, we can set $\beta_i = \mathbf{u}_i^T \boldsymbol{\varepsilon} = 0$, $i = 1, .., K$ and then $\alpha_i = \mathbf{u}_i^T \mathbf{x} = \mathbf{u}_i^T \mathbf{y} / \lambda_i$. For the remaining terms, , there is no choice, $\beta_i = \mathbf{u}_i^T \boldsymbol{\varepsilon} = \mathbf{u}_i^T \mathbf{y}$. Using the matrix notation,

$$\begin{aligned} \mathbf{x} &= \mathbf{V} \left(\mathbf{V}^T \mathbf{x} \right) = \mathbf{V}_K \left(\mathbf{V}_K^T \mathbf{x} \right) + \mathbf{V}_{null} \left(\mathbf{V}_{null}^T \mathbf{x} \right) \\ &= \mathbf{V}_K \left(\boldsymbol{\Lambda}_K^{-1} \mathbf{U}_K^T \mathbf{y} \right) + \mathbf{V}_{null} \left(\mathbf{V}_{null}^T \mathbf{x} \right) \end{aligned}$$

if, $\mathbf{U}_K^T \boldsymbol{\varepsilon} = \mathbf{0}$. But $\mathbf{U}_K^T \boldsymbol{\varepsilon}$ are the first K expansion coefficients of $\boldsymbol{\varepsilon}$ in the complete basis set $\mathbf{u}_i$. Clearly we can make $\boldsymbol{\varepsilon}$ as small as possible by assuming that these coefficients vanish.

But suppose $M > N$ which means that there are more $\mathbf{u}_i$ vectors than $\mathbf{v}_i$ vectors and there are at most $K = N$ of them. There's no way that the solution could fully reproduce $\mathbf{y}$ which is in the M dimensional space. Thus we are left in that case with,

$$\mathbf{U}_{null}^T \boldsymbol{\varepsilon} = \mathbf{U}_{null}^T \mathbf{y}$$

and

$$\boldsymbol{\varepsilon} = \sum_{K+1}^{N} \mathbf{u}_i \left(\mathbf{u}_i^T \mathbf{y} \right) .$$

This is the residual of ordinary least-squares, making $\boldsymbol{\varepsilon}$ as small as possible. On the other hand, if $M < N$, then there will be more $\mathbf{v}_i$ than $\mathbf{u}_i$, and if one wanted (does it make sense?), the residuals could be set to zero. More generally, $K < \min(M, N)$ and so there are null spaces in both $\mathbf{u}_i, \mathbf{v}_i$. One can now solve problems of *any* dimension and write explicit statements about *all* possible solutions, including the structure of the residuals.

Recapitulation:
Forward problems can always be written as a set of simultaneous equations (they may be nonlinear). Problems involving data of any sort always have errors in them. Any equation involving observations always has an unknown error term in it. Generic, linear, (or linearized) form,

$$\mathbf{E}\mathbf{x}+\boldsymbol{\varepsilon}=\mathbf{y}, \quad M \times N$$

Underdetermined problems (not always obvious) will have $M \neq N$, either bigger or smaller. Even if, $M = N$, $\mathbf{E}^{-1}$ usually won't exist. All of the fanciest methods are algorithms, some of them very sophisticated (e.g. a Kalman filter—which doesn't actually solve it), for obtaining solutions to problems like this, usually without having to write them all down at once or having to invert big matrices. (But A. Ganachaud's PhD thesis about 12 years ago involved about 5000 equations in 10,000 unknowns and was solved all-at-once.) Often the matrices are very sparse. This short course is essentially how to exploit systems like this of any dimension, to cope with contradictions, and to figure out what is determined and what isn't in any procedure.

Simultaneous equations are so ubiquitous, it's worth a simple-minded look at what they say:

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \quad M \times N$$

A is made up of a series of known row vectors:

$$\left\{ \begin{array}{c} \mathbf{a}_1^T \\ \mathbf{a}_2^T \\ \\ \\ \mathbf{a}_M^T \end{array} \right\}$$

so the simultaneous equations provide information about M dot products with M known vectors. If $M \geq N$, and the $\mathbf{a}_j$ are independent, more than enough information to find $\mathbf{x}$. (Unless some of the information is contradictory). If $M < N$, not enough information to fully determine $\mathbf{x}$ under all circumstances.

Alternatively, $\mathbf{A}$ is a set of known column vectors,

$$\mathbf{A} = \left\{ \begin{array}{cccccc} \mathbf{c}_1 & \mathbf{c}_2 & . & . & \mathbf{c}_N \end{array} \right\}$$

and the equations are

$$x_1 \mathbf{c}_1 + x_2 \mathbf{c}_2 + ... + x_N \mathbf{c}_N = \mathbf{b}$$

So if over determined, too few vectors $\mathbf{c}_j$ are given to fully describe $\mathbf{b}$. If $M < N$, too many vectors are available to describe it. Etc.

Inverse Problems, Inverse Methods, State Estimation, Data Assimilation, and All That

January 2013 Lecture 3

Carl Wunsch
Harvard University

January 24, 2013

Consider the conventional problem

$$\mathbf{E}\mathbf{x} = \mathbf{b}$$

for a square, $N \times N$ not necessarily symmetric, $\mathbf{E}$. One writes

$$\mathbf{x} = \mathbf{E}^{-1}\mathbf{b}$$

Does this always work?
Consider $\mathbf{E} = \mathbf{U}\Lambda\mathbf{V}^T$. Then

$$\mathbf{x} = \left(\mathbf{U}\Lambda\mathbf{V}^T\right)^{-1}\mathbf{b}.$$

By inspection, $\left(\mathbf{U}\Lambda\mathbf{V}^T\right)^{-1} = \mathbf{V}\Lambda^{-1}\mathbf{U}^T$ (check it). Then

$$\mathbf{x} = \mathbf{V}\Lambda^{-1}\mathbf{U}^T\mathbf{b} = \sum_{j=1}^{N}\left(\frac{\mathbf{u}_j^T\mathbf{b}}{\lambda_j}\right)\mathbf{v}_j$$

if no non-zero λ_j. Confirm $\mathbf{b} = \mathbf{E}\mathbf{x}$, with $\mathbf{b} = \sum_{j=1}^{N}\left(\mathbf{u}_j^T\mathbf{b}\right)\mathbf{u}_j$.

If one or more does vanish, the inverse does not exist, and at best,

$$\tilde{\mathbf{x}} = \sum_{j=1}^{K} \left(\frac{\mathbf{u}_j^T \mathbf{b}}{\lambda_j} \right) \mathbf{v}_j + \sum_{j=K+1}^{N} \alpha_j \mathbf{v}_j$$

$$\mathbf{E}\tilde{\mathbf{x}} = \tilde{\mathbf{b}} = \sum_{j=1}^{L} \left(\mathbf{u}_j^T \mathbf{b} \right) \mathbf{u}_j \neq \mathbf{b}$$

and there is a residual (noise), $\boldsymbol{\varepsilon} = \sum_{j=K+1}^{N} \left(\mathbf{u}_j^T \mathbf{b} \right) u_j$. Noise has zero projection onto the range vectors. Why? Usually put the $\alpha_j = 0$ as no information about them is available. (Ocham's Razor: don't introduce any structure not required by the data.)

Some statistical notation. $\langle \cdot \rangle$ is used to denote expected value in the theoretical sense,

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

Here, $p_x(X)$ is the probability density for variable x. The *variance* is,

$$\left\langle (x - \langle x \rangle)^2 \right\rangle = \int_{-\infty}^{\infty} (X - m)\, p_x(X - m)\, dX = \sigma^2.$$

The *covariance* is,

$$\left\langle (x - \langle x \rangle)(y - \langle y \rangle)^T \right\rangle = \int\int_{-\infty}^{\infty} (X - m_x)(Y - m_y)\, p_{xy}(X - m_x, Y - m_y)\, dX dY$$

where $p_{xy}(X, Y)$ is the joint probability density for x, y, or in general vector form,

$$\left\langle (\mathbf{x} - \langle \mathbf{x} \rangle)(\mathbf{x} - \langle \mathbf{x} \rangle)^T \right\rangle = \int \cdots \int (\mathbf{X} - \mathbf{m})(\mathbf{X} - \mathbf{m})^T\, p_{\mathbf{x}}(\mathbf{X} - \mathbf{m})\, d\mathbf{X}$$

For "unimodal" distributions (the Gaussian or normal is one), finding an estimate of something, $\tilde{x}$, that minimizes either the variance about its mean value ($\langle \tilde{x} \rangle$), or sometimes about the true value, x, is sensible (but arbitrary).

Suppose $\mathbf{Ex} = \mathbf{b}$, but there are known to be errors in $\mathbf{y}$. Can account for that by writing explicitly

$$\mathbf{Ex} + \boldsymbol{\varepsilon} = \mathbf{b}.$$

(Could absorb this, by writing again,

$$\mathbf{E}_1 \boldsymbol{\xi} = \mathbf{b}$$

but leave as is for now. Implies $\mathbf{b} = \mathbf{b}_0 - \boldsymbol{\varepsilon}$. Thus

$$\begin{aligned} \tilde{\mathbf{x}} &= \sum_{j=1}^{K} \left(\frac{\mathbf{u}_j^T \mathbf{b}}{\lambda_j} \right) \mathbf{v}_j + 0 \\ &= \sum_{j=1}^{K} \left(\frac{\mathbf{u}_j^T \mathbf{b}_0 - \mathbf{u}_j^T \boldsymbol{\varepsilon}}{\lambda_j} \right) \mathbf{v}_j - \sum_{j=K+1}^{N} \alpha_j \mathbf{v}_j \end{aligned}$$

The expected difference is,

$$\langle \tilde{\mathbf{x}} - \mathbf{x} \rangle = \sum_{j=1}^{K} \left(\frac{\left\langle -\mathbf{u}_j^T \boldsymbol{\varepsilon} \right\rangle}{\lambda_j} \right) \mathbf{v}_j - \sum_{j=K+1}^{N} \alpha_j \mathbf{v}_j$$

Perhaps, $\langle \varepsilon \rangle = 0, \left\langle -\mathbf{u}_j^T \boldsymbol{\varepsilon} \right\rangle = -\mathbf{u}_j^T \langle \boldsymbol{\varepsilon} \rangle = 0$. There is only a bias error due to the missing null space.

What about $\left\langle (\tilde{\mathbf{x}} - \mathbf{x}) (\tilde{\mathbf{x}} - \mathbf{x})^T \right\rangle$?

$$\left\langle (\tilde{\mathbf{x}} - \mathbf{x}) (\tilde{\mathbf{x}} - \mathbf{x})^T \right\rangle = \sum_{j=1}^{K} \left(\frac{\left\langle \left(-\mathbf{u}_j^T \boldsymbol{\varepsilon} \right)^2 \right\rangle}{\lambda_j^2} \right) \mathbf{v}_j \mathbf{v}_j^T + \sum_{j=K+1}^{N} \alpha_j^2 \mathbf{v}_j \mathbf{v}_j^T$$

assuming no cross-covariances amongst the $\alpha_{i,j}$.
For resolution, one wants K to be as large as possible. But for small variance, want to omit small λ_j. A classical statistical tradeoff.

Back to least-squares for the moment. Consider a general problem: $\mathbf{E}\mathbf{x}+\boldsymbol{\varepsilon}=\mathbf{y}$ where we arbitrarily decide to minimize,

$$J = \boldsymbol{\varepsilon}^T \mathbf{W}^{-1} \boldsymbol{\varepsilon} + \mathbf{x}^T \mathbf{S}^{-1} \mathbf{x} \tag{1}$$

$$= (\mathbf{y} - \mathbf{E}\mathbf{x})^T \mathbf{W}^{-1} (\mathbf{y} - \mathbf{E}\mathbf{x}) + \mathbf{x}^T \mathbf{S}^{-1} \mathbf{x}\,, \tag{2}$$

Setting the derivatives with respect to $\mathbf{x}$ to zero results in,

---

$$\tilde{\mathbf{x}} = \left(\mathbf{E}^T \mathbf{W}^{-1} \mathbf{E} + \mathbf{S}^{-1} \right)^{-1} \mathbf{E}^T \mathbf{W}^{-1} \mathbf{y}, \tag{3}$$

$$\tilde{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}}, \tag{4}$$

$$\mathbf{C}_{xx} = \tag{5}$$

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

$\mathbf{C}_{xx}$ is the covariance about the expected value of the estimate. The *matrix inversion lemma* permits rewriting Eqs. $(19-21)$:

$$\tilde{\mathbf{x}} = \mathbf{S}\mathbf{E}^T \left(\mathbf{E}\mathbf{S}\mathbf{E}^T + \mathbf{W} \right)^{-1} \mathbf{y}, \tag{6}$$

$$\tilde{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}}, \tag{7}$$

$$\mathbf{C}_{xx} = \mathbf{S}\mathbf{E}^T \left(\mathbf{E}\mathbf{S}\mathbf{E}^T + \mathbf{W} \right)^{-1} \mathbf{R}_{nn} \left(\mathbf{E}\mathbf{S}\mathbf{E}^T + \mathbf{W} \right)^{-1} \mathbf{E}\mathbf{S}. \tag{8}$$

The different forms make it easy to let $\mathbf{W}^{-1} \rightarrow 0$, etc.

