A Brief
(hopefully entertaining)
Historical Introduction to
Estimation Theory with Eyes on
Weather Prediction

Ricardo Todling
NASA Global Modeling and Assimilation Office

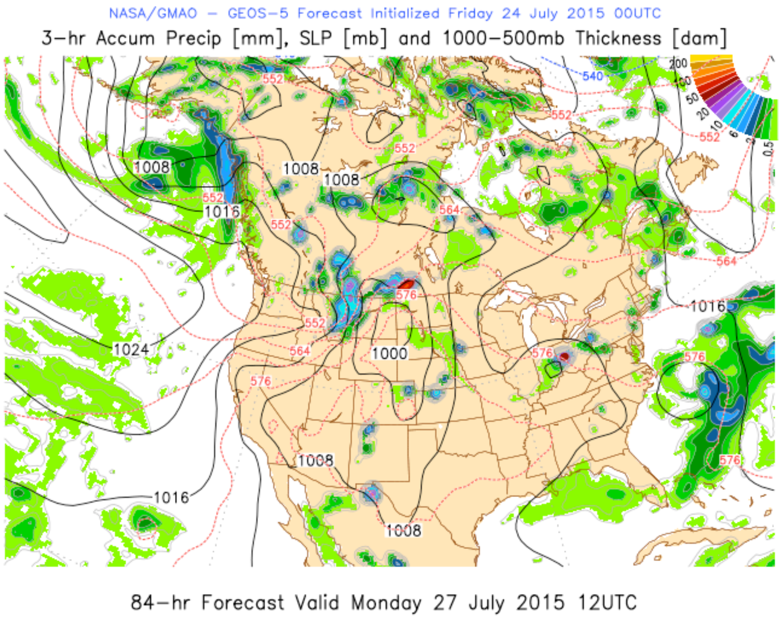
The JCSDA Summer Colloquium on Satellite Data Assimilation
Ft. Collins, CO, 27 July to 7 August 2015

First Presented to the Participants of the 2013
UMD Summer School on Data Assimilation during their visit to NASA

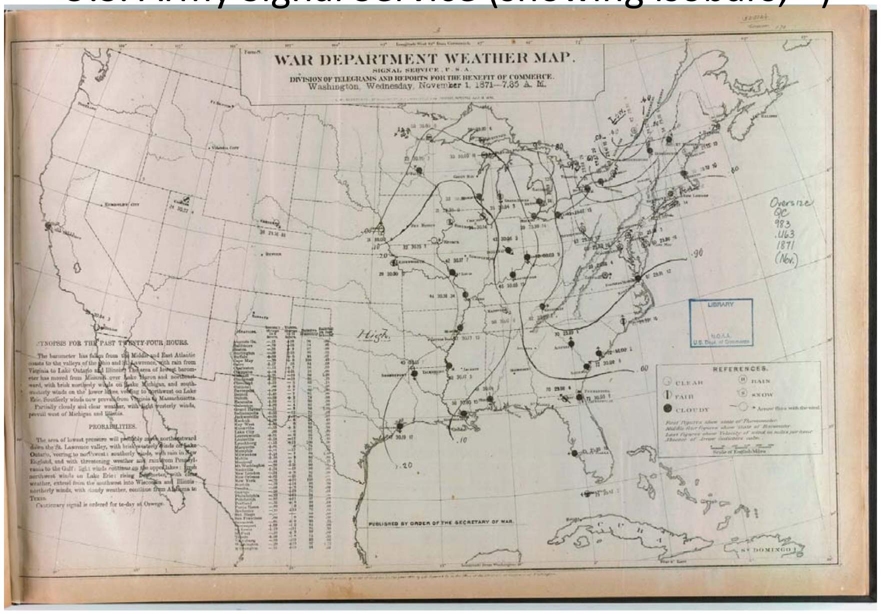
Warning: This is by no means an exhaustive introduction to the subject.

Estimating the Weather

84-hr Fcst for Today's 12 UTC


NASA/GMAO – GEOS-5 Forecast Initialized Friday 24 July 2015 00UTC
3–hr Accum Precip [mm], SLP [mb] and 1000–500mb Thickness [dam]
1008
1016
1024
1000
552
564
576
540
84–hr Forecast Valid Monday 27 July 2015 12UTC

1 Nov 1871: First weather map, issued by U.S. Army Signal Service (showing isobars; *)


WAR DEPARTMENT WEATHER MAP.
SIGNAL SERVICE, U.S.A.
DIVISION OF TELEGRAMS AND REPORTS FOR THE BENEFIT OF COMMERCE.
Washington, Wednesday, November 1, 1871—7.35 A. M.
SYNOPSIS FOR THE PAST TWENTY-FOUR HOURS.
PROBABILITIES.
REFERENCES.
High
LIBRARY

1686 Edmund Halley first map of the trade winds (*): connecting general circulation with solar heating distribution

AFRICA
AMERICA

(*) source: *www.shorstmeyer.com/msj/geo165/met_hist.pdf* - Steve Horstmeyer

When did it all begin?

Steven Horstmeyer's: *"An Outline of the History of Meteorology"* is a wonderful presentation you should consult.

The presentation is a way more modest illustrative short history of estimation for NWP.

“We know today, mainly due to the work of J. Charney, that we can predict by calculation the weather over an area like that of the United States for a duration like 24 hours [. . .]. We know that this gives results which are, by and large, as good as what an experienced ‘subjective’ forecaster can achieve, and this is very respectable.”

John von Neumann, 1954.

Data Assimilation or … ?

- Inverse Problems
- Stochastic Estimation
- Distributed Parameter Estimation
- Lumped Parameter Estimation
- Optimal Filtering and Smoothing
- Bayesian Estimation
- Least Squares Estimation
- Absolute Averaging

When did ideas on estimation emerge?

Minimization
Uncertainty
Probability

My main sources: Hacking, Franklin, Lanczos, McGee & Schmidt and sprinkles from many others.

Outline

1. Least; Extremum; Mimimum
2. Uncertainty
3. Probability
4. Two Real-Like Applications
    - The Apollo Missions
    - Predicting the Weather
5. Closing Remarks

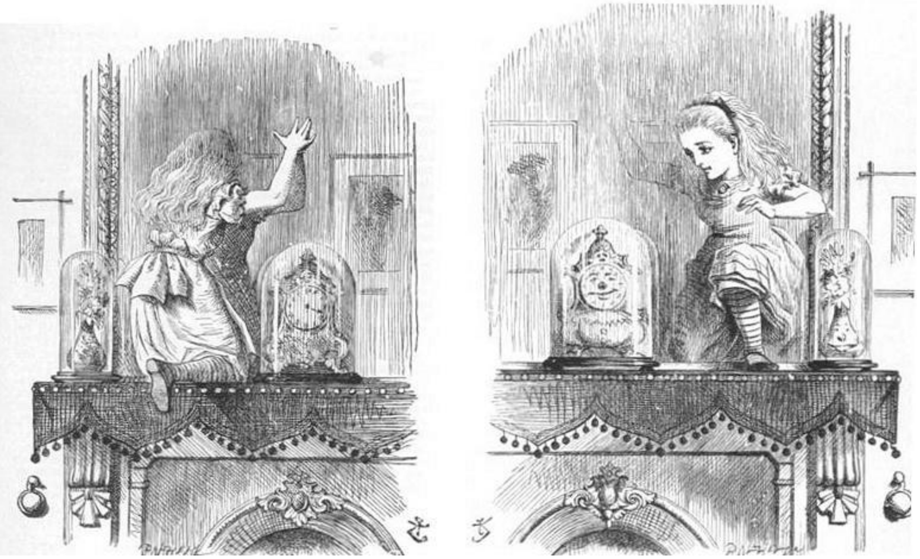
Least

Extremum

Minimum

The Shortest Path

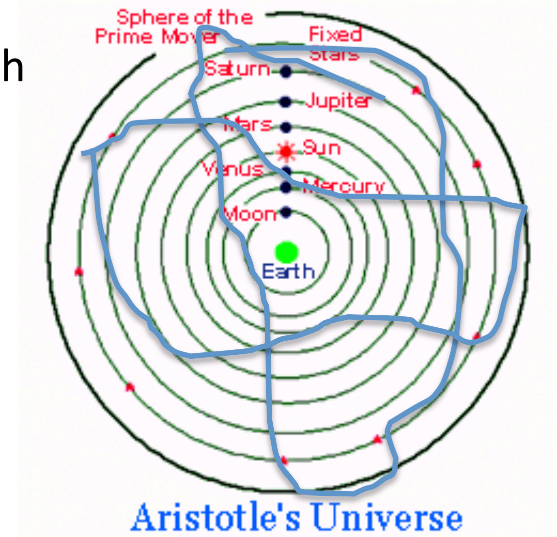
Hero of Alexandria
B. c 10 AD



Hero showed that the path taken by a light ray going from an object to a mirror and from the mirror to an observer, is the *shortest* of any *path* going from the object to the eye of the observer via the mirror. He derived the *law of reflection*.

Aristotle
384-322 BC

Hero's thinking was consistent with that of ***Aristotle***, who thought that planets moved in circles because they were the shortest closed path an object could trace when going around another.



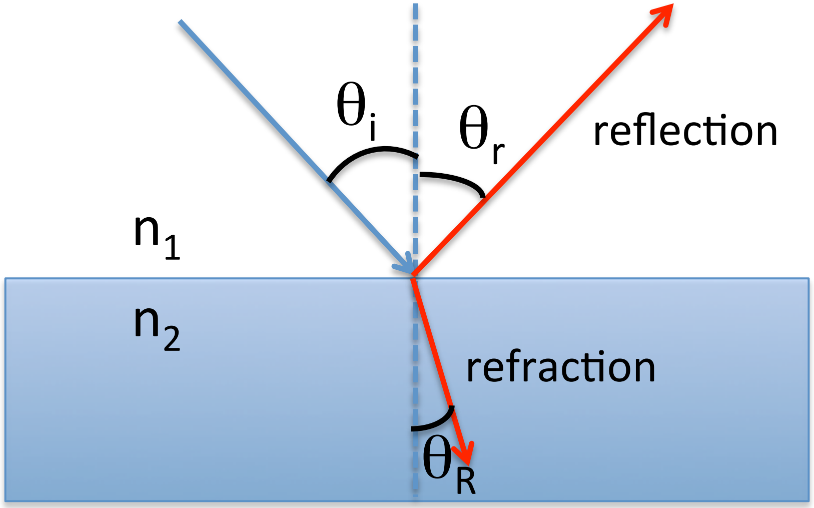
Modified from starchild.gsfc.nasa.gov

Combined with the *maximum speed of motion*, Hero's thinking leads to the concept of *the shortest time traveled*.

The Principle of Least Time

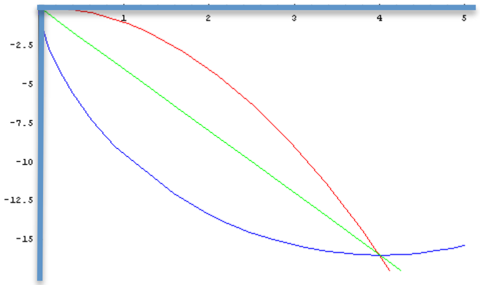
Pierre de Fermat
e1600-1655

$$\frac{\sin\theta_i}{\sin\theta_r} = \frac{n_r}{n_i} = \frac{v_i}{v_r}$$

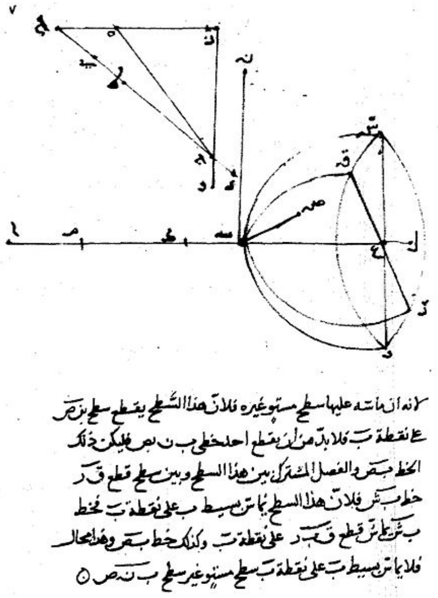


Fermat derived *the law of refraction* by using Hero’s principle of shortest time traveled.

A similar problem of interest was that of the *brachistochrone* – the curve of quickest descent – proposed by ***Johann Bernoulli***, and solved by Newton, Jakob Bernoulli (brother), Leibniz, Tschirnhaus, and l’Hopital. Jakob B.’s solution was based on Fermat’s least time traveled.



Simulation from
http://curvebank.calstatela.edu/brach/brach.htm



لانه ان ماسته عليها سطح مستو غيره فلان هذا السطح يقطع سطح ب ن ص
على نقطة ب فلا بد من ان يقطع احد خطي ب ن ب ص فليكن ذلك
الخط ب ص والفصل المشترك بين هذا السطح وبين سطح قطع ق ر
خط ب ش فلان هذا السطح يماس بسيط ب على نقطة ب فخط
ب ش يماس قطع ق ب ر على نقطة ب وكذلك خط ب ص وهذا محال
فلا يماس بسيط ب على نقطة ب سطح مستو غير سطح ب ن ص ○

Ibn Sahl manuscript of 984,
describing the law of refraction
six centuries before **Snell-Descartes**

The Principle of Least Action

Based on Lanczos (1949)

Pierre-Louis de Maupertuis
1698-1759

Leibniz argued that the principles of nature could be expressed in the terms of minimum principles. This went along with his vision that we live in the 'best of all possible worlds'.

But it was ***Maupertuis*** who explained the impact of bodies by assuming the product *mvs* to be a minimum following ***D'Alembert***'s principle. The quantities *mvs* was named *action*. He showed how ***Fermat's*** *principle of least time* can be replaced by the *principle of least action*.

Euler generalized Maupertuis principle into an integral theorem applicable to motion of particles subjected to a conservative force. The action principle was recognized to be a principle of *extremum*.

Lagrange extended Euler's principle introducing the feature of invariance with respect to arbitrary change of coordinates, and developed along the way the *calculus of variations*. He set the foundations of analytic mechanics.

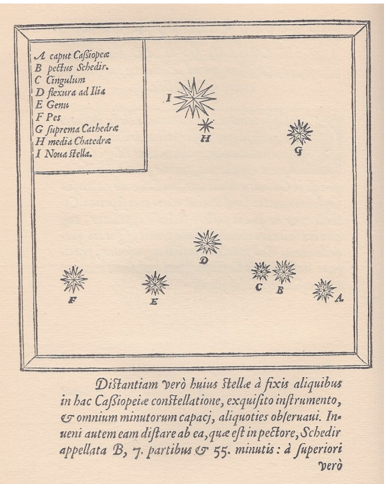
Hamilton transformed the second order differential equations of Lagrange into a more desirable set of first order differential equations with double the number of variables – called "*canonical form*" – prompting a new world of discoveries.

Uncertainty

Accounting for Uncertainty

Based on
Franklin (2001)

Another contribution from Galileo



A caput Caſsiopeæ
B pectus Schedir.
C Cingulum
D flexura ad Ilia
E Genu
F Pes
G ſuprema Cathedræ
H media Chatedræ
I Noua ſtella.

Diſtantiam verò huius ſtellæ à fixis aliquibus in hac Caſsiopeiæ conſtellatione, exquiſito inſtrumento, & omnium minutorum capacj, aliquoties obſeruaui. Inueni autem eam diſtare ab ea, quæ eſt in pectore, Schedir appellata B, 7. partibus & 55. minutis: à ſuperiori

verò

Constellation of Cassiopeia showing Tycho Brahe's nova of 1572.

Tycho Brahe refuted the *Aristolelian belief* in the unchanged sphere of the fixed start (beyond the Moon), but controversy remained.

In 1621 ***Scipione Chiaramonti*** published results from a comparative study examining observations of star elevation made by 13 astronomers. He looked at 12 pairs of observations and concluded the estimated distances from each measurement to be less than the distance of the moon. Being an Aristotelian, he wanted to show the heavens to be unchanging.

Galileo points out that of the 65 possible pairs Chiaramonti chose only those supporting his belief. Galileo re-evaluates the matter by realizing that observations are:
(i) "equally prone to err in one direction and the other"; and that
(ii) carefully taken measurements are "more likely to err little than much"

Galileo's solution is to choose the position that makes the *sum of the corrections least.*

Galileo was then able to show that indeed, Tycho Brahe was right in saying the nova had appeared in the unchanging sphere of the stars!

The Principle of Least Constraint & The Least Squares Method

Based on Lanczos (1949)

Gauss: from extremum to minimum

Johann Carl Friedrich Gauss
1777-1855

Up to about the time of Gauss all principles of action led to an extremum solution, not necessarily a minimum. Starting from ***D'Alembert***'s *principle of equilibrium* of forces acting on a system

$$\sum_{i=1}^{N} (\mathbf{f}_i - m_i \mathbf{a}_i)^T \delta \mathbf{r}_i = 0$$

Gauss showed it to be equivalent to the *principle of least constraint*

$$\delta \sum_{i=1}^{N} \frac{1}{2m_i} (\mathbf{f}_i - m_i \mathbf{a}_i)^2 = 0$$

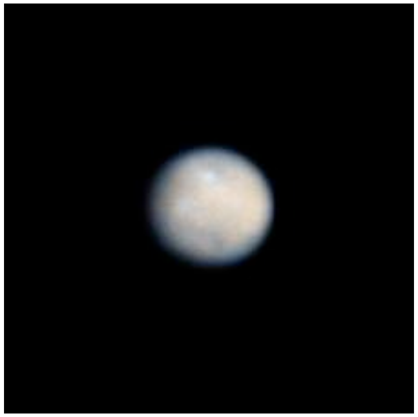
which has the advantage of its *stationary solution* being automatically a *minimum* – essentially because $m_i > 0$.

Though this does provide a more complicated solution to the problem requiring evaluation of the accelerations Gauss was particularly married to this principle since it directly related to his formulation of the **least squares method**. Here, the external forces could be thought as observations, the force of inertia as the true forces, and the mass could be interpreted as weights given accounting for different quality of the measurements.

The Least Squares Method

Based on
Lewis et al. (2006)

Laplace and Gauss: orbit of celestial objects from observations

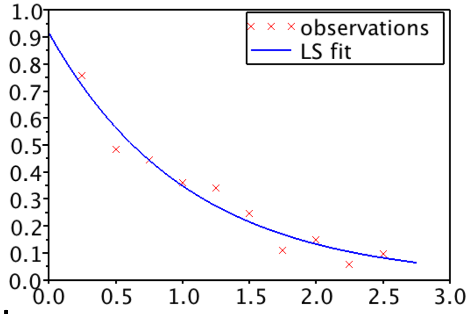


Ceres from Hubble. Today known as a planetoid in the asteroid belt.
Photo from jpl.nasa.gov

The desire to determine present and future position of celestial bodies has been with us since we first wondered about the heavens. The Babylonians and Greeks kept extensive observations of the skies. Galileo, Kepler, and Newton made known breakthroughs in using observations and to explain the skies. Laplace, Lagrange and others provide us with profound insights in methods to determine the path of comets from observations. ***Laplace***, in particular, introduced concepts fundamental to our story:

(a) the algebraic sum of residuals should vanish, and
(b) the sum of the absolute values of the residuals should be a minimum [*recall our story on Galileo a few slides back]*.

By the late 1700s, early 1800s, the race was on to predict the reappearance of *Ceres,* a planet discovered between Mars and Jupiter. On November 1801, ***Gauss*** predicted the planet's future path. His results were confirmed on January 1, 1802 by **Franz Zach** and **Heinrich Olbers** at two different observatories in Germany.

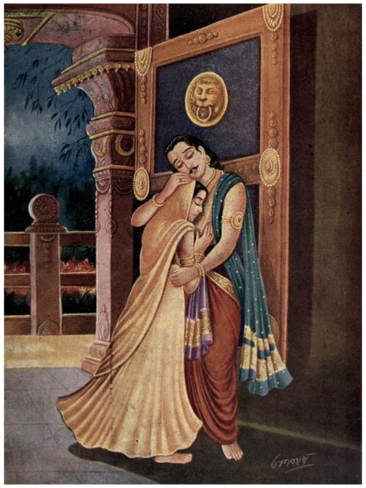


Gauss solution combines Newton's iterative method to solve nonlinear eqs, with his own development of the Least Squares Method.

Probability

The Concepts of Probability Become Mathematical

Based on
Hacking (1975)
&
Franklin (2001)



Nala meets his beloved Damayanti who's chosen him over the Gods. c A.D. 400; From http://en.wikipedia.org/wiki/Nala

From India to Pascal and Fermat

Though concepts of probability only started to mature after the mathematical forms more familiar to us, the story of ***Nala***, told in the Indian Sunskrit epic ***Mahābhārata***, who possessed by a rival demigod loses his empire to gambling. Only after coming across ***Rturpana*** and learning the science of estimation is Nala able to regain his empire and his beloved ***Damayanti*** in a game.