Consider the two limiting forms, $J = \min : \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}$. Then,

$$\tilde{\mathbf{x}} = \left(\mathbf{E}^T \mathbf{E} \right)^{-1} \mathbf{E}^T \mathbf{y}.$$

if the inverse exists (?). Suppose we have the SVD of $\mathbf{E} = \mathbf{U}_K \boldsymbol{\Lambda}_K \mathbf{V}_K^T$. Then

$$\begin{aligned} \tilde{\mathbf{x}} &= \left(\left(\mathbf{U}_K \boldsymbol{\Lambda}_K \mathbf{V}_K^T \right)^T \mathbf{U}_K \boldsymbol{\Lambda}_K \mathbf{V}_K^T \right)^{-1} \left(\mathbf{U}_K \boldsymbol{\Lambda}_K \mathbf{V}_K^T \right)^T \mathbf{y} \\ &= \left(\mathbf{V}_K \boldsymbol{\Lambda}_K \mathbf{U}_K^T \mathbf{U}_K \boldsymbol{\Lambda}_K \mathbf{V}_K^T \right)^{-1} \left(\mathbf{U}_K \boldsymbol{\Lambda}_K \mathbf{V}_K^T \right)^T \mathbf{y} \\ &= (\mathbf{V}_K \boldsymbol{\Lambda}_K^2 \mathbf{V}_K^T)^{-1} \mathbf{V}_K \boldsymbol{\Lambda}_K \mathbf{U}_K^T \mathbf{y} \end{aligned}$$

When does the inverse exist? By inspection,

$$(\mathbf{V}_K \boldsymbol{\Lambda}_K^2 \mathbf{V}_K^T) \left(\mathbf{V}_K \boldsymbol{\Lambda}_K^{-2} \mathbf{V}_K^T \right) = \mathbf{V}_K \mathbf{V}_K^T$$

which is the identity if, and only if, $K = N$ (the dimension of the $\mathbf{v}_i$). (If one subsituted the full matrices for $\mathbf{V}_K$ etc., the inverse "blows up" and the least-squares solution does not exist. Now we know when that is.) Suppose $K < N$ (meaning that there are fewer effective equations than unknowns).

Let us use $\mathbf{U}_K \mathbf{\Lambda}_K^2 \mathbf{V}_K^T$ as the best we can do:

$$\tilde{\mathbf{x}} = \left(\mathbf{V}_K \mathbf{\Lambda}_K^{-2} \mathbf{V}_K^T \right) \mathbf{V}_K \mathbf{\Lambda}_K \mathbf{U}_K^T \mathbf{y} = \mathbf{U}_K \mathbf{\Lambda}_K^{-1} \mathbf{U}_K^T \mathbf{y} = \sum_{i=1}^{K} \mathbf{v}_i \lambda_i^{-1} \left(\mathbf{u}_i^T \mathbf{y} \right)$$

and there is a "missing" null space, $\mathbf{x}_{null} = \sum_{j=K+1}^{N} \alpha_j \mathbf{v}_j$.

Now look at the other limit, $J = \mathbf{x}^T\mathbf{x}$, commonly used for underdetermined problems,

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

$$\tilde{\mathbf{x}} = \mathbf{V}_K \mathbf{\Lambda}_K \mathbf{U}_K^T (\mathbf{U}_K \mathbf{\Lambda}_K \mathbf{V}_K^T \mathbf{V}_K \mathbf{\Lambda}_K^T \mathbf{U}_K^T)^{-1} \mathbf{y}$$

$$= \mathbf{V}_K \mathbf{\Lambda}_K \mathbf{U}_K^T \left(\mathbf{U}_K \mathbf{\Lambda}_K \mathbf{U}_K^T \right)^{-1} \mathbf{y}$$

Again, by inspection, $\left(\mathbf{U}_K \mathbf{\Lambda}_K \mathbf{U}_K^T\right)^{-1} = \mathbf{U}_K \mathbf{\Lambda}_K^{-1} \mathbf{U}_K^T$, if and only if, $K = M$. (That is, the effective number of equations is equal to the total number.) Then,

$$\tilde{\mathbf{x}} = \mathbf{V}_M \mathbf{\Lambda}_M^{-1} \mathbf{U}_M^T \mathbf{y} = \sum_{i=1}^{M} \mathbf{v}_i \lambda_i^{-1} \left(\mathbf{u}_i^T \mathbf{y}\right) + \sum_{j=M+1}^{N} \alpha_j \mathbf{v}_j$$

as before. In this case, (*full rank underdetermined)*

$$\begin{aligned} \tilde{\varepsilon} &= \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}} = \mathbf{y} - \mathbf{E}\mathbf{E}^T \left(\mathbf{E}\mathbf{E}^T\right)^{-1} \mathbf{y} \\ &= \mathbf{0} \end{aligned}$$

If $K < N$, the best we can do is,

$$\tilde{\mathbf{x}} = \mathbf{V}_K \mathbf{\Lambda}_K^{-1} \mathbf{U}_K^T \mathbf{y} = \sum_{i=1}^{K} \mathbf{v}_i \lambda_i^{-1} \left(\mathbf{u}^T \mathbf{y} \right) + \sum_{j=K+1}^{N} \alpha_j \mathbf{v}_j$$

But since, $\left(\mathbf{E}\mathbf{E}^T \right)^{-1}$ does not exist,

$$\begin{aligned} \tilde{\varepsilon} &= \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}} = \mathbf{y} - \left(\mathbf{U}_K \mathbf{\Lambda}_K \mathbf{V}_K^T \right) \left(\mathbf{V}_K \mathbf{\Lambda}_K^{-1} \mathbf{U}_K^T \mathbf{y} \right) \\ &= \mathbf{y} - \mathbf{U}_K \mathbf{U}_K^T \mathbf{y} \end{aligned}$$

and there is a residual despite there being fewer equations than unknowns. (Many more things can be done.) No solution can reduce this residual.

Consider minimizing $J = \gamma^2 \mathbf{x}^T \mathbf{x} + \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}$. Then, with a little algebra or substituting, $\mathbf{S}^{-1} = \gamma^2 \mathbf{I}$, $\mathbf{W}^{-1} = \mathbf{I}$,

$$\begin{aligned} \tilde{\mathbf{x}} &= \mathbf{V}(\boldsymbol{\Lambda}^T \boldsymbol{\Lambda} + \gamma^2 \mathbf{I})^{-1} \mathbf{V}^T \mathbf{V} \boldsymbol{\Lambda}^T \mathbf{U}^T \mathbf{y} \\ &= \mathbf{V} \operatorname{diag} \left(\lambda_i^2 + \gamma^2 \right)^{-1} \boldsymbol{\Lambda}^T \mathbf{U}^T \mathbf{y}, \end{aligned} \tag{9}$$

or,

$$\tilde{\mathbf{x}} = \sum_{i=1}^{N} \frac{\lambda_i (\mathbf{u}_i^T \mathbf{y})}{\lambda_i^2 + \gamma^2} \mathbf{v}_i \, . \tag{10}$$

called "tapered least-squares". Note,

$$\begin{aligned} \mathbf{C}_{xx} &= \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\lambda_i \lambda_j \mathbf{u}_i^T \mathbf{R}_{nn} \mathbf{u}_j^T}{(\lambda_i^2 + \gamma^2)(\lambda_j^2 + \gamma^2)} \mathbf{v}_i \mathbf{v}_j^T \\ &= \sigma_n^2 \sum_{i=1}^{N} \frac{\lambda_i^2}{(\lambda_i^2 + \gamma^2)^2} \mathbf{v}_i \mathbf{v}_i^T \\ &= \sigma_n^2 \mathbf{V} (\boldsymbol{\Lambda}^T \boldsymbol{\Lambda} + \gamma^2 \mathbf{I}_N)^{-1} \boldsymbol{\Lambda}^T \boldsymbol{\Lambda} (\boldsymbol{\Lambda}^T \boldsymbol{\Lambda} + \gamma^2 \mathbf{I}_N)^{-1} \mathbf{V}^T, \end{aligned} \tag{11}$$

and consider the limit as $\gamma^2 \to 0$ when one or more λ_i is small. Even if some λ_i vanishes, there is still a finite solution. For λ_i small compared to γ, the contribution of the term to the solution is reduced from what it ought to be. The reader might want to confirm that this solution is also obtained by adding γ^2 to the diagonal of the matrix $\mathbf{E}^T\mathbf{E}$ which has to be inverted in conventional least-squares. This action assures the inverse exists; it represents a form of "regularization" which can be recognized as just reducing the influence of small λ_i.

Summary:
Given any set of linear simultaneous equations, $\mathbf{Ex} + \boldsymbol{\varepsilon} = \mathbf{y}$, of arbitrary dimension, one can characterize solutions $\tilde{\mathbf{x}}$ in terms of the their resolution, $\tilde{\mathbf{x}} = \mathbf{V}_K \mathbf{V}_K^T \mathbf{x}$, and the residual, $\tilde{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}}$, in terms of the $\mathbf{u}_i$ vectors. It remains to understand the statistical uncertainty of these solutions. The central point is that you can handle almost anything — if you know what you are doing....Given any set of linear constraints, of dimension $M \times N$, it's possible to find the components that are determinable, which ones are not, how stable are the ones that are known, which structures of the "data" are explainable, which structures cannot be explained by any solution, and which elements of the solution are determined by which data.

Thus for the least-squares problem,

$$\langle \tilde{\mathbf{x}} \rangle = \left\langle \mathbf{S}\mathbf{E}^T \left(\mathbf{E}\mathbf{S}\mathbf{E}^T + \mathbf{W} \right)^{-1} (\mathbf{y}_0 + \boldsymbol{\varepsilon}) \right\rangle = \mathbf{S}\mathbf{E}^T \left(\mathbf{E}\mathbf{S}\mathbf{E}^T + \mathbf{W} \right)^{-1} \mathbf{y}_0$$

assuming $\langle \boldsymbol{\varepsilon} \rangle = 0$. (*Note that in general,* $\langle \tilde{\mathbf{x}} \rangle \neq \mathbf{x}$.) Then,

$$\begin{aligned}
\mathbf{C}_{xx} &= \left\langle (\tilde{\mathbf{x}} - \langle \tilde{\mathbf{x}} \rangle) (\tilde{\mathbf{x}} - \langle \tilde{\mathbf{x}} \rangle)^T \right\rangle \\
&= \left\langle \left(\mathbf{S}\mathbf{E}^T \left(\mathbf{E}\mathbf{S}\mathbf{E}^T + \mathbf{W} \right)^{-1} \boldsymbol{\varepsilon} \right) \left(\mathbf{S}\mathbf{E}^T \left(\mathbf{E}\mathbf{S}\mathbf{E}^T + \mathbf{W} \right)^{-1} \boldsymbol{\varepsilon} \right)^T \right\rangle \\
&= \mathbf{S}\mathbf{E}^T \left(\mathbf{E}\mathbf{S}\mathbf{E}^T + \mathbf{W} \right)^{-1} \left\langle \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T \right\rangle \left(\mathbf{E}\mathbf{S}\mathbf{E}^T + \mathbf{W} \right)^{-1} \mathbf{E}\mathbf{S} \\
&= \mathbf{S}\mathbf{E}^T \left(\mathbf{E}\mathbf{S}\mathbf{E}^T + \mathbf{W} \right)^{-1} \mathbf{R}_{nn} \left(\mathbf{E}\mathbf{S}\mathbf{E}^T + \mathbf{W} \right)^{-1} \mathbf{E}\mathbf{S} \\
&= \left(\mathbf{E}^T \mathbf{W}^{-1} \mathbf{E} + \mathbf{S}^{-1} \right)^{-1} \mathbf{E}^T \mathbf{W}^{-1} \mathbf{R}_{nn} \mathbf{W}^{-1} \mathbf{E} \left(\mathbf{E}^T \mathbf{W}^{-1} \mathbf{E} + \mathbf{S}^{-1} \right)^{-1}
\end{aligned}$$

(and can take limits as $\left\| \mathbf{W}^{-1} \right\| \to 0$, etc. $\mathbf{C}_{xx}$ is the covariance about the expected value of the estimate. It is distinguished from the later use of $\mathbf{P}$ to denote the covariance about the true value. If $\langle \tilde{\mathbf{x}} \rangle = \langle \mathbf{x} \rangle$, one has a special case of an *unbiassed* estimator.

What do $\mathbf{S}, \mathbf{W}$ signify? In least-squares they are arbitrary (as long as they have inverses). You can choose anything you want. The motivation for a *special* choice comes from a digression—the Gauss-Markov Theorem.

Gauss-Markov

Consider e.g., a map-making problem in one-dimension to start. We have some data $\mathbf{y} = \{y(r_i)\} = \{x + n_i\} = \mathbf{x} + \boldsymbol{\varepsilon}$, where the measurement noise has zero mean, and a known covariance, $\langle \varepsilon_i \varepsilon_j \rangle = R_{nn,ij}$ (it might be diagonal). We'd like to "map" y onto a regular grid, $\tilde{r}_\alpha$. Call the mapped value $\tilde{x} = \tilde{x}(\tilde{r}_\alpha)$. We think that x has zero mean and has a covariance $\langle x(\tilde{r}_\alpha) x(r_\beta) \rangle = R(\tilde{r}_\alpha, r_\beta)$. (Maybe we think it has a large-scale.) The mapped value $\tilde{x}(\tilde{r}_\alpha)$ is assumed to be some linear combination of the data points,

$$\tilde{x}(\tilde{r}_\alpha) = \sum_j b_j(\tilde{r}_\alpha) y(r_j) = \mathbf{b}^T(\tilde{r}_\alpha) \mathbf{y} = \mathbf{y}^T \mathbf{b}.$$

(Linear interpolation is a special case, as is a spline, etc.). How should we choose the b_j? Let's demand that

$$\left\langle (\tilde{x}(\tilde{r}_\alpha) - x(\tilde{r}_\alpha))^2 \right\rangle$$

is to be as small as possible (an arbitrary, but reasonable choice). Then we want to choose $\mathbf{b}$ to obtain the minimum of

$$\begin{aligned} J &= \left\langle \left(\mathbf{b}^T (\tilde{r}_\alpha) \mathbf{y} - x(\tilde{r}_\alpha) \right)^2 \right\rangle \\ &= \mathbf{b}^T (\tilde{r}_\alpha) \left\langle \mathbf{y}\mathbf{y}^T \right\rangle \mathbf{b} - 2\mathbf{b}^T (\tilde{r}_\alpha) \langle \mathbf{y} x(\tilde{r}_\alpha) \rangle + \left\langle x(\tilde{r}_\alpha)^2 \right\rangle . \end{aligned}$$

So take the partial derivatives,

$$\frac{\partial J}{\partial \mathbf{b}} = 2 \left\langle \mathbf{y}\mathbf{y}^T \right\rangle \mathbf{b} - 2 \langle \mathbf{y} x(\tilde{r}_\alpha) \rangle = 0,$$

or

$$\begin{aligned} &\left\langle \mathbf{y}\mathbf{y}^T \right\rangle \mathbf{b}(\tilde{r}_\alpha) = \langle \mathbf{y} x(\tilde{r}_\alpha) \rangle = \langle (\mathbf{x} + \boldsymbol{\varepsilon}) x(\tilde{r}_\alpha) \rangle \\ = &\left\langle \mathbf{x}\mathbf{x}^T + \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T \right\rangle \mathbf{b}(\tilde{r}_\alpha) = \langle \mathbf{x} x(\tilde{r}_\alpha) \rangle \end{aligned}$$

assuming, $\left\langle \boldsymbol{\varepsilon}\mathbf{x}^T \right\rangle = 0$. Or,

$$\left(\mathbf{R}_{xx}(r_i, r_j) + \mathbf{R}_{nn}(r_i, r_j) \right) \mathbf{b} = \mathbf{R}_{xx}(r_i, \tilde{r}_\alpha)$$

and so,

$$\mathbf{b} = \left(\mathbf{R}_{xx}(r_i, r_j) + \mathbf{R}_{nn}(r_i, r_j) \right)^{-1} \mathbf{R}_{xx}(r_i, \tilde{r}_\alpha)$$

Usually want to find the best value at several points. Let's do all at once, by considering a collection of points $\tilde{r}_\alpha, \quad \alpha = 1, 2, ..., N$. Make a matrix whose *rows* correspond to each point,

$$\mathbf{B} = \left\{ \begin{array}{c} \mathbf{b}\left(\tilde{r}_\alpha\right)^T \\ \mathbf{b}\left(\tilde{r}_{\alpha+1}\right)^T \\ . \\ \mathbf{b}\left(\tilde{r}_M\right)^T \end{array} \right\}$$

Then can do all at once by writing $\tilde{\mathbf{x}} = \mathbf{By}$, and minimizing the *diagonal* elements (so it's M separate problems) of

$$\mathbf{P}\left(\tilde{\mathbf{r}}_\alpha, \mathbf{r}_j\right) = \left\langle \left(\tilde{\mathbf{x}} - \mathbf{x}\right)\left(\tilde{\mathbf{x}} - \mathbf{x}\right)^T \right\rangle = \left\langle \left(\mathbf{By} - \mathbf{x}\right)\left(\mathbf{By} - \mathbf{x}\right)^T \right\rangle$$

$$= \mathbf{B}\left\langle \mathbf{y}\mathbf{y}^T \right\rangle \mathbf{B}^T - \left\langle \mathbf{x}\mathbf{y}^T \right\rangle \mathbf{B}^T - \mathbf{B}\left\langle \mathbf{y}\mathbf{x}^T \right\rangle + \left\langle \mathbf{x}\mathbf{x}^T \right\rangle$$

$$\begin{aligned} &= \mathbf{B}\mathbf{R}_{yy}\mathbf{B}^T - \mathbf{R}_{xy}\mathbf{B}^T - \mathbf{B}\mathbf{R}_{xy}^T + \mathbf{R}_{xx} \\ &= (\mathbf{B} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1})\mathbf{R}_{yy}(\mathbf{B} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1})^T - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{xy}^T + \mathbf{R}_{xx} \end{aligned}$$

Skipping some steps, the minimizer of the diagonal is actually the obvious choice,

$$\mathbf{B} = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}$$

The subcripts indicate generally, $\mathbf{R}_{xy} = \left\langle \mathbf{x}\mathbf{y}^T \right\rangle$, etc.

For mapping, we have $\mathbf{R}_{yy} = \mathbf{R}_{xx} + \mathbf{R}_{nn}$, $\mathbf{R}_{xy} = \mathbf{R}_{xx}$ (assuming $\mathbf{R}_{xn} = 0$). So the mapping problem is solved by

$$\tilde{\mathbf{x}}(\tilde{\mathbf{r}}_\alpha) = \mathbf{R}_{xx}\left(\mathbf{R}_{xx} + \mathbf{R}_{nn}\right)^{-1}\mathbf{y}\left(\mathbf{r}_j\right)$$

and we know the uncertainty,

$$\mathbf{P} = \mathbf{R}_{xx} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{xy}^T$$

(the various arguments in the $\mathbf{R}$ matrices involves both $\tilde{r}_\alpha$, and r_i. Extending it to two or more dimensions changes nothing except that the scalar r_i becomes a vector $\mathbf{r}_i$. $\mathbf{P}$ shows that the maximum possible variance is $\mathbf{R}_{xx}$. This procedure is called *objective mapping.* Meteorologists mis-call it OI for *optimal interpolation.* But it's neither of those things. (Also called 3DVAR.) Easy to see that $\tilde{x}\left(r_i\right) \neq y\left(r_i\right)$ ("interpolation" as in linear or splines or anything else requires that at the data point, the estimated value is the same as the measured one). It also

wouldn't be optimal unless one really knew the $\mathbf{R}$ matrices—which is almost never true. Two advantages: (1) it is repeatable, and (2), there is an estimate of the accuracy. (It is common for people to simply guess covariances written as functions, like $\mathbf{R}_{xx,ij} = A \exp\left(-\left(r_i - r_j\right)^2 / b^2\right)$), etc.)

This Gauss-Markov result is very general, not confined to mapping. Let's use it on $\mathbf{y} = \mathbf{E}\mathbf{x}+\boldsymbol{\varepsilon}$, so that e.g., $\mathbf{R}_{yy} = \left\langle (\mathbf{E}\mathbf{x}+\boldsymbol{\varepsilon}) (\mathbf{E}\mathbf{x}+\boldsymbol{\varepsilon})^T \right\rangle = \mathbf{E}\mathbf{R}_{xx}\mathbf{E}^T + \mathbf{R}_{nn}$, $\mathbf{R}_{xy} = \langle \mathbf{x}(\mathbf{E}\mathbf{x}+\boldsymbol{\varepsilon})^T \rangle = \mathbf{R}_{xx}\mathbf{E}^T$,and we now have,

$$\tilde{\mathbf{x}} = \mathbf{R}_{xx}\mathbf{E}^T \left(\mathbf{E}\mathbf{R}_{xx}\mathbf{E}^T + \mathbf{R}_{nn}\right)^{-1} \mathbf{y} \tag{12}$$

$$\tilde{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}} \tag{13}$$

$$\mathbf{P} = \mathbf{R}_{xx} - \mathbf{R}_{xx}\mathbf{E}^T \left(\mathbf{E}\mathbf{R}_{xx}\mathbf{E}^T + \mathbf{R}_{nn}\right)^{-1} \mathbf{E}\mathbf{R}_{xx} \tag{14}$$

$$= \left(\mathbf{R}_{xx}^{-1} + \mathbf{E}^T\mathbf{R}_{nn}^{-1}\mathbf{E}\right)^{-1} \tag{15}$$

where the last relationship is from the matrix inversion lemma (MIL).
This is the minimum variance (maximum likelihood) solution to a set of simultaneous equations.

IMPORTANT INFERENCE:

The solution is *identical* to the least-squares solution for $J = \boldsymbol{\varepsilon}^T\mathbf{W}^{-1}\boldsymbol{\varepsilon} + \mathbf{x}^T\mathbf{S}^{-1}\mathbf{x}$:

$$\tilde{\mathbf{x}} = \mathbf{S}\mathbf{E}^T \left(\mathbf{E}\mathbf{S}\mathbf{E}^T + \mathbf{W} \right)^{-1} \mathbf{y}, \tag{16}$$

$$\tilde{\varepsilon} = \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}}, \tag{17}$$

$$\mathbf{C}_{xx} = \mathbf{S}\mathbf{E}^T \left(\mathbf{E}\mathbf{S}\mathbf{E}^T + \mathbf{W} \right)^{-1} \mathbf{R}_{nn} \left(\mathbf{E}\mathbf{S}\mathbf{E}^T + \mathbf{W} \right)^{-1} \mathbf{E}\mathbf{S}. \tag{18}$$

if we choose $\mathbf{S} = \mathbf{R}_{xx}, \mathbf{W} = \mathbf{R}_{nn}$ (note that $\mathbf{P}$ and $\mathbf{C}_{xx}$ are different) and must also be identical to the alternate least-squares solution (again using the MIL),

$$\tilde{\mathbf{x}} = \left(\mathbf{E}^T \mathbf{R}_{nn}^{-1} \mathbf{E} + \mathbf{R}_{xx}^{-1} \right)^{-1} \mathbf{E}^T \mathbf{R}_{nn}^{-1} \mathbf{y}, \tag{19}$$

$$\tilde{\varepsilon} = \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}}, \tag{20}$$

$$\mathbf{C}_{xx} = \tag{21}$$

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

The form is usually chosen so as to minimize the dimension of matrices being inverted—assuming that dominates the calculation (it need not). Notice that if $\|\mathbf{R}_{nn}\| \to \infty$, that $\tilde{\mathbf{x}} \to 0$ and $\mathbf{P} = \mathbf{R}_{xx}$ (cannot be any worse than its own variance). Etc.

The logic of the derivations is entirely different. *It is a great convenience* that least-squares produces the Gauss-Markov solution, but one must understand why one is using it—not because it's least-squares but usually because it reproduces the G-M solution—by construction.

Inverse Problems, Inverse Methods, State Estimation, Data Assimilation, and All That

Short Course 2013-Lecture 4.