In the words of Ian ***Hacking*** ``*That is evidence that in India, long ago, it was recognized that there was a genuine science to master …*"

Basic concepts of averaging go as far back as the Greeks. ***Hipparchus***, about 150 BCE, was able to develop geometric models to fit the vast Babylonian observations of the stars. His eccentric circles with epicycles are made to fit the observations in a method close to what we call *regression*. But the link between averaging and probability didn't come until later.

Formal understanding of the concept of averaging (*expectation*) is relatively new, dating from the 1650's and the correspondences between ***Fermat*** and ***Pascal***. Our present-day concept of probability dates back from that period. Thoughts and needs in various areas from *law*, *gambling*, *economics*, *agriculture*, and *theology* all combined to form what we know today.

The Concepts of Probability Become Mathematical

Based on Hacking (1975)

Becoming Bayesian

In ``When Did Bayesian Inference Become "Bayesian"?'', **Stephen Fienberg** traces the roots of our present-day referencing to Thomas Bayes approach to probability problems.

Rev. Thomas Bayes
c1701-1761

What does it mean to be ``Bayesian''? It means we ``believe'' we can use the outcome of past events to infer the chances of a certain outcome in the next trail.

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)}$$

Actually due to Laplace (1774)

Laplace played a fundamental role in solidifying concepts in probability, but it wasn't until early in the 1900s that Bayes thinking gained momentum, and eventually influenced a huge body of work: **Fisher**, **Neyman**, **Pearson**, **Carnap**, **Kolmogorov**, **Turing**, **Keynes**, & others.

For us, our main interest in Bayesian probability is that it essentially provides the proper link among various formulations of the estimation problem.

L^p Norms in Estimation

We have seen that **Galileo** and **Laplace** have chosen the requirement that the absolute-value of the residual error be minimal when trying to come up with the estimates they sought.

$$min \sum_i \frac{|x_i - \mu_i|}{\sigma_i}$$

We have also seen that **Gauss** added an alternative requiring the square of the residual error to be a minimum.

We can show that **least-squares** is intimately related to **Gaussian probability distribution**.

$$min \sum_i \frac{(x_i - \mu_i)^2}{\sigma_i^2}$$

$$p \;\; \alpha \;\; \exp\left[-\sum_i \frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right]$$

When a traveler reaches a fork in the road, the L1-norm tells him to take either one way or the other, the L2-norm instructs him to head off into the bushes.

J. G. Clearbout & F. Muir, (1973); quoted in Tarantola (2005)

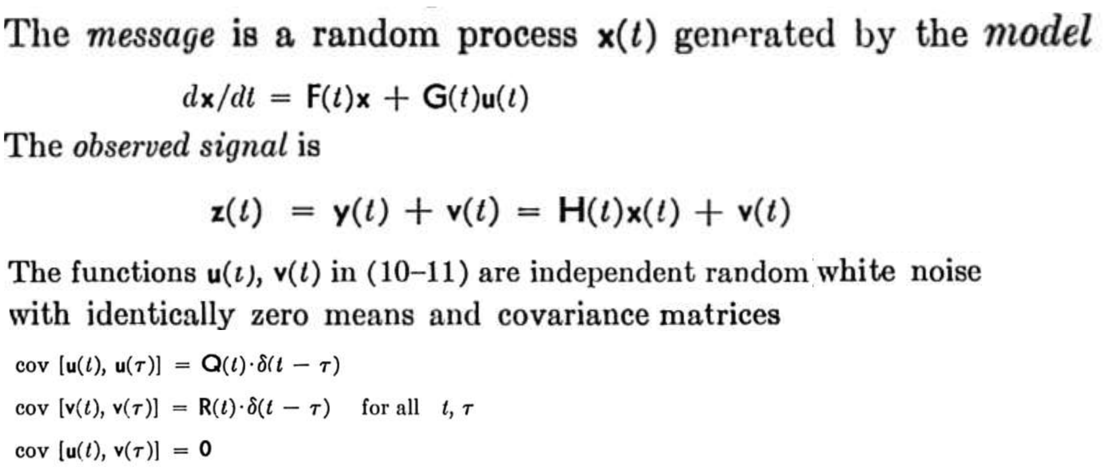
More generally, there is a class of problems based on L^p norms with associated probability distributions that are better than the familiar L^2 norm (Gaussian) to handle certain types of situations, e.g., outliers.

$$||x||_p = \left(\sum_i \frac{|x_i|^p}{\sigma_i^p}\right)^{1/p}$$

$$||x||_\infty = \max_i \left(\frac{|x_i|}{\sigma_i}\right)$$

e.g., Tarantola (2005)

The Kalman Filter: Problem



The *message* is a random process $\mathbf{x}(t)$ generated by the *model*

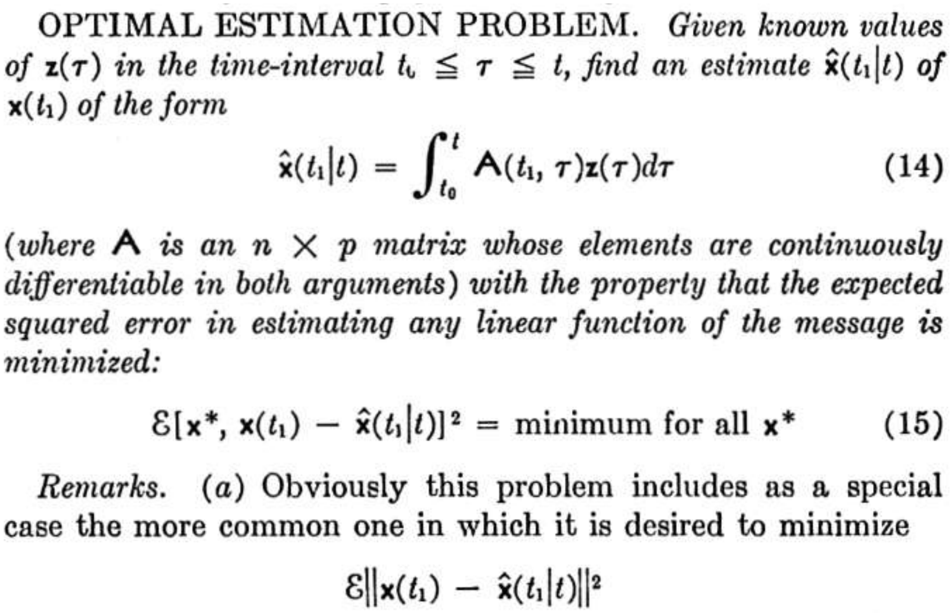
$$d\mathbf{x}/dt = \mathsf{F}(t)\mathbf{x} + \mathsf{G}(t)\mathbf{u}(t)$$

The *observed signal* is

$$\mathbf{z}(t) = \mathbf{y}(t) + \mathbf{v}(t) = \mathsf{H}(t)\mathbf{x}(t) + \mathbf{v}(t)$$

The functions $\mathbf{u}(t)$, $\mathbf{v}(t)$ in (10–11) are independent random white noise with identically zero means and covariance matrices

$$\begin{aligned} \operatorname{cov}\,[\mathbf{u}(t), \mathbf{u}(\tau)] &= \mathsf{Q}(t)\cdot\delta(t-\tau) \\ \operatorname{cov}\,[\mathbf{v}(t), \mathbf{v}(\tau)] &= \mathsf{R}(t)\cdot\delta(t-\tau) \quad \text{for all} \quad t, \tau \\ \operatorname{cov}\,[\mathbf{u}(t), \mathbf{v}(\tau)] &= 0 \end{aligned}$$



OPTIMAL ESTIMATION PROBLEM. *Given known values of $\mathbf{z}(\tau)$ in the time-interval $t_0 \leqq \tau \leqq t$, find an estimate $\hat{\mathbf{x}}(t_1|t)$ of $\mathbf{x}(t_1)$ of the form*

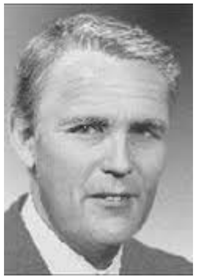
$$\hat{\mathbf{x}}(t_1|t) = \int_{t_0}^{t} \mathsf{A}(t_1, \tau)\mathbf{z}(\tau)d\tau \qquad (14)$$

(where A is an $n \times p$ matrix whose elements are continuously differentiable in both arguments) with the property that the expected squared error in estimating any linear function of the message is minimized:

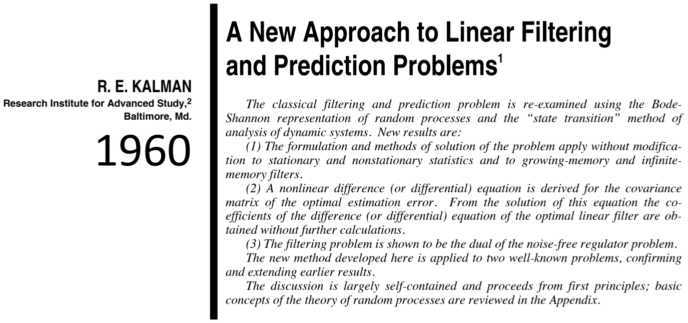
$$\mathcal{E}[\mathbf{x}^*, \mathbf{x}(t_1) - \hat{\mathbf{x}}(t_1|t)]^2 = \text{minimum for all } \mathbf{x}^* \qquad (15)$$

Remarks. (*a*) Obviously this problem includes as a special case the more common one in which it is desired to minimize

$$\mathcal{E}\|\mathbf{x}(t_1) - \hat{\mathbf{x}}(t_1|t)\|^2$$



R. E. Kalman



A New Approach to Linear Filtering and Prediction Problems[1]

R. E. KALMAN
Research Institute for Advanced Study,[2]
Baltimore, Md.

1960

The classical filtering and prediction problem is re-examined using the Bode-Shannon representation of random processes and the "state transition" method of analysis of dynamic systems. New results are:

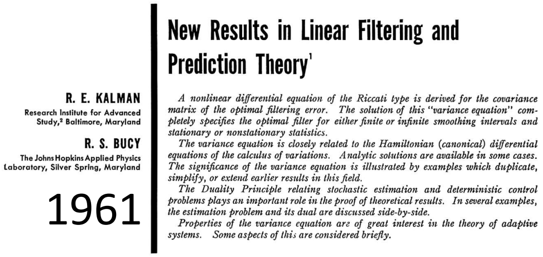
(1) The formulation and methods of solution of the problem apply without modification to stationary and nonstationary statistics and to growing-memory and infinite-memory filters.

(2) A nonlinear difference (or differential) equation is derived for the covariance matrix of the optimal estimation error. From the solution of this equation the coefficients of the difference (or differential) equation of the optimal linear filter are obtained without further calculations.

(3) The filtering problem is shown to be the dual of the noise-free regulator problem.

The new method developed here is applied to two well-known problems, confirming and extending earlier results.

The discussion is largely self-contained and proceeds from first principles; basic concepts of the theory of random processes are reviewed in the Appendix.



New Results in Linear Filtering and Prediction Theory[1]

R. E. KALMAN
Research Institute for Advanced
Study,[2] Baltimore, Maryland

R. S. BUCY
The Johns Hopkins Applied Physics
Laboratory, Silver Spring, Maryland

1961

A nonlinear differential equation of the Riccati type is derived for the covariance matrix of the optimal filtering error. The solution of this "variance equation" completely specifies the optimal filter for either finite or infinite smoothing intervals and stationary or nonstationary statistics.

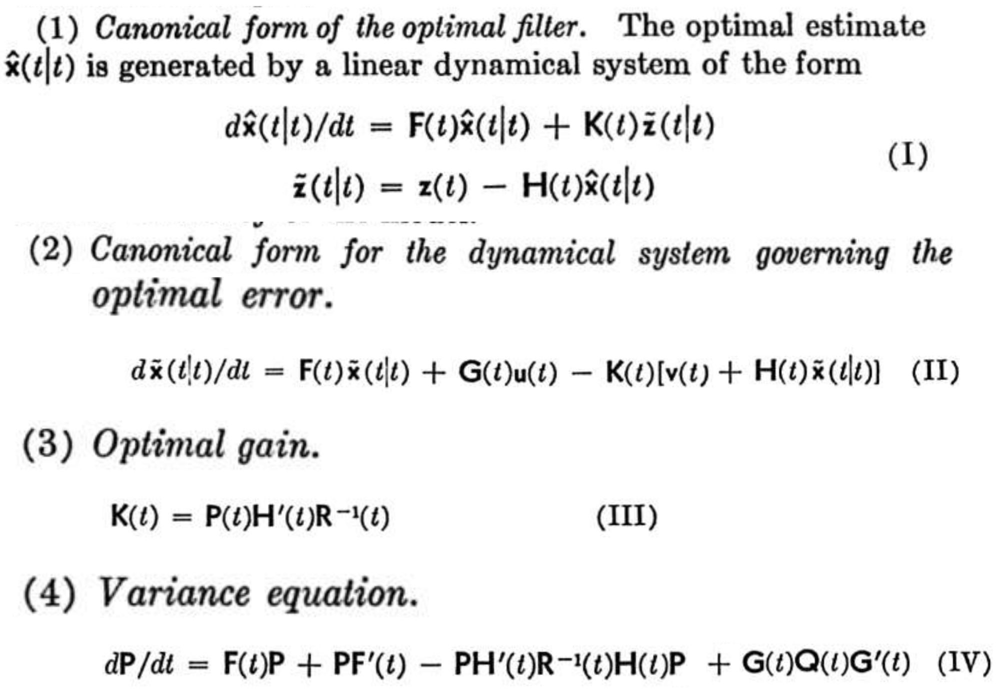
The variance equation is closely related to the Hamiltonian (canonical) differential equations of the calculus of variations. Analytic solutions are available in some cases. The significance of the variance equation is illustrated by examples which duplicate, simplify, or extend earlier results in this field.

The Duality Principle relating stochastic estimation and deterministic control problems plays an important role in the proof of theoretical results. In several examples, the estimation problem and its dual are discussed side-by-side.

Properties of the variance equation are of great interest in the theory of adaptive systems. Some aspects of this are considered briefly.

Excerpts from Kalman & Bucy (1961) original work

The Kalman Filter: Solution



(1) *Canonical form of the optimal filter.* The optimal estimate $\hat{x}(t|t)$ is generated by a linear dynamical system of the form

$$\begin{aligned} d\hat{x}(t|t)/dt &= F(t)\hat{x}(t|t) + K(t)\tilde{z}(t|t) \\ \tilde{z}(t|t) &= z(t) - H(t)\hat{x}(t|t) \end{aligned} \qquad \text{(I)}$$

(2) *Canonical form for the dynamical system governing the optimal error.*

$$d\tilde{x}(t|t)/dt = F(t)\tilde{x}(t|t) + G(t)u(t) - K(t)[v(t) + H(t)\tilde{x}(t|t)] \qquad \text{(II)}$$

(3) *Optimal gain.*

$$K(t) = P(t)H'(t)R^{-1}(t) \qquad \text{(III)}$$

(4) *Variance equation.*

$$dP/dt = F(t)P + PF'(t) - PH'(t)R^{-1}(t)H(t)P + G(t)Q(t)G'(t) \qquad \text{(IV)}$$

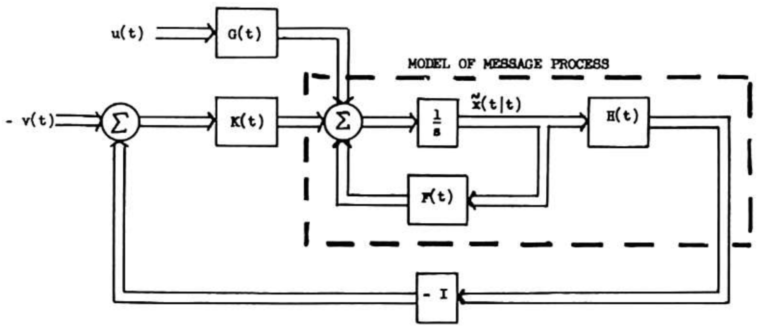
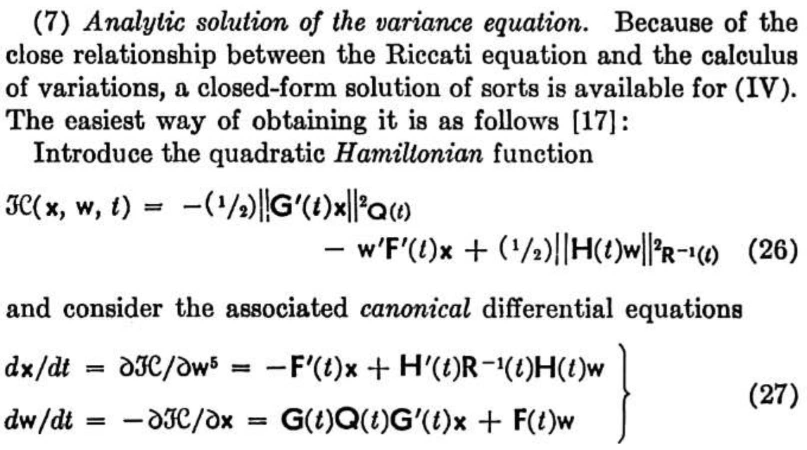


Fig. 10 General block diagram of optimal estimation error

Excerpts from Kalman & Bucy (1961) original work

Why canonical? KB61 answer it:



(7) *Analytic solution of the variance equation.* Because of the close relationship between the Riccati equation and the calculus of variations, a closed-form solution of sorts is available for (IV). The easiest way of obtaining it is as follows [17]:

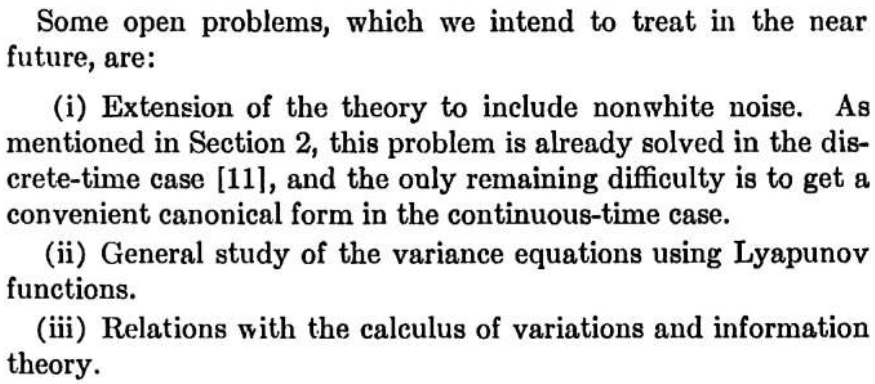
Introduce the quadratic *Hamiltonian* function

$$\mathcal{H}(x, w, t) = -(1/2)\|G'(t)x\|^2_{Q(t)} - w'F'(t)x + (1/2)\|H(t)w\|^2_{R^{-1}(t)} \qquad (26)$$

and consider the associated *canonical* differential equations

$$\left.\begin{aligned} dx/dt &= \partial\mathcal{H}/\partial w^5 = -F'(t)x + H'(t)R^{-1}(t)H(t)w \\ dw/dt &= -\partial\mathcal{H}/\partial x = G(t)Q(t)G'(t)x + F(t)w \end{aligned}\right\} \qquad (27)$$

This is a very important ; note that it provides a link to the adjoint-method employed in meteorology (e.g., Talagrand & Courtier, 1987).