Carl Wunsch
Harvard University

January 23, 2013

Consider *any* problem in which an estimate $\tilde{\mathbf{x}}$ might logically be written as a linear combination of some observations $\mathbf{y}$, $\tilde{\mathbf{x}} = \mathbf{B}\mathbf{y}$. Logical to minimize the *diagonal* elements (so it's M separate problems for each element x_i) of,

$$\mathbf{P}\left(\tilde{\mathbf{r}}_\alpha, \mathbf{r}_j\right) = \left\langle (\tilde{\mathbf{x}} - \mathbf{x})\,(\tilde{\mathbf{x}} - \mathbf{x})^T \right\rangle = \left\langle (\mathbf{B}\mathbf{y} - \mathbf{x})\,(\mathbf{B}\mathbf{y} - \mathbf{x})^T \right\rangle$$

$$= \mathbf{B}\left\langle \mathbf{y}\mathbf{y}^T \right\rangle \mathbf{B}^T - \left\langle \mathbf{x}\mathbf{y}^T \right\rangle \mathbf{B}^T - \mathbf{B}\left\langle \mathbf{y}\mathbf{x}^T \right\rangle + \left\langle \mathbf{x}\mathbf{x}^T \right\rangle$$

$$\begin{aligned} &= \mathbf{B}\mathbf{R}_{yy}\mathbf{B}^T - \mathbf{R}_{xy}\mathbf{B}^T - \mathbf{B}\mathbf{R}_{xy}^T + \mathbf{R}_{xx} \\ &= (\mathbf{B} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1})\mathbf{R}_{yy}(\mathbf{B} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1})^T - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{xy}^T + \mathbf{R}_{xx} \end{aligned}$$

Skipping some steps, the minimizer of the diagonal is actually the obvious choice,

$$\mathbf{B} = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}$$

The subcripts indicate generally, $\mathbf{R}_{xy} = \left\langle \mathbf{x}\mathbf{y}^T \right\rangle$, etc.

For mapping, we have $\mathbf{R}_{yy} = \mathbf{R}_{xx} + \mathbf{R}_{nn}$, $\mathbf{R}_{xy} = \mathbf{R}_{xx}$ (assuming $\mathbf{R}_{xn} = 0$). So

$$\tilde{\mathbf{x}} = \mathbf{R}_{xx}\left(\mathbf{R}_{xx} + \mathbf{R}_{nn}\right)^{-1}\mathbf{y}$$

and we know the uncertainty,

$$\mathbf{P} = \mathbf{R}_{xx} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{xy}^{T}$$

$\mathbf{P}$ shows that the maximum possible variance is $\mathbf{R}_{xx}$—if you say that the best estimate is zero—there being no additional information contained in $\mathbf{y}$.

Try it on $\mathbf{y} = \mathbf{E}\mathbf{x}+\boldsymbol{\varepsilon}$, where it's reasonable to think that $\mathbf{x}$ depends linearly on $\mathbf{y}$.
Then $\mathbf{R}_{yy} = \left\langle (\mathbf{E}\mathbf{x}+\boldsymbol{\varepsilon})(\mathbf{E}\mathbf{x}+\boldsymbol{\varepsilon})^T \right\rangle = \mathbf{E}\mathbf{R}_{xx}\mathbf{E}^T + \mathbf{R}_{nn}$,
$\mathbf{R}_{xy} = \langle \mathbf{x}(\mathbf{E}\mathbf{x}+\boldsymbol{\varepsilon})^T \rangle = \mathbf{R}_{xx}\mathbf{E}^T$, and

$$\tilde{\mathbf{x}} = \mathbf{R}_{xx}\mathbf{E}^T \left(\mathbf{E}\mathbf{R}_{xx}\mathbf{E}^T + \mathbf{R}_{nn}\right)^{-1} \mathbf{y} \tag{1}$$

$$\tilde{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}} \tag{2}$$

$$\mathbf{P} = \mathbf{R}_{xx} - \mathbf{R}_{xx}\mathbf{E}^T \left(\mathbf{E}\mathbf{R}_{xx}\mathbf{E}^T + \mathbf{R}_{nn}\right)^{-1} \mathbf{E}\mathbf{R}_{xx} \tag{3}$$

$$= \left(\mathbf{R}_{xx}^{-1} + \mathbf{E}^T\mathbf{R}_{nn}^{-1}\mathbf{E}\right)^{-1} \tag{4}$$

where the last relationship is from the matrix inversion lemma (MIL).
This is the minimum variance (maximum likelihood) solution to a set of simultaneous equations.

IMPORTANT INFERENCE:

The solution is *identical* to the least-squares solution for $J = \boldsymbol{\varepsilon}^T\mathbf{W}^{-1}\boldsymbol{\varepsilon} + \mathbf{x}^T\mathbf{S}^{-1}\mathbf{x}$:

$$\tilde{\mathbf{x}} = \mathbf{S}\mathbf{E}^T \left(\mathbf{E}\mathbf{S}\mathbf{E}^T + \mathbf{W} \right)^{-1} \mathbf{y}, \tag{5}$$

$$\tilde{\varepsilon} = \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}}, \tag{6}$$

$$\mathbf{C}_{xx} = \mathbf{S}\mathbf{E}^T \left(\mathbf{E}\mathbf{S}\mathbf{E}^T + \mathbf{W} \right)^{-1} \mathbf{R}_{nn} \left(\mathbf{E}\mathbf{S}\mathbf{E}^T + \mathbf{W} \right)^{-1} \mathbf{E}\mathbf{S}. \tag{7}$$

if we choose $\mathbf{S} = \mathbf{R}_{xx}, \mathbf{W} = \mathbf{R}_{nn}$ (note that $\mathbf{P}$ and $\mathbf{C}_{xx}$ are different) and must also be identical to the alternate least-squares solution (again using the MIL),

$$\tilde{\mathbf{x}} = \left(\mathbf{E}^T \mathbf{R}_{nn}^{-1} \mathbf{E} + \mathbf{R}_{xx}^{-1} \right)^{-1} \mathbf{E}^T \mathbf{R}_{nn}^{-1} \mathbf{y}, \tag{8}$$

$$\tilde{\varepsilon} = \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}}, \tag{9}$$

$$\mathbf{C}_{xx} = \tag{10}$$

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

The form is usually chosen so as to minimize the dimension of matrices being inverted—assuming that dominates the calculation (it need not). Notice that if $\|\mathbf{R}_{nn}\| \to \infty$, that $\tilde{\mathbf{x}} \to 0$ and $\mathbf{P} = \mathbf{R}_{xx}$ (cannot be any worse than its own variance). Etc.

The logic of the derivations is entirely different. *It is a great convenience* that least-squares produces the Gauss-Markov solution, but one must understand why one is using it—not because it's least-squares but usually because it reproduces the G-M solution—by construction.

Example.
Consider the problem of estimating a mean value. Suppose that we have a set of measurements, $y_i = m + \varepsilon_i$ where the "noise" is probably the signal, but never mind. This is a set of simultaneous equations

$$\left\{ \begin{array}{c} 1 \\ 1 \\ . \\ . \\ 1 \\ 1 \end{array} \right\} m + \left[\begin{array}{c} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_{M-1} \\ \varepsilon_M \end{array} \right] = \left[\begin{array}{c} y_1 \\ y_2 \\ . \\ . \\ y_{M-1} \\ y_M \end{array} \right]$$

Suppose that $R_{xx} = m_0^2$ (perhaps very large). Then the solution formula produces

$$\tilde{m} = \left\{ \frac{1}{m_0^2} + \mathbf{E}^T \mathbf{R}_{nn}^{-1} \mathbf{E} \right\}^{-1} \mathbf{E}^T \mathbf{R}_{nn}^{-1} \mathbf{y}$$

(using the alternate form because only a scalar needs to be inverted).

Now suppose $\mathbf{R}_{nn} = \sigma_\varepsilon^2 \mathbf{I}$ ("white noise"), then

$$\begin{aligned} \tilde{m} &= \left\{ \frac{1}{m_0^2} + \sigma_\varepsilon^2 \mathbf{E}^T \mathbf{E} \right\}^{-1} \mathbf{E}^T \mathbf{R}_{nn}^{-1} \mathbf{y} \\ &= \frac{\sigma_\varepsilon^2}{1/m_0^2 + M\sigma_\varepsilon^2} \sum_{j=1}^{M} y_j \end{aligned}$$

and if $m_0^2 \to \infty$, $\tilde{m} = 1/M \sum_{j=1}^{M} y_j$, the ordinary average. The uncertainty of the estimate is

$$P = \left\{ \frac{1}{m_0^2} + \mathbf{E}^T \mathbf{R}_{nn}^{-1} \mathbf{E} \right\}^{-1} = \frac{1}{1/m_0^2 + \mathbf{E}^T \mathbf{R}_{nn}^{-1} \mathbf{E}} \to \sigma_\varepsilon^2 / M$$

in the same limit and using the alternate form. This last result is the classical statement that the standard error of the mean goes like $\sigma_\varepsilon / \sqrt{M}$.
(A good exercise is to work this problem through for trend determination—that is fitting a straight line and computing the uncertainty of the result.)

A further comment about the estimate of the mean. The example used prior information only about the variance of x (m_0^2) which says nothing about whether it might be positive or negative. In practice, one might suspect that $m = m_{prior}$. The easiest way to use that information is to subtract m_{prior} from the data, and then solve for its correction, $m = m_{prior} + \Delta m$. The value of m_0^2 would then be the estimate of the variance of Δm for which the sign would be irrelevant.

Recursive Least-Squares

Suppose we have solved a big, $M \times N$, least-squares problem,

$$\mathbf{E}\mathbf{x} + \boldsymbol{\varepsilon} = \mathbf{y}$$

and gotten a solution $\tilde{\mathbf{x}} = \tilde{\mathbf{x}}(1)$, along with its uncertainty, $\mathbf{P} = \mathbf{P}(1)$. Someone offers us another measurement, $\mathbf{e}_{M+1}^T \mathbf{x} + \varepsilon_{M+1} = y_{N+1}$, where $\left\langle \varepsilon_{M+1}^2 \right\rangle = R(M+1)$. One approach to using it is simply to make a new $(M+1) \times N$ problem,

$$\mathbf{E}_1 \mathbf{x} + \boldsymbol{\varepsilon} = \mathbf{y}_1, \quad \mathbf{E}_1 = \left\{ \begin{array}{c} \mathbf{E} \\ \mathbf{e}_{M+1}^T \end{array} \right\}, \text{ etc.} \tag{11}$$

and recompute. But having done all that work, do we really need to repeat everything? The answer is no, and leads to the notion of *recursive least-squares.*

The derivation proceeds by manipulating the partitioned problem (11) and produces (see P. 137 of DISEP):

$$\begin{aligned}
& \tilde{\mathbf{x}}(2) = \tilde{\mathbf{x}}(1) + \\
& \mathbf{P}(1)\mathbf{e}_{M+1}\left[\mathbf{e}_{M+1}\mathbf{P}(1)\mathbf{e}_{M+1}^T + R(M+1)\right]^{-1}\left(q_{N+1} - \mathbf{e}_{M+1}^T\tilde{\mathbf{x}}(1)\right) \\
= \quad & \tilde{\mathbf{x}}(1) + \mathbf{K}(2)\left(q_{N+1} - \mathbf{e}_{M+1}^T\tilde{\mathbf{x}}(1)\right) \\
\mathbf{P}(2) = \quad & \mathbf{P}(1) - \mathbf{K}(2)\mathbf{e}_{M+1}\mathbf{P}(1)
\end{aligned}$$

(NOTICE THAT THE PREVIOUS DATA HAVE DISAPPEARED! ALL INFORMATION IS CONTAINED IN $\tilde{\mathbf{x}}(1)$ and $\mathbf{P}(1)$.)
More generally, if there is a previous solution $\tilde{\mathbf{x}}(1)$ of uncertainty $\mathbf{P}(1)$, obtained from anywhere, and if there are $N(2)$ additional measurements

$$\mathbf{E}(2)\mathbf{x} + \varepsilon(2) = \mathbf{y}(2)$$

of noise covariance $\mathbf{R}(2)$, the same logic produces,

$$\begin{aligned}
\tilde{\mathbf{x}}(2) &= \tilde{\mathbf{x}}(1) + \mathbf{P}(1)\mathbf{E}(2)^T \left[\mathbf{E}(2)\mathbf{P}(1)\mathbf{E}(2)^T + \mathbf{R}(2)\right]^{-1} (\mathbf{y}(2) - \mathbf{E}(2)\tilde{\mathbf{x}}(1)) \\
&= \tilde{\mathbf{x}}(1) + \mathbf{K}(2)(\mathbf{y}(2) - \mathbf{E}(2)\tilde{\mathbf{x}}(1)) \\
\mathbf{P}(2) &= \mathbf{P}(1) - \mathbf{K}(2)\mathbf{E}(2)\mathbf{P}(1)
\end{aligned}$$

And do it again....$\mathbf{K}(2)$ is the *gain matrix.* Notice that $\tilde{\mathbf{x}}(2)$ is unchanged from $\tilde{\mathbf{x}}(1)$, if the predicted extra observations. $\mathbf{E}(2)\tilde{\mathbf{x}}(1)$ coincide with what is observed, $\mathbf{y}(2)$. The uncertainty would, however, sensibly change.
This result is almost the Kalman filter. Coming... Note that (1), (2) do *not* imply successive times, only two different sets of observations.