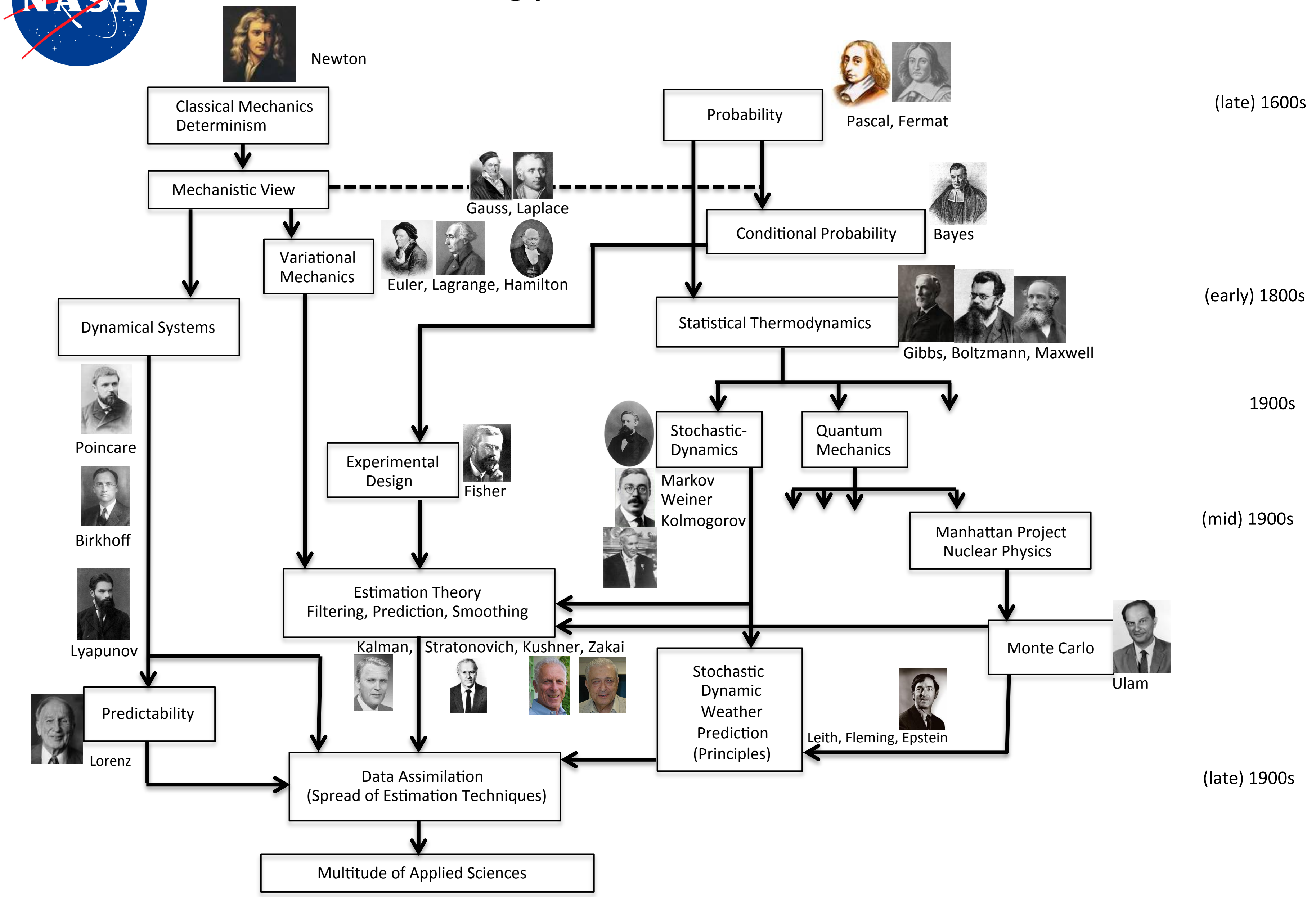
Some open problems, which we intend to treat in the near future, are:

(i) Extension of the theory to include nonwhite noise. As mentioned in Section 2, this problem is already solved in the discrete-time case [11], and the only remaining difficulty is to get a convenient canonical form in the continuous-time case.

(ii) General study of the variance equations using Lyapunov functions.

(iii) Relations with the calculus of variations and information theory.

Genealogy of Data Assimilation



Newton

Classical Mechanics
Determinism

Probability

Pascal, Fermat

(late) 1600s

Mechanistic View

Gauss, Laplace

Conditional Probability

Bayes

Variational
Mechanics

Euler, Lagrange, Hamilton

Dynamical Systems

Statistical Thermodynamics

Gibbs, Boltzmann, Maxwell

(early) 1800s

1900s

Poincare

Experimental
Design

Fisher

Stochastic-
Dynamics

Markov
Weiner
Kolmogorov

Quantum
Mechanics

Birkhoff

Manhattan Project
Nuclear Physics

(mid) 1900s

Lyapunov

Estimation Theory
Filtering, Prediction, Smoothing

Monte Carlo

Ulam

Kalman, Stratonovich, Kushner, Zakai

Predictability

Lorenz

Stochastic
Dynamic
Weather
Prediction
(Principles)

Leith, Fleming, Epstein

Data Assimilation
(Spread of Estimation Techniques)

(late) 1900s

Multitude of Applied Sciences

Based on Lewis, Lakshmivarahan, and Dhall (2006; Fig. 4.6.1)

Two Real-Life Applications

The Apollo Missions

Predicting the Weather

Digression …

The Apollo Missions

The First Real-Life Application of the KF

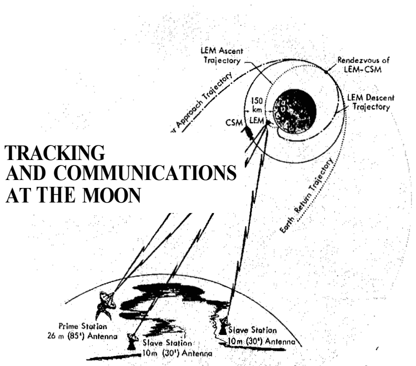
Discovery of the Kalman Filter as a Practical Tool for Aerospace and Industry

In 1985 ***McGee & Schmidt*** published a NASA Tech Memo telling the story of how "the Kalman filter first application was made at NASA Ames during feasibility studies for circumlunar navigation and control of the **Apollo** space capsule".

McGee & Schmidt (1985)

The article describes how:

- The need for something like the Kalman filter arose
- Extensions required to Kalman's work for use in real-life problems
- Various stability tricks were designed and employed
- The need for a stable reformulation leading to the square-root KF
- Various efficient formulations derived to fit the computing real-time-application constraints

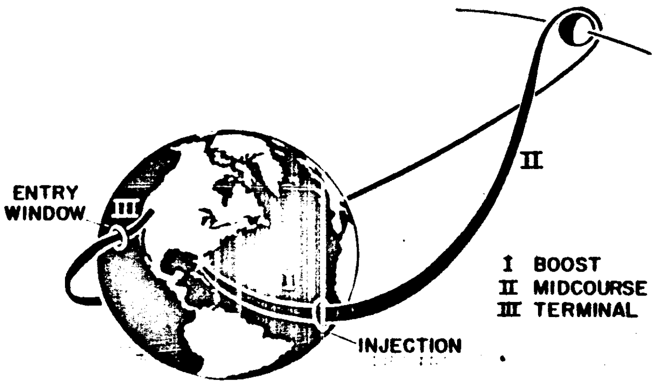

LEM Ascent Trajectory
Rendezvous of LEM-CSM
Approach Trajectory
150 km
CSM
LEM
LEM Descent Trajectory
TRACKING AND COMMUNICATIONS AT THE MOON
Earth Return Trajectory
Prime Station 26 m (85') Antenna
Slave Station 10m (30') Antenna
Slave Station 10m (30') Antenna

From F. O. Vonbun, (1966)

NASA worldwide tracking network (c. 1965)

WHITE SANDS
CORPUS CHRISTI
EGLIN
BERMUDA
CANARY IS.
PT. ARGUELLO
PLA
PLA
PLA
PLA
CSQ
RTK
HAWAII
GUAYMAS
CAPE KENNEDY
GRAND BAHAMA
GRAND TURK
ANTIGUA
KANO
CANTON IS.
RKV
ASCENSION
TANANARIVE
CARNARVON

Source http://history.nasa.gov/SP-4002/p2b.htm



From Schmidt & McLean (1962)

Discovery of the Kalman Filter as a Practical Tool for Aerospace and Industry

McGee & Schmidt (1985)

Support for the Apollo Mission (from mid-1962 to mid-1964)

Three areas of study were the focus:
(1) effect of modeling errors and suboptimal space vehicle trajectory.
(2) effect of short-word length in the airborne computers.
(3) effect of combining ground-base and on-board observational data.

The first **stability issues** with the **Kalman-Schmidt filter** were encountered while studying (3). Earlier investigations apparently involved systems that were less sensitive to nonlinearities.

Part of the stability issue was attributed to computer round-off problems. Initial attempts to address the issue involved (the now familiar) forcing P to be symmetric by:
(a) using only its upper (or lower) triangle to form a symmetric matrix.
(b) averaging its off-diagonal terms.
(c) applying (b), then computing correlations coeffs, if any > 1, stop.
(d) adding a small number to the diagonal of P after measurement and time update steps.

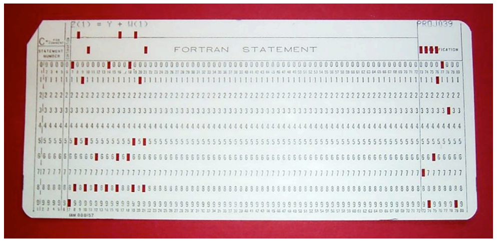
About this time is when **Joseph's update formula** came into play.

During this research **they learned**:
(1) how to handle uncertainties and biases
(2) when the error cov P is too-optimistic it mis-represents errors, leading to filter divergence
(3) ground-base radar obs were more effective, with onboard corrections only used as backup

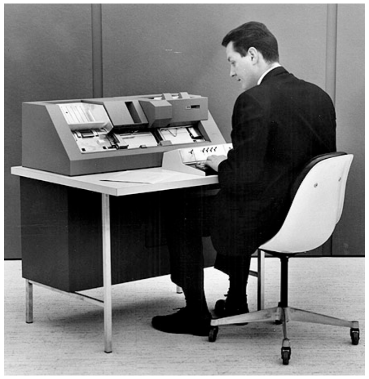
Discovery of the Kalman Filter as a Practical Tool for Aerospace and Industry

Back in the days …

- Computer programs were “typed” in punch cards
- Debugging was tough!
- On the IBM 704 (at Ames) matrix double indexing was slow; programs had to be rewritten with single indexing.
- IBM 704: 36-bit arithmetic; Apollo onboard: 15-bit



Source http://en.wikipedia.org/wiki/Punched_card



Source http://www-03.ibm.com/ibm/history/exhibits/vintage/vintage_4506VV4002.html

Initial Kalman filter studies used
IBM 704 Data Processing System

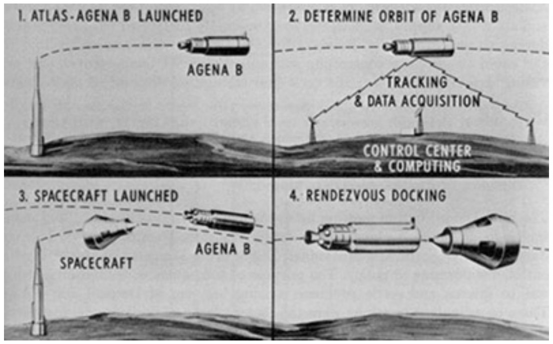
Source http://www-03.ibm.com/ibm/history/exhibits/mainframe/mainframe_2423PH704.html

Apollo 11 Mission Control
IBM’s Real-Time Computer
Complex at NASA, Houston

http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/apollo/breakthroughs/

Discovery of the Kalman Filter as a Practical Tool for Aerospace and Industry

Application of The Filter to the Agena Program (c. 1961)


1. ATLAS-AGENA B LAUNCHED
AGENA B
2. DETERMINE ORBIT OF AGENA B
TRACKING & DATA ACQUISITION
CONTROL CENTER & COMPUTING
3. SPACECRAFT LAUNCHED
AGENA B
SPACECRAFT
4. RENDEZVOUS DOCKING

Source http://history.nasa.gov/SP-4002/p1b.htm

Gemini Docking

Purpose: validate the Agena upper stage rendezvous and docking during Project Gemini.

Observations: downrange stations & in-flight telemetry.

Model: equations of motion of the vehicle predicting position and velocity.

Estimator: measurement biases, location, coefficients of propulsion model for the thrust of Agena upper stage.

Resulting techniques following from study:

- Quality control: data-rejected based on size of residual.
- KF used as data compression algorithm.
- Effect of nonlinearities handled with backward integration and forward filtering.
- KF used to estimate parameters in measurement and model.

McGee & Schmidt (1985)

Discovery of the Kalman Filter as a Practical Tool for Aerospace and Industry

The square-root filter

Schmidt and his group continued to applied the **extended KF** to various navigation problems:

(a) the development of the C-5A aircraft navigation system

(b) flight test of the RAINPAL system for approach and landing

Still having stability issues Schmidt, knowing about the recently developed algorithm by ***Potter***, implemented the first square-root filtering for real on-board aircraft applications. Potter's procedure uses a Cholesky factorization of the error covariance matrix, which by construction maintains positiveness, and results in a more stable implementation then the more direct Extended KF..

The group of ***Eldon Hall*** implemented Potter's algorithm in the Apollo Guidance Computer.

Potter's original algorithm neglects model error. Various generalizations become available in the late 60s and during the 70s that by then took into account factorizations the model error covariance – amount the great contributors where ***Carlson, Bierman, & Thornton***. The most reliable and computationally efficient schemes are based on a **U-D decomposition** of the error covariance and a **modified Gram-Schmidt orthogonalization**.

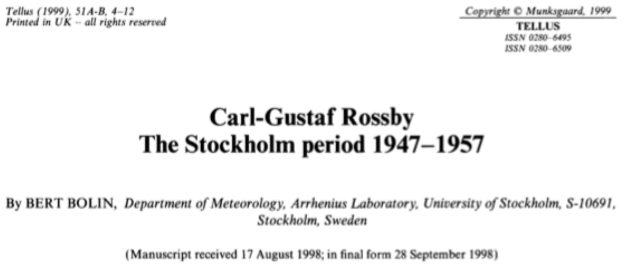
McGee & Schmidt (1985)

Predicting the Weather

From First Principles to the KF for NWP

Meteorology & Weather Forecasting

There is a number of articles that tell the history of meteorology, weather forecasting, & of those who pioneered the field.



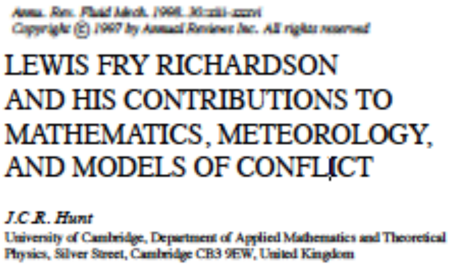
Tellus (1999), 51A-B, 4–12
Printed in UK – all rights reserved

Copyright © Munksgaard, 1999
TELLUS
ISSN 0280–6495
ISSN 0280–6509

Carl-Gustaf Rossby
The Stockholm period 1947–1957

By BERT BOLIN, *Department of Meteorology, Arrhenius Laboratory, University of Stockholm, S-10691, Stockholm, Sweden*

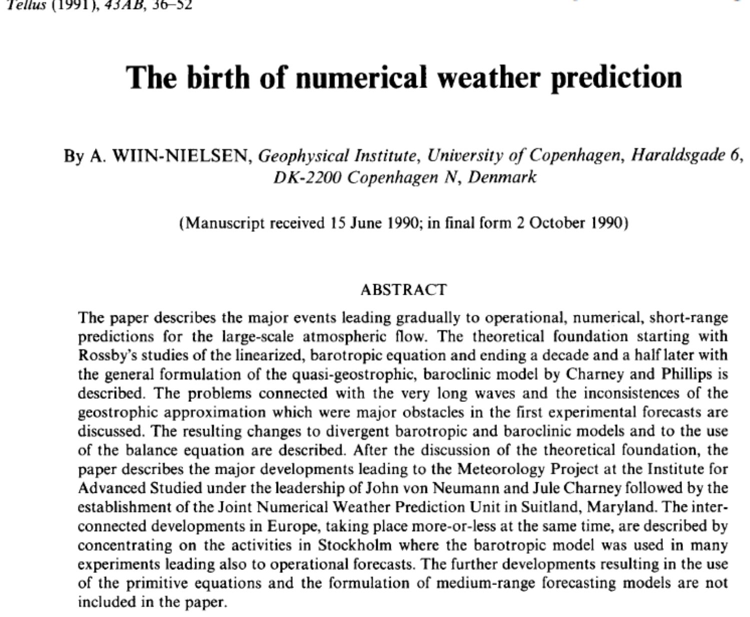
(Manuscript received 17 August 1998; in final form 28 September 1998)



Annu. Rev. Fluid Mech. 1998. 30:xiii–xxxvi
Copyright © 1997 by Annual Reviews Inc. All rights reserved

LEWIS FRY RICHARDSON
AND HIS CONTRIBUTIONS TO
MATHEMATICS, METEOROLOGY,
AND MODELS OF CONFLICT

J.C.R. Hunt
University of Cambridge, Department of Applied Mathematics and Theoretical Physics, Silver Street, Cambridge CB3 9EW, United Kingdom



Tellus (1991), *43AB*, 36–52

The birth of numerical weather prediction

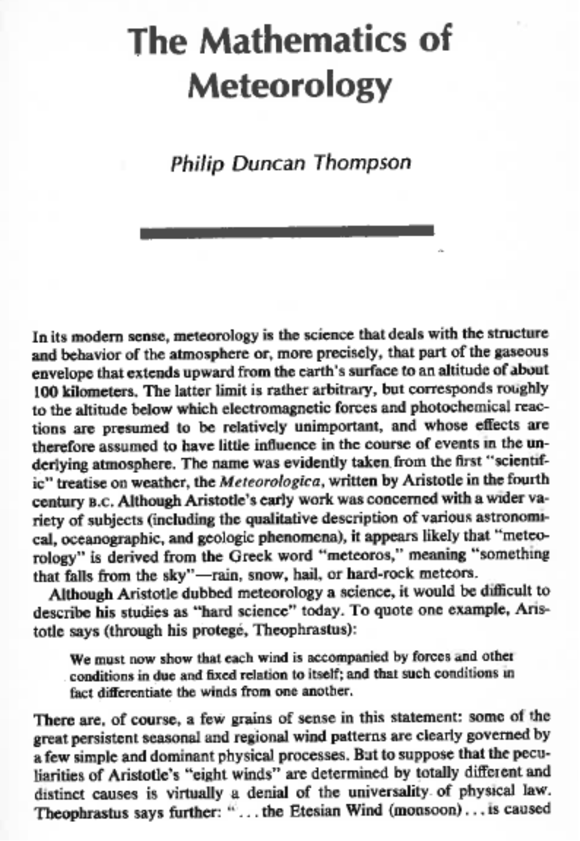
By A. WIIN-NIELSEN, *Geophysical Institute, University of Copenhagen, Haraldsgade 6, DK-2200 Copenhagen N, Denmark*

(Manuscript received 15 June 1990; in final form 2 October 1990)

ABSTRACT

The paper describes the major events leading gradually to operational, numerical, short-range predictions for the large-scale atmospheric flow. The theoretical foundation starting with Rossby's studies of the linearized, barotropic equation and ending a decade and a half later with the general formulation of the quasi-geostrophic, baroclinic model by Charney and Phillips is described. The problems connected with the very long waves and the inconsistences of the geostrophic approximation which were major obstacles in the first experimental forecasts are discussed. The resulting changes to divergent barotropic and baroclinic models and to the use of the balance equation are described. After the discussion of the theoretical foundation, the paper describes the major developments leading to the Meteorology Project at the Institute for Advanced Studied under the leadership of John von Neumann and Jule Charney followed by the establishment of the Joint Numerical Weather Prediction Unit in Suitland, Maryland. The interconnected developments in Europe, taking place more-or-less at the same time, are described by concentrating on the activities in Stockholm where the barotropic model was used in many experiments leading also to operational forecasts. The further developments resulting in the use of the primitive equations and the formulation of medium-range forecasting models are not included in the paper.

Mathematics Today, 1978, L. A. Steen, Ed.127-152



The Mathematics of Meteorology

Philip Duncan Thompson

In its modern sense, meteorology is the science that deals with the structure and behavior of the atmosphere or, more precisely, that part of the gaseous envelope that extends upward from the earth's surface to an altitude of about 100 kilometers. The latter limit is rather arbitrary, but corresponds roughly to the altitude below which electromagnetic forces and photochemical reactions are presumed to be relatively unimportant, and whose effects are therefore assumed to have little influence in the course of events in the underlying atmosphere. The name was evidently taken from the first "scientific" treatise on weather, the *Meteorologica*, written by Aristotle in the fourth century B.C. Although Aristotle's early work was concerned with a wider variety of subjects (including the qualitative description of various astronomical, oceanographic, and geologic phenomena), it appears likely that "meteorology" is derived from the Greek word "meteoros," meaning "something that falls from the sky"—rain, snow, hail, or hard-rock meteors.