Recapitulation: To find an estimate of $\mathbf{x}$,$\boldsymbol{\varepsilon}$, given

$$\mathbf{E}\mathbf{x}+\boldsymbol{\varepsilon}=\mathbf{y}$$

$$\langle \mathbf{x} \rangle = 0, \ \langle \boldsymbol{\varepsilon} \rangle = \mathbf{0}, \ \mathbf{R}_{xx} = \left\langle \mathbf{x}\mathbf{x}^T \right\rangle, \ \mathbf{R}_{nn} = \left\langle \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T \right\rangle$$

can minimize

$$J = \mathbf{x}^T \mathbf{R}_{xx}^{-1} \mathbf{x} + \boldsymbol{\varepsilon}^T \mathbf{R}_{nn}^{-1} \boldsymbol{\varepsilon}$$

and which will coincide with the Gauss-Markov or minimum variance solution.

Let's look at a similar problem in the G-M framework. Suppose there is an estimated solution with known uncertainty, $\tilde{\mathbf{x}}_1, \mathbf{P}_1$.An independent second estimate $\tilde{\mathbf{x}}_2, \mathbf{P}_2$ becomes available, and one seeks to combine them into a best solution. It is not difficult to show that the minimum variance estimate is the weighted average,

$$\begin{aligned} \tilde{\mathbf{x}}_3 &= \mathbf{P}_2 \left(\mathbf{P}_1 + \mathbf{P}_2\right)^{-1} \tilde{\mathbf{x}}_1 + \mathbf{P}_1 \left(\mathbf{P}_1 + \mathbf{P}_2\right)^{-1} \tilde{\mathbf{x}}_2 \\ &= \tilde{\mathbf{x}}_1 + \mathbf{P}_1 \left(\mathbf{P}_1 + \mathbf{P}_2\right)^{-1} \left(\tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2\right) \\ &= \tilde{\mathbf{x}}_1 + \mathbf{K}_2 \left(\tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2\right) \end{aligned} \tag{12}$$

with new uncertainty,

$$\begin{aligned} \mathbf{P}_3 &= \mathbf{P}_1 - \mathbf{P}_1 \left(\mathbf{P}_1 + \mathbf{P}_2\right) \mathbf{P}_1 \\ &= \mathbf{P}_1 - \mathbf{K}\left(2\right) \mathbf{P}_1 \\ &= \left(\mathbf{P}_1^{-1} + \mathbf{P}_2^{-1}\right)^{-1} \end{aligned}$$

Notice that the combined solution behaves sensibly as the norm of $\mathbf{P}_1$ or of $\mathbf{P}_2$ go to zero or infinity. The 3rd line of Eq. (12) reduces properly, as it must, but somewhat hides what is still a simple weighted average. The recursive least-squares and G-M updated solutions are identical if the weights are properly chosen in the former.

The Kalman Filter

Suppose we have a system that evolves in time (but it could equally well be space. "Time" is purely conventional.) . Define the system *state* as that information required to make a prediction an infinitesmal time, Δt, in the future, along with any externally prescribed conditions. So e.g., for the mass-spring oscillator,

$$m\frac{d^2y\left(t\right)}{dt^2}+r\frac{dy\left(t\right)}{dt}+ky\left(t\right)=q(t), \tag{13}$$

$y(t)$ is the state variable (the position) at a continuum of times so that it's first derivative, the velocity, $y'\left(t\right)$ is also known. With knowledge of $q\left(t\right)$ (externally prescribed), the equation permits the calculation dt into the future. As with everything else in these lectures, we prefer to assume that there is an adequate discretization e.g. (not unique),

$$\xi(t+\Delta t) = \left(2 - \frac{r\Delta t}{m} - \frac{k(\Delta t)^2}{m}\right) \xi(t) + \left(\frac{r\Delta t}{m} - 1\right) \xi(t-\Delta t) + (\Delta t)^2 \frac{q(t)}{m}$$

This says given,$\xi(t), \xi(t-\Delta t)$we can compute $\xi(t+\Delta t)$, and at any future time. So the *state vector* is $\mathbf{x}(t) = [\xi(t), \xi(t-\Delta t)]^T$ (because the discretization is not unique, neither is the state vector). The above equation permits time-stepping (if two starting conditions are known) and is, of course, nothing but a set of simultaneous equations (for all of the $\xi(q)$,all q), even if the extra conditions are not specified at $t = 0$. From here, set $\Delta t = 1$.

Rewrite this in *state space form:*

$$\begin{bmatrix} \xi(t) \\ \xi(t-\Delta t) \end{bmatrix} = \begin{Bmatrix} 2 - \frac{r}{m}\Delta t - \frac{k}{m}(\Delta t)^2 & \frac{r\Delta t}{m} - 1 \\ 1 & 0 \end{Bmatrix} \begin{bmatrix} \xi(t-\Delta t) \\ \xi(t-2\Delta t) \end{bmatrix} + \begin{bmatrix} (\Delta t)^2 \frac{q(t-\Delta t)}{m} \\ 0 \end{bmatrix}$$

$$\mathbf{x}(t) = \begin{bmatrix} \xi(t) & \xi(t-\Delta t) \end{bmatrix}^T, \qquad \mathbf{B}(t)\mathbf{q}(t) = \begin{bmatrix} (\Delta t)^2 q(t)/m & 0 \end{bmatrix}^T.$$

This is a canonical form:

$$\mathbf{x}(t) = \mathbf{A}(t)\mathbf{x}(t-1) + \mathbf{B}\mathbf{q}(t-1)$$

(may prefer to write as $\mathbf{x}(t+1) = \ldots$) Can show that *any* linear model can be put into this form (try it with any of your favorite ordinary or partial differential equations).

Note that *any* non-linear model is still a time-stepping rule:

$$\mathbf{x}(t) = \mathbf{L}(\mathbf{x}(t-1), \mathbf{Bq}(t-1))$$

and, as with the linear model, is again just a set of simultaneous equations:

$$\begin{aligned}
\mathbf{x}(1) &= \mathbf{L}(\mathbf{x}(0), \mathbf{Bq}(0)) \\
\mathbf{x}(2) &= \mathbf{L}(\mathbf{x}(1), \mathbf{Bq}(1)) \\
&\cdot \\
\mathbf{x}(t_f) &= \mathbf{L}(\mathbf{x}(t_f-1), \mathbf{Bq}(t_f-1))
\end{aligned}$$

although non-linear. (For now, we confine ourselves to the linear system, and with $\mathbf{A}(t), \mathbf{B}(t)$ being time-independent: time-independence saves a lot of writing without any essential loss of generality).

Suppose, somehow, we have an estimate $\tilde{\mathbf{x}}(t-1)$, with uncertainty, $\mathbf{P}(t-1)$. At time t, we have some observations of $\mathbf{x}(t)$, with error, written as,

$$\mathbf{E}(t)\mathbf{x}(t)+\boldsymbol{\varepsilon}(t)=\mathbf{y}(t)$$

Can we combine the observations with the state vector at the previous time to take advantage of the information in both?
Let's use the model to make a prediction of the state vector at the observation time. Suppose too, that the $\mathbf{Bq}(t)$ are partially known, but have an unpredictable component, so that the model is of the form,

$$\mathbf{x}(t)=\mathbf{A}(t)\mathbf{x}(t-1)+\mathbf{Bq}(t-1)+\boldsymbol{\Gamma}\mathbf{u}(t-1)$$

where $\langle \mathbf{u}(t)\rangle = 0$, $\left\langle \mathbf{u}(t)\mathbf{u}(t)^T\right\rangle = \mathbf{Q}(t)$. Will drop the t in $\mathbf{Q}$, for convenience. ($\boldsymbol{\Gamma}$ can be used to describe known spatial dependencies in the "control" $\mathbf{u}(t)$. At $t-1$, we don't know $\mathbf{x}(t)=\tilde{\mathbf{x}}(t-1)+\boldsymbol{\gamma}(t-1)$, where by definition,

$$\left\langle \boldsymbol{\gamma}(t-1)\boldsymbol{\gamma}(t-1)^T\right\rangle = \mathbf{P}(t-1)$$

The best prediction we can make is probably,

$$\tilde{\mathbf{x}}(t,-) = \mathbf{A}\tilde{\mathbf{x}}(t-1) + \mathbf{B}\mathbf{q}(t-1) + \mathbf{0},$$

replacing the unknown control by its zero mean. The minus sign is put into the argument to show that it is a pure prediction (the observations not yet used). How good is this? Let's find the expected error in the prediction,

$$\begin{aligned} & \left\langle (\tilde{\mathbf{x}}(t,-) - \mathbf{x}(t)) (\tilde{\mathbf{x}}(t,-) - \mathbf{x}(t))^T \right\rangle = \mathbf{P}(t,-) \\ = \quad & \mathbf{A} \left\langle \boldsymbol{\gamma}(t-1) \boldsymbol{\gamma}(t-1)^T \right\rangle \mathbf{A}^T + \boldsymbol{\Gamma} \left\langle \mathbf{u}(t-1) \mathbf{u}(t-1)^T \right\rangle \boldsymbol{\Gamma}^T \\ = \quad & \mathbf{A}\mathbf{P}(t-1)\mathbf{A}^T + \boldsymbol{\Gamma}\mathbf{Q}(t-1)\boldsymbol{\Gamma}^T \end{aligned}$$

The "prediction error" is made up of two pieces—a part from the erroneous "initial condition" and a part from the unknown external driving (control). If one must integrate multiple time steps into the future before observations become available, one simply loops on the above equation, propagating the error forward.

Now make an estimate of $\tilde{\mathbf{x}}(t)$ from the data alone using e.g., the Gauss-Markov estimate, but with $\mathbf{R}_{xx}^{-1} = 0$ (so it's independent of anything coming before—complete initial uncertainty). Then from the previous recursive least-squares, or recursive Gauss-Markov, solution,

$$\begin{aligned}
\tilde{\mathbf{x}}^{+}(t) &= \tilde{\mathbf{x}}(t,-) + \\
&\quad \mathbf{P}(t,-)\mathbf{E}(t)^{T}[\mathbf{E}(t)\mathbf{P}(t,-)\mathbf{E}(t)^{T} + \mathbf{R}_{nn}(t)]^{-1}(\mathbf{y}(t) - \mathbf{E}(t)\tilde{\mathbf{x}}(t,-)), \\
\mathbf{P}^{+}(t) &= \left(\mathbf{P}(t,-)^{-1} + \mathbf{E}(t)^{T}\mathbf{R}_{nn}(t)^{-1}\mathbf{E}(t)\right)^{-1} = \mathbf{P}(t) \\
&= \mathbf{P}(t,-) - \mathbf{K}(t)\mathbf{E}(t)\mathbf{P}(t,-) \\
\mathbf{K}(t) &= \mathbf{P}(t,-)\mathbf{E}(t)^{T}[\mathbf{E}(t)\mathbf{P}(t,-)\mathbf{E}(t)^{T} + \mathbf{R}_{nn}(t)]^{-1}
\end{aligned}$$

This algorithm is the famous Kalman filter (KF). (In the estimation literature, a *filter, usually* refers to the best estimate of the "now" state or to prediction—the "prediction filter".)

$$\tilde{\mathbf{x}}(t) = \tilde{\mathbf{x}}(t,-) + \mathbf{K}(t)\left(\mathbf{y}(t) - \mathbf{E}(t)\tilde{\mathbf{x}}(t,-)\right),$$
$$\mathbf{P}(t,-) = \mathbf{A}\mathbf{P}(t-1)\mathbf{A}^T + \mathbf{\Gamma}\mathbf{Q}(t-1)\mathbf{\Gamma}^T$$
$$\mathbf{P}(t) = \mathbf{P}(t-1) - \mathbf{K}(t)\mathbf{E}(t)\mathbf{P}(t-1)$$
$$\mathbf{K}(t) = \mathbf{P}(t,-)\mathbf{E}(t)^T\left[\mathbf{E}(t)\mathbf{P}(t,-)\mathbf{E}(t)^T + \mathbf{R}_{nn}(t)\right]^{-1}$$

Can be rewritten in different ways for accuracy, efficiency etc. Tens of thousands of papers on it. Note that $\mathbf{P}(t)$ and hence, $\mathbf{K}(t)$, evolves with time in a way dependent, among other elements $\mathbf{E}(t)$ and $\mathbf{R}_{nn}(t)$.
Sometimes with a steady observational stream, $\mathbf{E}(t) = \mathbf{E}$, $\mathbf{P}(t)$ ultimately becomes steady,

$$\mathbf{P}_\infty = \mathbf{P}_\infty - \mathbf{K}_\infty\mathbf{E}\mathbf{P}_\infty$$

a matrix Riccati equation with its own extensive literature. $\mathbf{K}$ is then also steady. But with time dependencies in the observation distribution or technology, such a steady-state will not normally be obtained.

There are ways of deriving the KF that look entirely different and more impressive. In practice, it is nothing but a linear, weighted combination of a model forecast, with an estimate made from observations. It is not hard to prove that $\tilde{\mathbf{x}}(t,-)$ is the prediction with the minimum error variance (the best possible predictor in the G-M sense). This scheme was published by Kalman (1960) to predict locations of ballistic missile impacts from observations. It is part of the *control problem*: what should you do to the system to make it land in the right place (or better, to land an airplane or a drone or lunar lander?) You need to know where it will be if you don't intervene (prediction).

On the other hand, look at the general problem from the point of view of a finite time span, $0 \leq t \leq t_f$:

$$\begin{aligned}
\mathbf{x}(0) &= \mathbf{x}_0, \quad \text{with } \mathbf{P}(0) \\
\mathbf{x}(1) &= \mathbf{A}\mathbf{x}(0) + \mathbf{B}\mathbf{q}(0) + \boldsymbol{\Gamma}\mathbf{u}(0) \\
\mathbf{x}(2) &= \mathbf{A}\mathbf{x}(1) + \mathbf{B}\mathbf{q}(1) + \boldsymbol{\Gamma}\mathbf{u}(1) \\
&\cdots \\
\mathbf{x}(t_f) &= \mathbf{A}\mathbf{x}(t_f - 1) + \mathbf{B}\mathbf{q}(t_f - 1) + \boldsymbol{\Gamma}\mathbf{u}(t_f - 1) \\
\mathbf{E}(1)\mathbf{x}(1) + \boldsymbol{\varepsilon}(1) &= \mathbf{y}(1) \\
\mathbf{E}(2)\mathbf{x}(2) + \boldsymbol{\varepsilon}(2) &= \mathbf{y}(2) \\
&\cdot \\
\mathbf{E}(t_f)\mathbf{x}(t_f) + \boldsymbol{\varepsilon}(t_f) &= \mathbf{y}(t_f)
\end{aligned}$$

Might want to re-write the first equation, instead, as $\mathbf{x}(0) + \boldsymbol{\varepsilon}(0) = \mathbf{x}_0, \mathbf{R}_{nn}(0) = \mathbf{P}(0)$. Each $\boldsymbol{\varepsilon}(t)$ is accompanied by an error covariance, as is each $\mathbf{u}(t)$ and it is assumed that there is no *time* correlation in either $\mathbf{u}(t)$ or $\boldsymbol{\varepsilon}(t)$ (it there is, various methods exist for handling it). The most efficient way to solve the above set of equations, which represents everything one knows, is to proceed with weighted least-squares or G-M solution, if it will fit into your computer. The KF *does not* solve these equations. Consider that it doesn't produce any estimate of $\mathbf{u}(t)$, and that e.g., the estimate made previously for time $\tilde{\mathbf{x}}(t-1)$ is never changed when data come in at time t or later. It's fairly obvious that the early solution might look very different with a later measurement. (Consider the situation when a hurricane is suddenly observed in the middle of an atmospheric region, an observation that was not predicted by the model an hour earlier. One would probably want to modify that earlier estimate.) *But if all one cares about is the best prediction, there is no point in worrying about the earlier state (who cares where the ballistic missile was?)*

The best estimate of the state, $\mathbf{x}(t)$ and of the unknown $\mathbf{u}(t)$ comes from solving the whole equation set as simultaneous equations. If that approach is impractical, how to do that? A *smoother* refers to estimation of the state in the past. An important alternative *algorithm* is called the *RTS smoother*. (RTS is for Rauch, Tung, Striebel who first worked it out.) It consists of running the KF all the way to the end of the data set. There is then no future data to modify that last point. Thus $\tilde{\mathbf{x}}(t_f, +) = \tilde{\mathbf{x}}(t_f)$, with uncertainty, $\mathbf{P}(t_f)$, the $+$ being used to show that future data have been used. One then goes one step backwards in time, knowing that the discrepancy between the value predicted at t_f, which was $\tilde{\mathbf{x}}(t_f, -)$ and the value finally estimated, $\tilde{\mathbf{x}}(t_f)$, had to be due to (A) the missing $\mathbf{u}(t_f)$, and/or the error in $\tilde{\mathbf{x}}(t_f - 1)$ .The RTS smoothing algorithm partitions the corrections between them in a second recursion,

RTS algorithm (see DISEP),

$$
\begin{aligned}
\tilde{\mathbf{x}}(t,+) &= \tilde{\mathbf{x}}(t) + \mathbf{L}(t+1)\left[\tilde{\mathbf{x}}(t+1,+) - \tilde{\mathbf{x}}(t+1,-)\right] \\
\mathbf{L}(t+1) &= \mathbf{P}(t)\mathbf{A}(t)^T \mathbf{P}(t+1,-)^{-1} \\
\mathbf{u}(t,+) &= \mathbf{M}(t+1)\left\{\mathbf{x}(t+1,+) - \mathbf{x}(t+1,-)\right\} \\
\mathbf{M}(t+1) &= \mathbf{Q}(t)\,\mathbf{\Gamma}(t)^T \mathbf{P}(t+1,-)^{-1} \\
\mathbf{P}(t,+) &= \mathbf{P}(t) + \mathbf{L}(t+1)\left[\mathbf{P}(t+1,+) - \mathbf{P}(t+1,-)\right]\mathbf{L}(t+1)^T \\
\mathbf{Q}(t,+) &= \mathbf{Q}(t) + \mathbf{M}(t+1)\left[\mathbf{P}(t+1,+) - \mathbf{P}(t+1,-)\right]\mathbf{M}(t+1)^T
\end{aligned}
$$

The formulas are a bit complex appearing because successive estimates with the KF have temporally correlated errors. The smoothing rotates and stretches the KF solutions to render them uncorrelated with equal variance, averages them appropriately, then rotates and stretches back. The estimated control $\mathbf{u}(t)$ is proportional to the estimated covariance $\mathbf{Q}(t)$ and the value of $\tilde{\mathbf{x}}(t,+)$ is a covariance-weighted average of the KF prediction and the estimated new value using the formally future data. Consider $\mathbf{Q}(t) \rightarrow 0$, etc.

The Kalman filter is in very wide use (it's used to land airplanes and is part of the control circuits of robotic systems everywhere). Its use has remained the goal of many weather forecasters. For a linear system with known covariances it is truly optimal. One pays a huge price for its use however (here, a practical issue!): Consider the computation of $\mathbf{P}(t,-) = \mathbf{A}\mathbf{P}(t-1)\mathbf{A}^T + \mathbf{\Gamma}\mathbf{Q}\mathbf{\Gamma}^T$. The $\mathbf{P}(t)$ matrices are square of the dimension of the state vector. A meteorological model with 4 degree spatial resolution, will have about 22 meridional grid points, 90 zonal ones, and perhaps 25 levels. At each grid point, one must compute 3 components of velocity, the ambient pressure, the water content etc., thus the state vector will be of order $(22)(90)(25)(5) = 10^5 - 10^6$ elements, minimally. Thus the $\mathbf{P}$ matrix is $10^6 \times 10^6$. Running the model one time step into the future (if it were linear), involves multiplication of $\mathbf{x}(t)$ by $\mathbf{A}(t)$ which is also $10^6 \times 10^6$at each time step. Forming $\mathbf{A}\mathbf{P}(t-1)\mathbf{A}^T = \mathbf{A}\left(\mathbf{A}\mathbf{P}(t-1)\right)^T$ is equivalent to running the model $2N$ times at every time step (2×10^6 times). Such a requirement remains far beyond reach (ocean models can have state vectors 100+ times larger).
What to do? (Are still ignoring the nonlinearity.)

Inverse Problems, Inverse Methods, State Estimation, Data Assimilation, and All That. Lecture 5

Carl Wunsch
Harvard University

January 26, 2013

Consider again the Kalman filter (KF), the wholly grail of *prediction*,

$$\underset{forecast}{\tilde{\mathbf{x}}(t,-)} = \mathbf{A} \underset{\text{"initial" condition}}{\tilde{\mathbf{x}}(t-\Delta t)} + \mathbf{B}\mathbf{q}(t-\Delta t) + \underset{\text{unknown control}}{\mathbf{0}},$$

$$\underset{forecast}{\tilde{\mathbf{x}}(t)} = \underset{forecast}{\tilde{\mathbf{x}}(t,-)} + \mathbf{K}(t)\underset{correction-from-obs}{(\mathbf{y}(t) - \mathbf{E}(t)\mathbf{x}(t,-))}$$

$$\begin{aligned}
\underset{forecast-uncertainty}{\mathbf{P}(t,-)} &= \underset{error-from-prev-est}{\mathbf{A}\mathbf{P}(t-\Delta t)\mathbf{A}^T} + \underset{error-from-unknown-control}{\mathbf{\Gamma}\mathbf{Q}(t-\Delta t)\mathbf{\Gamma}^T} \\
\underset{uncertainty-of-combined-est}{\mathbf{P}(t)} &= \mathbf{P}(t,-) - \mathbf{K}(t)\mathbf{E}(t)\mathbf{P}(t,-) \\
\underset{Kalman-gain}{\mathbf{K}(t)} &= \mathbf{P}(t,-)\mathbf{E}(t)^T[\mathbf{E}(t)\mathbf{P}(t,-)\mathbf{E}(t)^T + \mathbf{R}_{nn}(t)]^{-1}
\end{aligned}$$

(These notes are not entirely consistent in the use of the tilde notation. But if $\mathbf{x}(t)$ or $\mathbf{u}(t)$ are calculated, they should technically be $\tilde{\mathbf{x}}(t), \tilde{\mathbf{u}}(t)$ to distinguish them from the true values. Usually the context makes it clear.)

One can try to guess the gain matrix $\mathbf{K}(t)$, typically by assuming it is in a steady-state, and not bother with the error covariances. In weather forecasting, one finds out rather quickly if the forecast was any good and can readily experiment with different guesses. Unfortunately, this is not an attractive option in climate forecasting! Many other approximations have been tried. Some seem to work usefully. These ad hoc weather forecast schemes are what are called *data assimilation*. They are mostly about *prediction* and involve a lot of numerical experimentation and engineering. Textbooks are devoted to their use. They are *not* directed at the problem of making a best estimate of the state vector or the controls and if one is not doing prediction, they are not the right choice. (Furthermore, smoothing, or interpolation, is a far more forgiving process than is prediction or extrapolation.)

The Kalman filter is a (hypothetically) optimum *prediction* method. From a general estimation point of view, it has some extremely undesirable properties:

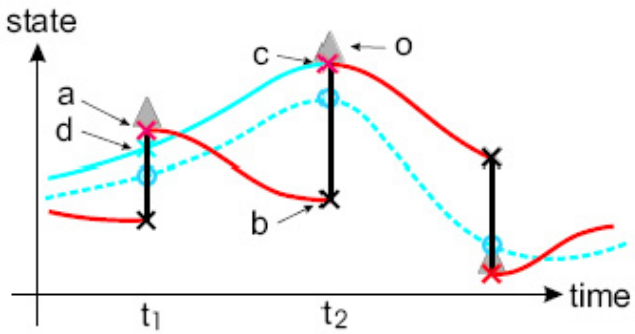


Figure: In a KF-like structure, the solution "jumps" at the analysis times. Introduces artificial sources and sinks. (From I. Fukumori)

The update step at the "analysis time", in going from $\tilde{\mathbf{x}}(t,-)$ to $\tilde{\mathbf{x}}(t)$, even with a rigorously correct $\mathbf{K}(t)$, forces the model state to "jump" from the prediction to the corrected values—thus without a properly determined control vector, unphysical sources and sinks of momentum, energy, freshwater, vorticity, etc. are being introduced. One thus cannot do physically meaningful budgets of these quantities—the essence of climate change.