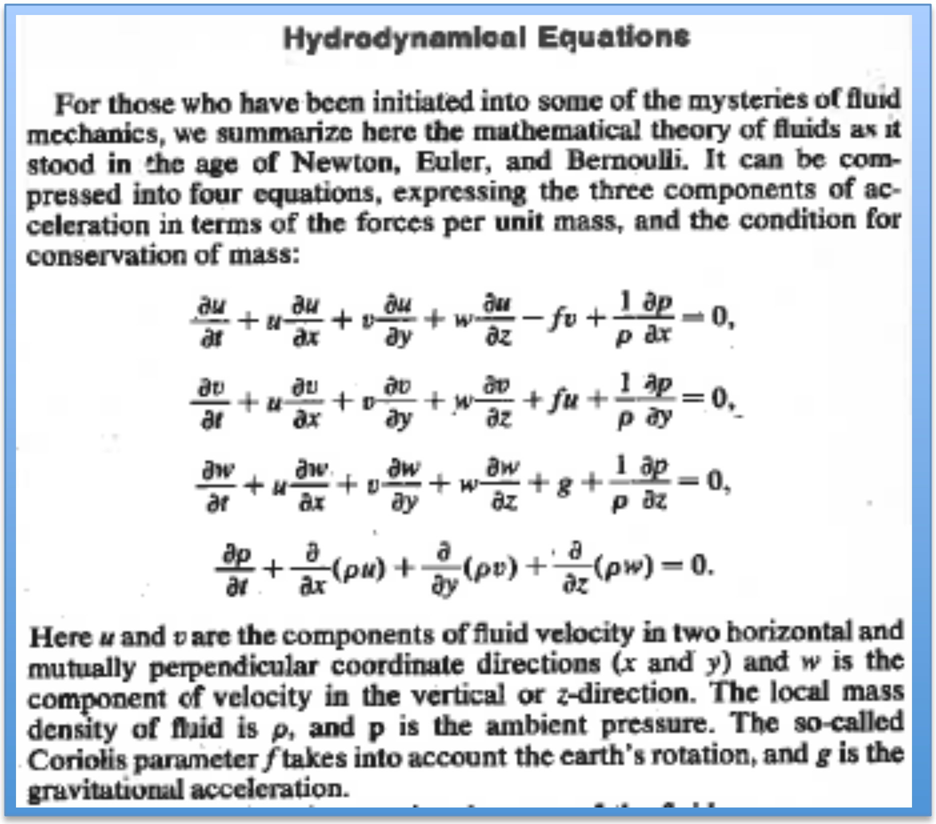
Although Aristotle dubbed meteorology a science, it would be difficult to describe his studies as "hard science" today. To quote one example, Aristotle says (through his protegé, Theophrastus):

> We must now show that each wind is accompanied by forces and other conditions in due and fixed relation to itself; and that such conditions in fact differentiate the winds from one another.

There are, of course, a few grains of sense in this statement: some of the great persistent seasonal and regional wind patterns are clearly governed by a few simple and dominant physical processes. But to suppose that the peculiarities of Aristotle's "eight winds" are determined by totally different and distinct causes is virtually a denial of the universality of physical law. Theophrastus says further: "... the Etesian Wind (monsoon) ... is caused

Meteorology & Weather Forecasting

Why did it take so long for meteorology to become a science?



Hydrodynamical Equations

For those who have been initiated into some of the mysteries of fluid mechanics, we summarize here the mathematical theory of fluids as it stood in the age of Newton, Euler, and Bernoulli. It can be compressed into four equations, expressing the three components of acceleration in terms of the forces per unit mass, and the condition for conservation of mass:

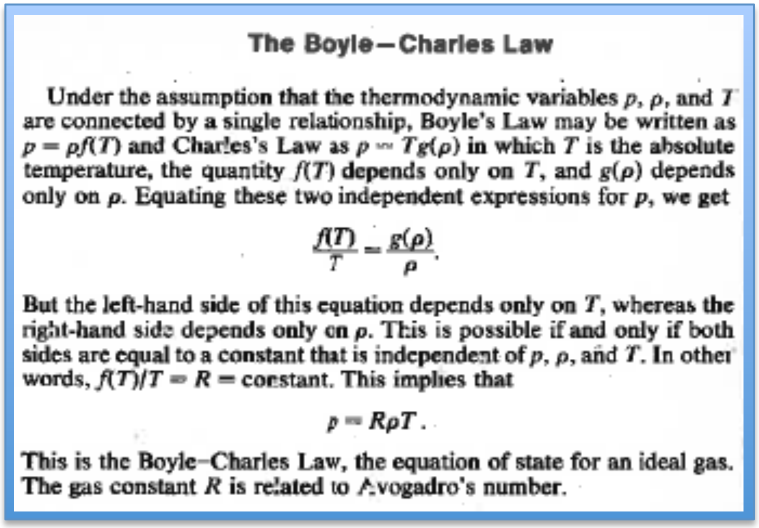
$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} + w\frac{\partial u}{\partial z} - fv + \frac{1}{\rho}\frac{\partial p}{\partial x} = 0,$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} + w\frac{\partial v}{\partial z} + fu + \frac{1}{\rho}\frac{\partial p}{\partial y} = 0,$$

$$\frac{\partial w}{\partial t} + u\frac{\partial w}{\partial x} + v\frac{\partial w}{\partial y} + w\frac{\partial w}{\partial z} + g + \frac{1}{\rho}\frac{\partial p}{\partial z} = 0,$$

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) + \frac{\partial}{\partial y}(\rho v) + \frac{\partial}{\partial z}(\rho w) = 0.$$

Here u and v are the components of fluid velocity in two horizontal and mutually perpendicular coordinate directions (x and y) and w is the component of velocity in the vertical or z-direction. The local mass density of fluid is ρ, and p is the ambient pressure. The so-called Coriolis parameter f takes into account the earth's rotation, and g is the gravitational acceleration.



The Boyle–Charles Law

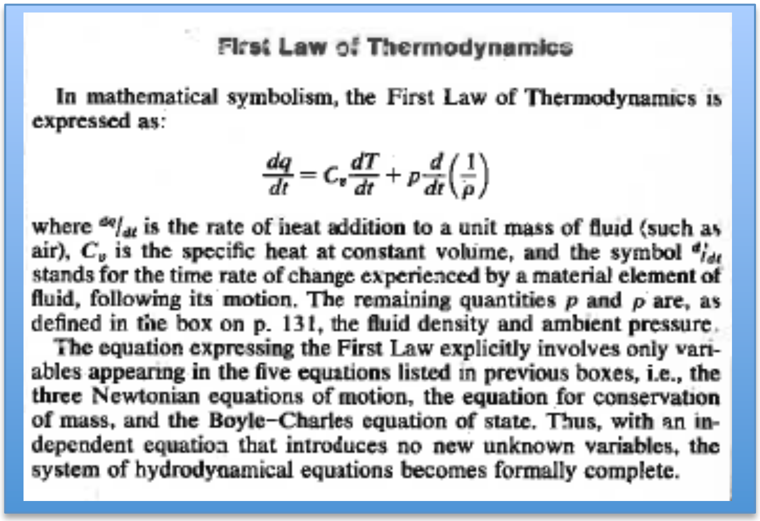
Under the assumption that the thermodynamic variables p, ρ, and T are connected by a single relationship, Boyle's Law may be written as $p = \rho f(T)$ and Charles's Law as $p = Tg(\rho)$ in which T is the absolute temperature, the quantity $f(T)$ depends only on T, and $g(\rho)$ depends only on ρ. Equating these two independent expressions for p, we get

$$\frac{f(T)}{T} = \frac{g(\rho)}{\rho}.$$

But the left-hand side of this equation depends only on T, whereas the right-hand side depends only on ρ. This is possible if and only if both sides are equal to a constant that is independent of p, ρ, and T. In other words, $f(T)/T = R =$ constant. This implies that

$$p = R\rho T.$$

This is the Boyle–Charles Law, the equation of state for an ideal gas. The gas constant R is related to Avogadro's number.



First Law of Thermodynamics

In mathematical symbolism, the First Law of Thermodynamics is expressed as:

$$\frac{dq}{dt} = C_v\frac{dT}{dt} + p\frac{d}{dt}\left(\frac{1}{\rho}\right)$$

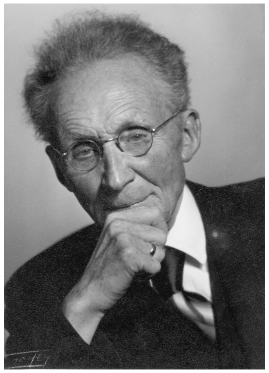
where ${}^{dq}/_{dt}$ is the rate of heat addition to a unit mass of fluid (such as air), C_v is the specific heat at constant volume, and the symbol ${}^{d}/_{dt}$ stands for the time rate of change experienced by a material element of fluid, following its motion. The remaining quantities p and ρ are, as defined in the box on p. 131, the fluid density and ambient pressure.

The equation expressing the First Law explicitly involves only variables appearing in the five equations listed in previous boxes, i.e., the three Newtonian equations of motion, the equation for conservation of mass, and the Boyle–Charles equation of state. Thus, with an independent equation that introduces no new unknown variables, the system of hydrodynamical equations becomes formally complete.

Note: thought the system of equations becomes formally complete with six equations, for it to describe a meaningful atmosphere, it also needs a seventh eq. provided by the **Second Law of Thermodynamics** – leading to inclusion of water-vapor.

From Thompson; in Mathematics Today 1978

Meteorology & Weather Forecasting



Vilhelm Bjerknes
From www.uib.no

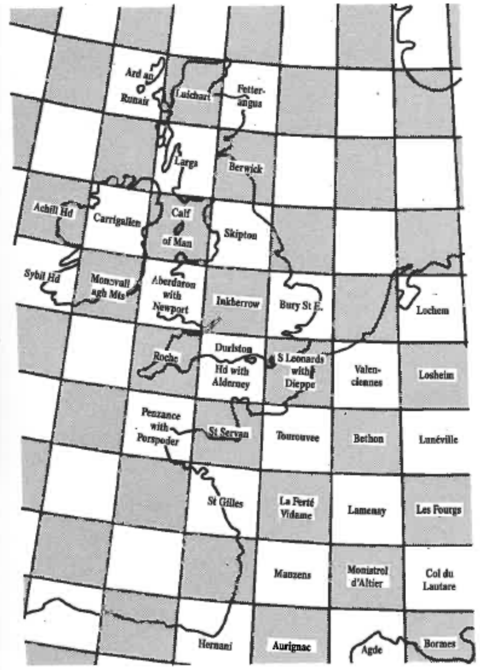
“If it is true, as every scientist believes, that atmospheric states develop from the preceding ones according to physical laws, then it is apparent that the necessary and sufficient conditions for the rational solution of forecasting problems are the following:

1. A sufficiently accurate knowledge of the state of the atmosphere at the initial time.
2. A sufficiently accurate knowledge of the laws according to which one state of the atmosphere develops from another.“

Bjerknes (1904; Meteor. Zeitschrift)

“Perhaps some day in the dim future it will be possible to advance the computations faster than the weather advances and at a cost less than the saving to mankind due to the information gained. But that is a dream.”