The best estimate of the state, $\mathbf{x}(t)$ and of the unknown $\mathbf{u}(t)$ comes from solving the whole equation set as simultaneous equations. If that approach is impractical, how to do that? A *smoother* refers to estimation of the state in the past. An important alternative *algorithm* is called the *RTS smoother*. (RTS is for Rauch, Tung, Striebel who first worked it out.) It consists of running the KF all the way to the end of the data set. There is then no future data to modify that last point. Thus $\tilde{\mathbf{x}}(t_f, +) = \tilde{\mathbf{x}}(t_f)$, with uncertainty, $\mathbf{P}(t_f)$, the $+$ being used to show that future data have been used. One then goes one step backwards in time, knowing that the discrepancy between the value predicted at t_f, which was $\tilde{\mathbf{x}}(t_f, -)$ and the value finally estimated, $\tilde{\mathbf{x}}(t_f)$, had to be due to (A) the missing $\mathbf{u}(t_f)$, and/or the error in $\tilde{\mathbf{x}}(t_f - 1)$ .The RTS smoothing algorithm partitions the corrections between them in a second recursion,

RTS algorithm (see DISEP),

$$
\begin{aligned}
\tilde{\mathbf{x}}(t,+) &= \tilde{\mathbf{x}}(t) + \mathbf{L}(t+\Delta t)\left[\tilde{\mathbf{x}}(t+\Delta t,+) - \tilde{\mathbf{x}}(t+\Delta t,-)\right] \\
\mathbf{L}(t+\Delta t) &= \mathbf{P}(t)\mathbf{A}(t)^T \mathbf{P}(t+\Delta t,-)^{-1} \\
\mathbf{u}(t,+) &= \mathbf{M}(t+\Delta t)\left\{\mathbf{x}(t+\Delta t,+) - \mathbf{x}(t+\Delta t,-)\right\} \\
\mathbf{M}(t+\Delta t) &= \mathbf{Q}(t)\mathbf{\Gamma}(t)^T \mathbf{P}(t+\Delta t,-)^{-1} \\
\mathbf{P}(t,+) &= \mathbf{P}(t) + \mathbf{L}(t+\Delta t)\left[\mathbf{P}(t+\Delta t,+) - \mathbf{P}(t+\Delta t,-)\right]\mathbf{L}(t+\Delta t)^T \\
\mathbf{Q}(t,+) &= \mathbf{Q}(t) + \mathbf{M}(t+\Delta t)\left[\mathbf{P}(t+\Delta t,+) - \mathbf{P}(t+\Delta t,-)\right]\mathbf{M}(t+\Delta t)^T
\end{aligned}
$$

The formulas are a bit complex appearing because successive estimates with the KF have temporally correlated errors. The smoothing rotates and stretches the KF solutions to render them uncorrelated with equal variance, averages them appropriately, then rotates and stretches back. The estimated control $\mathbf{u}(t)$ is proportional to the estimated covariance $\mathbf{Q}(t)$ and the value of $\tilde{\mathbf{x}}(t,+)$ is a covariance-weighted average of the KF prediction and the estimated new value using the formally future data. Consider $\mathbf{Q}(t) \to 0$, etc.
Other smoothing algorithms exist, too.

The origin of the computational load in the Kalman filter and in any smoothing algorithm lies with the sequential property: the forecast state has to be weighted appropriately relative to the data-based estimate of it in both the filter and smoother sequences. (If you don't do it right, you will get a wrong answer!) Can one find another way to solve the simultaneous equations without holding them all in the computer at once? (Repeating: if one can do it all at once, that's the most efficient way.)

It leads one to think about methods that might be called "whole domain" rather than "sequential," so that the uncertainties are not required, and in the spirit of least-squares, to the notion of Lagrange multipliers (and what will be called the *adjoint* method).

First, a bit more about "data assimilation" as practiced by meteorologists. The atmosphere isn't dynamically linear over weather forecast time scales. Using a nonlinear model is not, per se, an issue. It's only when combined with data at the "analysis" time of the update by data that one has trouble: one needs the uncertainty of the forecast. A critical step in the KF derivation was the assertion that the forecast uncertainty, $\mathbf{P}(t,-)$ was the sum of the error from that in the previous state plus that of the unknown controls, $\mathbf{AP}(t-\Delta t)\mathbf{A}^T+\Gamma\mathbf{Q}\Gamma^T$. But if the model is nonlinear, among other issues, *errors are not additive.* They are also very unlikely to be Gaussian or even unimodal, which raises all kinds of other problems when averaging the forecast with a data-based estimate which might be Gaussian.

Resort is commonly had to Monte Carlo or *ensemble* methods, a kind of brute force approach: Given $\tilde{\mathbf{x}}(t-\Delta t)$, $\mathbf{P}(t-\Delta t)$, and $\mathbf{Q}(t-\Delta t)$, generate an ensemble of forecasts by running the model forward on a whole series of initial and boundary conditions disturbed randomly, $\tilde{\mathbf{x}}(t-\Delta t)+\Delta\tilde{\mathbf{x}}(t-\Delta t)$, $\mathbf{Bq}(t-\Delta t)+\Delta\mathbf{q}(t-\Delta t)$, where the Δ fields are in principle consistent with $\mathbf{P}(t-\Delta t)$, and $\mathbf{Q}(t-\Delta t)$. One then has an ensemble of forecasts,$\tilde{\mathbf{x}}^{(i)}(t,-)$ which are then used to calculate an empirical covariance, $\mathbf{P}(t,-)$. Generating a useful realistic set of disturbances is not trivial, and ensemble sizes are usually orders of magnitude smaller than the dimension of $\mathbf{x}, \mathbf{P}$ so that the empirical covariance matrix is necessarily highly singular. One then assumes that the best estimate is still a *linear* combination of forecasts and estimate from the data alone. Nonethless, these methods are useful and Evensen (2007) is an entire book on the subject. Not pursued here.

The KF and smoothers can be extended in numerous ways, usefully remembering the underlying structure of trying to solve sets of simultaneous equations. The "linearized" filter is applied to the equations governing the deviation from a guessed state, $\mathbf{x}_0(t)$. The "extended" filter is applied to a linearization about the previous best estimate: $\tilde{\mathbf{x}}(t-\Delta t)$ (stability issues arise). The update of $\mathbf{P}(t)$ can be numerically inaccurate if done by successive subtractions over many time-steps. Can rewrite as a propagation in $\sqrt{\mathbf{P}(t)}$ (retains the non-negative definite property; the "square-root filter"), or in terms of $\mathbf{P}(t)^{-1}$ ("information matrix"). Etc. These are numerical algorithm problems, not conceptual ones.
If one of the parameters, e.g., k in the mass-spring oscillator, is to be estimated, one can augment the state vector, $[x(t), x(t-\Delta t), k(t-\Delta t)]^T$, perhaps with an extra equation, e.g., $k(t) = k(t-\Delta t)$, or something more interesting if the spring constant were time-dependent. The problem is now nonlinear, and one might start by writing $k = k_0 + \Delta k(t)$, thus linearizing, etc., etc.
All of these techniques can be applied to smoothing algorithms.

Let's go back and look briefly at the so-called reanalyses. The "analysis" is the original weather forecast done by some workable approximation to a best estimate (bridges are not normally built optimally—they are still useful). The models (operator $\mathbf{L}$) have changed greatly over time beginning at their origin around 1955. That generated sensible worries about a changing estimated atmospheric state arising artificially from model changes alone. So the model is "frozen" in the reanalysis, and is commonly the most recent weather forecast model.

The motivation for many people here is undoubtedly what the meteorologists call "data assimilation."
Example, Kalnay et al. (1996)

The NCEP/NCAR 40-Year Reanalysis Project

E. Kalnay,* M. Kanamitsu,* R. Kistler,* W. Collins,* D. Deaven,* L. Gandin,*
M. Iredell,* S. Saha,* G. White,* J. Woollen,* Y. Zhu,* M. Chelliah,* W. Ebisuzaki,*
W. Higgins, J. Janowiak,* K. C. Mo, C. Ropelewski,* J. Wang,*
A. Leetmaa,* R. Reynolds,* Roy Jenne,* and Dennis Joseph*

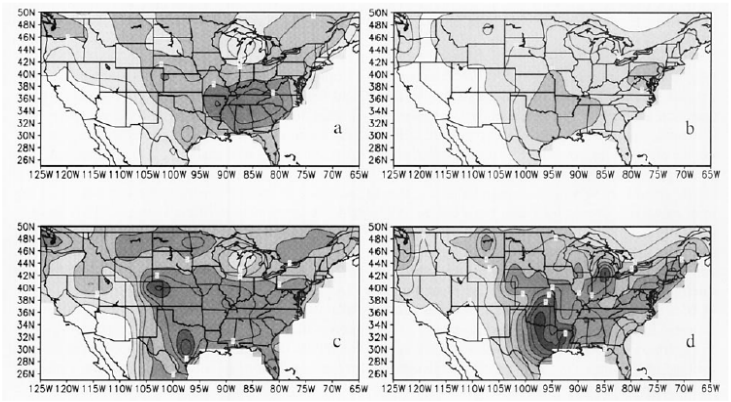
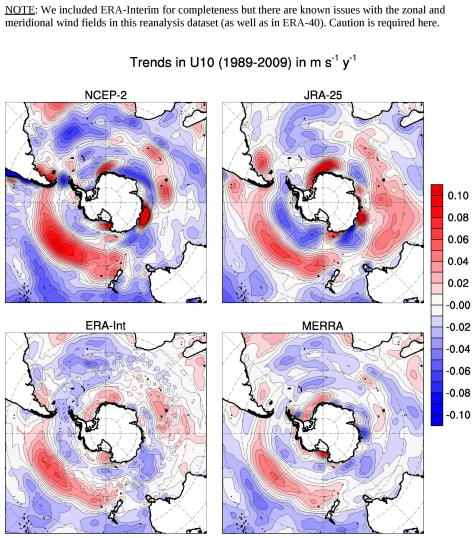


FIG. 8. Daily mean precipitation rates (mm day^{-1}) for May 1985–89 over the United States in (a) the NCEP/NCAR reanalysis and in (b) the observations. Standard deviation of the daily mean precipitation rates (mm day^{-1}) within May 1985–89 in (c) the NCEP/NCAR reanalysis and (d) the observations. Contour interval is 1 mm day^{-1}, greater than 1 mm day^{-1} is shaded.

Cited 9500+ times. What is this? How is it done? Citation index gives about 9000 papers on "data assimilation" (June 2012).



Trends in southern hemisphere winds. Same data, very similar models, similar methods. Why are they so different? Do you believe one of them? Why?

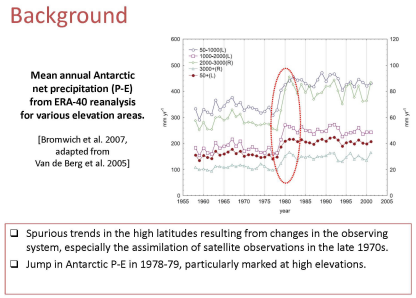
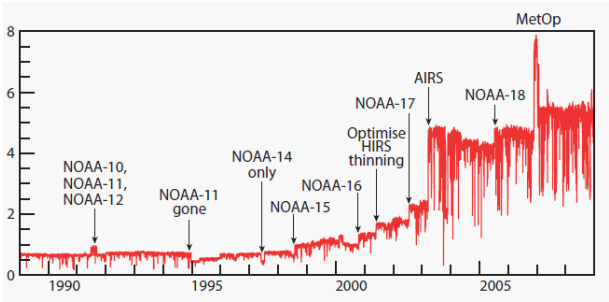
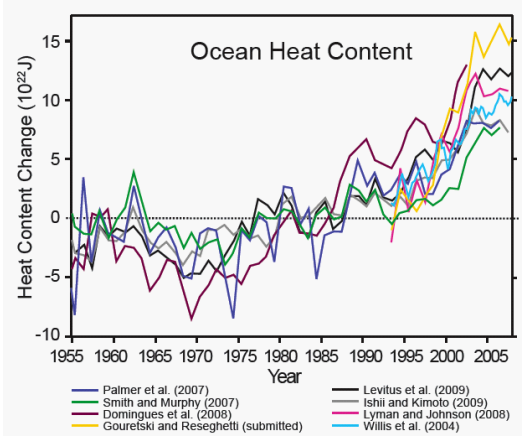


Figure: From D. Bromwich. The "jump" with the new data type shows that (A) either something is very wrong with the procedure and/or, (B) the uncertainty of the estimates exeeds the magnitude of the jump.



The error analyses are usually inadequate and unconvincing: The data base changes hugely over that same time period, but the equivalent of the $\mathbf{K}(t)$ matrix is (A) guessed at and (B) held constant in time. Because of the use of KF-like structures, the earlier estimates do not "know" about information in the later data. Why use a prediction method on what is obviously a smoothing (interpolation) problem? Would at least be useful were there error estimates.

A *major issue* is the many orders of magnitude change in the observation system since 1950 or 1870 or whenever the reanalysis begins. The sensitive dependence of the time evolution of $\mathbf{K}(t)$ (and through it of $\mathbf{P}(t,-1)$, $\mathbf{P}(t)$) on $\mathbf{E}(t)$ shows that there will be qualitative changes in the correct weighting of model and data over the decades. These changes are not normally accounted for—$\mathbf{K}(t)$ being held fixed.



(Courtesy P. Heimbach). Why does the model spread become greater toward the end when there are many more data? Answer is probably that with more data, more options are available for interpolation, smoothing, weighting etc.

Bengtsson, L., Hagemann, S., Hodges, K. I. Can climate trends be calculated from reanalysis data? JGR, 2004

The answer they give is "no" (widely ignored). Trends seen are dominated by changes in the observing system. But note the confusion in their title of the "reanalysis" with "data". Model outputs are *never* data.

Optimal sequential methods are beyond computational reach. What to do? Consider a toy least-squares problem:
minimize $J = x_1^2 + x_2^2$. (The solution is $x_1 = x_2 = 0$ and $J = 0$). But suppose one wanted the solution to also satisfy $x_1 - x_2 = 7$? One approach is to eliminate: $x_1 = 7 + x_2$, $J = (x_2 + 7)^2 + x_2^2$ and minimize with respect to x_2 (left to the reader; $x_2 = 7/2$, $x_1 = 7 + 7/2$). This is a *constrained optimization problem.* An alternative (apparently due to Lagrange, circa 1800): Introduce a new variable, μ. Modify J to

$$J' = J - 2\mu\,(x_1 - x_2 - 7)$$

Treat it as *unconstrained*, but with μ also to be determined:

$$\begin{aligned}
\frac{1}{2}\frac{\partial J'}{\partial x_1} &= x_1 - \mu = 0 \\
\frac{1}{2}\frac{\partial J'}{\partial x_2} &= x_2 + \mu = 0 \\
\frac{1}{2}\frac{\partial J'}{\partial \mu} &= -(x_1 - x_2 - 7) = 0
\end{aligned}$$

Three equations in three unknowns, which produces the same solution for x_1, x_2 as well as $\mu = 7/2$. The "trick" is explained in all textbooks on the calculus of variations, classical mechanics, and least-squares, including DISEP. It converts a constrained problem into an unconstrained one—which can be much simpler if it isn't so easy to eliminate enough variables. Price paid is the enlargement of the number of unknowns.

What problem are we trying to solve?
Minimize:

$$J = \sum_{t=1}^{t_f} \left(\mathbf{y}(t) - \mathbf{E}(t)\mathbf{x}(t)\right)^T \mathbf{R}_{nn}(t)^{-1} \left(\mathbf{y}(t) - \mathbf{E}(t)\mathbf{x}(t)\right)$$
$$+ \sum_{t=1}^{t_f} \mathbf{u}(t)^T \mathbf{Q}^{-1}\mathbf{u}(t) + \left[\tilde{\mathbf{x}}(0,+) - \tilde{\mathbf{x}}(0)\right]^T \mathbf{P}(0)^{-1} \left[\tilde{\mathbf{x}}(0,+) - \tilde{\mathbf{x}}(0)\right]$$

subject to the model:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t - \Delta t) + \mathbf{B}\mathbf{q}(t - \Delta t) + \Gamma\mathbf{u}(t - \Delta t)$$

So let's modfiy J using a Lagrange mutliplier for the model at each time step, $\boldsymbol{\mu}(t)$

Rewriting the above J,

$$\begin{aligned} J &= [\tilde{\mathbf{x}}(0,+) - \tilde{\mathbf{x}}(0)]^T \mathbf{P}(0)^{-1} [\tilde{\mathbf{x}}(0,+) - \tilde{\mathbf{x}}(0)] \\ &+ \sum_{t=1}^{t_f} [\mathbf{y}(t) - \mathbf{E}(t)\tilde{\mathbf{x}}(t,+)]^T \mathbf{R}(t)^{-1} [\mathbf{y}(t) - \mathbf{E}(t)\tilde{\mathbf{x}}(t,+)] \\ &+ \sum_{t=0}^{t_f-1} \tilde{\mathbf{u}}(t,+)^T \mathbf{Q}(t)^{-1} \tilde{\mathbf{u}}(t,+) \\ &-2 \sum_{t=1}^{t_f} \boldsymbol{\mu}(t)^T [\tilde{\mathbf{x}}(t,+) - \mathbf{A}\tilde{\mathbf{x}}(t-\Delta t,+) - \mathbf{B}\mathbf{q}(t-\Delta t,+) - \Gamma\tilde{\mathbf{u}}(t-\Delta t,+)] \end{aligned}$$

(the $+$ has been put into the arguments to emphasize that the solution uses all data, past and future to t). $\boldsymbol{\mu}(t)$ is now a vector. Take the partial derivatives and set them to zero:

$$\frac{1}{2}\frac{\partial J}{\partial \tilde{\mathbf{u}}(t,+)} = \mathbf{Q}(t)^{-1}\tilde{\mathbf{u}}(t,+) + \Gamma^T \boldsymbol{\mu}(t+\Delta t) = 0\,, \quad t = 0, 1, ..., t_f - 1 \tag{1}$$

$$\frac{1}{2}\frac{\partial J}{\partial \boldsymbol{\mu}(t)} = \tilde{\mathbf{x}}(t,+) - \mathbf{A}\tilde{\mathbf{x}}(t-\Delta t,+) - \mathbf{B}\mathbf{q}(t-\Delta t) - \Gamma\tilde{\mathbf{u}}(t-\Delta t,+) = 0,$$

$$t = 0, 1, ..., t_f \tag{2}$$

$$\frac{1}{2}\frac{\partial J}{\partial \tilde{\mathbf{x}}(0,+)} = \mathbf{P}(0)^{-1}\left[\tilde{\mathbf{x}}(0,+) - \tilde{\mathbf{x}}(0)\right] + \mathbf{A}^T\boldsymbol{\mu}(\Delta t) = 0\,, \tag{3}$$

$$\frac{1}{2}\frac{\partial J}{\partial \tilde{\mathbf{x}}(t,+)} = -\mathbf{E}(t)^T\mathbf{R}(t)^{-1}\left[\mathbf{y}(t) - \mathbf{E}(t)\tilde{\mathbf{x}}(t,+)\right] - \boldsymbol{\mu}(t) + \mathbf{A}^T\boldsymbol{\mu}(t+\Delta t) = 0,$$

$$t = 1, 2, ..., t_f - 1 \tag{4}$$

$$\frac{1}{2}\frac{\partial J}{\partial \tilde{\mathbf{x}}(t_f)} = -\mathbf{E}(t_f)^T\mathbf{R}(t_f)^{-1}\left[\mathbf{y}(t_f) - \mathbf{E}(t_f)\tilde{\mathbf{x}}(t_f)\right] - \boldsymbol{\mu}(t_f) = 0 \tag{5}$$

another set of $N \times N$ simultaneous equations.

The so-called *adjoint model* is the 4th equation,

$$\boldsymbol{\mu}(t) = \mathbf{A}^T \boldsymbol{\mu}(t+\Delta t) + \mathbf{E}(t)^T \mathbf{R}(t)^{-1} \left[\mathbf{E}(t)\,\tilde{\mathbf{x}}(t,+) - \mathbf{y}(t)\right],$$

with time (apparently, but not actually) running backwards and with the estimate-data misfit as a forcing term (the problem has a fixed time span—so time runs in both directions).
The "adjoint method" solves these equations (usually with a nonlinear model present). For a linear model, this system has an analytic solution (see DISEP). Computationally it is far smaller than the KF-RTS system or equivalent, because it does not require computation of the $\mathbf{P}$. On the other hand, one does not have the $\mathbf{P}$!

Meteorologists call their algorithms for solving this system iteratively, approximately, "4DVAR" (although it's neither (necessarily) four-dimensional nor is it variational); electrical engineers call it the *Pontryagin (minimum) Principle*. I call it the *method of Lagrange multipliers*, because that description is more meaningful to a wider community.

Nobody in his right mind writes a code for a real GCM with the **A** matrix being formed explicitly. Rather one time-steps each element according to some rule. **A** is implicit. So how does one get it? There are several ways, discussed in DISEP, but not pursued here.
The time-stepping is done in practice by a large Fortran code (circa 100,000 lines) and J is similarly another Fortran code.
The simultaneous equations are solved by *numerical search* for a stationary value of J, most commonly by a "downhill" search algorithm (quasi-Newton method). In particular, the $\boldsymbol{\mu}(t)$ are the partial derivatives of J with respect to the state variables. Set the downhill direction.
How do you take the partial derivatives of a Fortran code? The answer is *algorithmic differentiation* or AD. One feeds the Fortran code into the AD tool, and it produces a second Fortran code which represents the partial derivatives (work originally of R. Giering).

Assertions are commonly made that this method works only for "perfect models." But no limits exist on the magnitudes, nor the representation of $\mathbf{u}(t)$, and in principle the original model can be greatly modified. Note that after a solution is found, the *free-running,* adjusted model is run forward in time from the modified

initial conditions so that the solution one would analyze for its physics, biology, chemistry, etc. satisfies these new, exactly known governing equations.

This adjoint method is used in the ECCO project at MIT, in which there are about 2 billion data constraints, a one hour time-step over 20 years, a state vector of approximate dimension, $0.7(180)(360)(21)(6) \approx 6 \times 10^6$ at each hourly time step.

There's a lot more to it than that, but it really does work. (A lot of further information, particularly about the Lagrange multiplier/adjoint method, can be found on the ECCO webpage, (http://ecco.mit.edu/), that of OpenAD (http://www.mcs.anl.gov/OpenAD/), and Patrick Heimbach's MIT webpage.)

Some parting comments.
People make a career out of these methods and their applications, and we've just seen the tip of a large iceberg. There are many good textbooks, including those in control theory, in electrical engineering, pure mathematics, etc. They are very powerful methods, but like all powerful tools, can be dangerous to the user!
We know how to do a dynamically and statistically consistent, useful climate state estimate. Because of its enormous dimensions, it is a very challenging numerical engineering problem—one that must be solved, but that will not be done overnight.
Thanks for listening to the end.

Suggested Self-Teaching Exercises: Harvard University Shortcourse
Inverse Problems, Inverse Methods, State Estimation, Data Assimilation, and All That
Carl Wunsch January 2013

If you can work these through, you have a reasonable understanding of the short-course material. If you get stuck, come see me in Room 451, Geological Museum or MIT Rm 54-1426.

1. Discretize the Laplace/Poisson equation,

$$\nabla^2 \phi = \rho \tag{1}$$

in two-dimensions on a rectangular grid, i, j. Make the grid big enough that there is an interior separate from the boundaries. A minimum number of grid points of 6 in each direction is probably about right to keep the output manageable, but much larger is ok too. Let the coordinates be x, y. Let ρ be zero except on one interior grid point (say, 4,4). Let the boundaries be $x = 0, L, y = 0, L$ so that the domain is square (you can change that). Put $\phi(x = 0, y) = \phi(x = L, y) = \phi(x, y = L) = 0$, $\phi(x, y = 0) = 1$.

(a) Solve for ϕ_{ij} by matrix inversion. (Check that it works.)

(b) Solve for ϕ_{ij} using the SVD, increasing the rank from 1 to the maximum possible and calculating the residual at each rank choice.

(c) Using ϕ_{ij} from (a) calculate the boundary conditions and values of ρ that must have been imposed (you know them, but pretend you don't).

(d) Keeping the system of equations used in (a,b), add "observations" that $\phi_{3,4} = 2, \phi_{4,3} = 7$ (but use your own grid values—any interior points). Resolve the system and find the new values of ϕ_{ij} everywhere?. How do they differ? Use both some form of least-squares and the SVD. Then let there be some uncertainty: $\phi_{3,4} = 2 \pm 1, \phi_{4,3} = 7 \pm 10$ and make a new estimate of ϕ_{ij}.

(e) Suppose the boundary values $\phi(x, y = 0) = 1$ are not known, but you know those on the other three boundaries. And you know $\phi_{3,4} = 2 \pm 1, \phi_{4,3} = 7 \pm 10$. What can you say about the missing boundary conditions?

2. Now consider a time-dependent tracer equation,

$$\frac{\partial C}{\partial t} - k\nabla^2 C = 0 \tag{2}$$

discretized spatially as in Problem 1, and also in time over an interval Δt (your choice). Initial conditions are zero. Boundary conditions are also zero, except $\mathrm{C}(x, y = L) = 1$, $t > 0$. Write it as a state-space form,

$$\mathbf{x}\left(t + \Delta t\right) = \mathbf{A}\mathbf{x}\left(t\right) \tag{3}$$

(a) Write it out as a large set of simultaneous equations over some finite time $t = 0...M\Delta t$ and solve by matrix inversion. It should be an $N \times N$ system if done right. (Notice that the coefficient matrix is sparse—which you can exploit if you wish.)

(b) Take the solution at $t = M\Delta t$, add a bit of corrupting white noise to it, and solve for the initial conditions, pretending you don't know them, using least-squares.

(c) Using those corrupted data, use a Kalman filter to predict $\mathbf{x}\left(t\right), t = \left(M + 1\right)\Delta t, \left(M + 2\right)\Delta t, ..$ and calculate their uncertainty.

(d) From the corrupted "data" of (2) find the initial conditions using (i) any smoother of your choice; (ii) Lagrange multipliers.