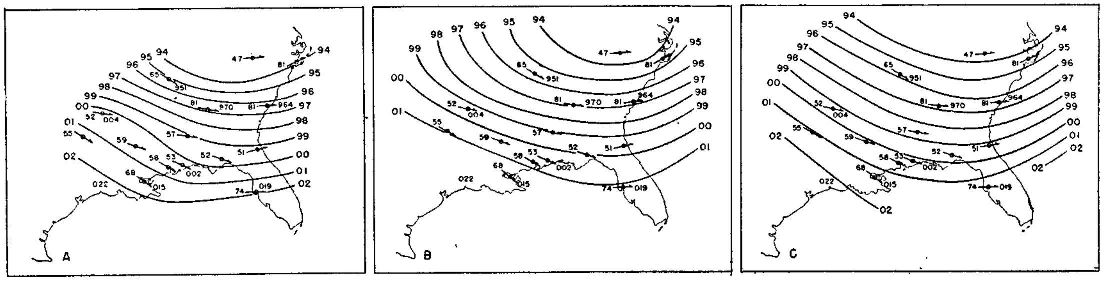
*Lewis Fry Richardson, 1922***.**



Richarson’s checkerboard grid with p and wind staggered at shaded and clear boxes.

Lewis F. Richardson
From www.wmo.int

Objective Analysis and The Variational Principle (c. 1950s)



Subjective (left) and two objective analyses of 700 mb height at 1500 GCT on 25 March 1947; From ***Panofsky*** (1949).

The objective analysis amounts to a third-order polynomial fit

$$p(x, y) = \sum_{i,j} a_{ij}x^i y^j, \qquad (i + j \leqq 3)$$

Many contributed to objective Analysis: ***Cressman, Charney, Platzman, Smagorinsky***, others.

It was quickly realized that simple objective analysis techniques would have to be made consistent with the physical constrains underlying the meteorological variables. ***Sasaki*** proposed using the variational principle to accomplish consistency.

Yoshikazu ("Yoshi") ***Sasaki***

$$I = \int_V \mathcal{E}^2 dx dy dp^* \qquad \delta I = 0$$

Governing QG and thermal wind eqs.:

$$u = -\frac{1}{f}\frac{\partial \phi}{\partial y}, \qquad v = \frac{1}{f}\frac{\partial \phi}{\partial x}$$

$$\frac{\partial u}{\partial p} = \frac{R}{pf}\frac{\partial T}{\partial y}, \qquad \frac{\partial v}{\partial p} = -\frac{R}{pf}\frac{\partial T}{\partial x}$$

Deviations from observations:

$$u' \equiv u - u_0$$
$$v' \equiv v - v_0$$
$$\phi' \equiv \phi - \phi_0$$
$$T' \equiv T - T_0$$

Corresponding error equations:

$$u' = -\frac{1}{f}\frac{\partial \phi'}{\partial y} - u_0 - \frac{1}{f}\frac{\partial \phi_0}{\partial y}$$
$$v' = \frac{1}{f}\frac{\partial \phi'}{\partial x} - v_0 + \frac{1}{f}\frac{\partial \phi_0}{\partial x}$$
$$T' = \frac{\partial \phi'}{\partial p^*} - T_0 + \frac{\partial \phi_0}{\partial p^*}$$

The total quadratic error in the integrand:

$$\mathcal{E}^2 \equiv \alpha_1{}^2 u'^2 + \alpha_1{}^2 v'^2 + \alpha_2{}^2 \phi'^2 + \alpha_3{}^2 T'^2$$

Remaining difficulties recognized:

- Extrapolation beyond area of available data.
- Specification of weights

From Sasaki (1958)

Discovery of the Kalman Filter as a Practical Tool for Numerical Weather Prediction

The earliest mention of the Kalman filter as a possible approach to initialize NWP models is found in a publication in the Journal of Atmospheric Sciences, by **Richard H. Jones**, in **1965**.

Computational complexity kept most from looking into the KF for real-time NWP applications.

It wasn't until the early 1980s that **M**. **Ghil, S. E. Cohn**, & **D. P**. **Dee** started looking at the problem and studying the KF properties for hyperbolic PDEs (associated with NWP).

Though works on KF for NWP started appearing more often, it wasn't until **1994** with **Geir Evensen's ensemble Kalman filter** that the feasibility of using the filter for real-time weather applications started sinking in.

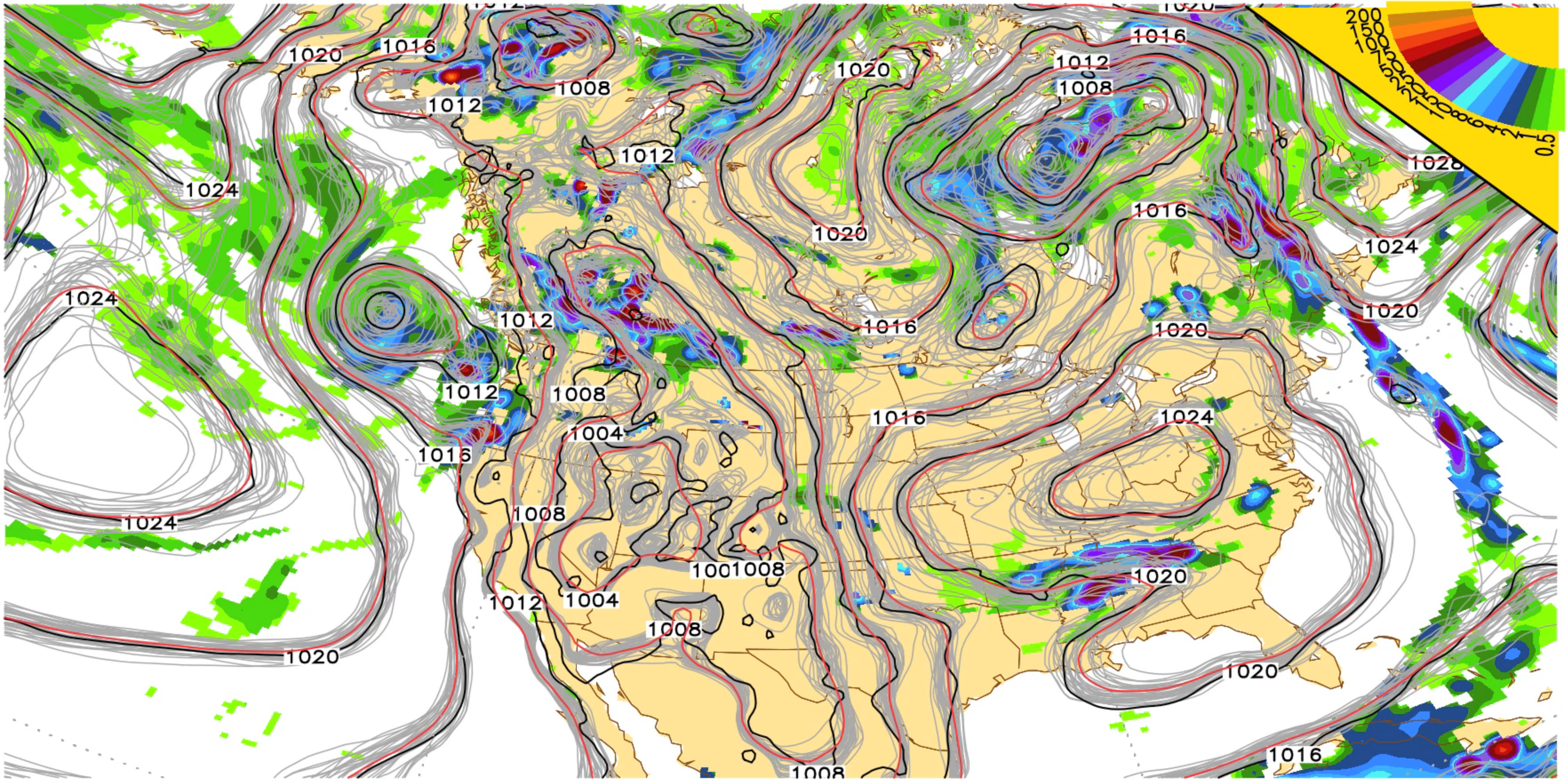
Since then, the literature on **Kalman filtering** related to **NWP** (and other Earth Sciences applications) has exploded. Many weather centers now have some version of an ensemble-based data assimilation procedure implemented; some of these being **EnKF's.**

Most interestingly, many of the EnKF's fit under the banner of **Square-Root Kalman Filters**. So, in some sense, it seems we have come all the way around to conclude (for somewhat slightly different reasons), that Square-Root filters are better suited for practical applications.

The NASA GMAO Variational-Ensemble Hybrid Data Assimilation System

NASA/GMAO – GEOS Central Analysis and 32–Member Ensemble Analysis

Central: Precip [mm], SLP [mb] (black); Ens Mean SLP [mb] (red)



Analysis and Ensemble Analysis on Sunday 2:00 AM EDT 2014–05–18

Closing Remarks

In the process of preparing this presentation I came across an article not too dissimilar from that of **McGee & Schmidt** (1985). This is the article of **Grewal & Andrews** (2010) which also provides a nice review of the use of Kalman filtering in Aerospace. It seems unfortunate, though, these authors are not aware of the earlier review of McGee & Schmidt.

In our Earth Science applications, the **square-root filter** formulation has become rather important as it is behind the ensemble-based formulations for the filtering (and smoothing) problem(s).

Just in our field of interest, the amount of literature on filtering and smoothing has explored. it is becoming very difficult to know all available variations of possible twists to the solution equations. But it seems that those who've made the larger strides in progress in our field have given special attention **not only** to the **assimilation strategy**, **but also** to **how to treat the observations** being assimilation:

(a) removal of **biases**
(b) specification of underlying **error statistics**
(c) treatment of **balance**
(d) and a host of other **details**

have all been fundamental to progress in Estimation Techniques for Earth Sciences.

Bibliographical Sources

Wikipedia: https://en.wikipedia.org

Barrow, J. D., & F. J. Tipler, 1986: The Anthropic Cosmological Principle, Oxford University Press.

Erhendorfer, M., 2007: Meteorologische Zeitschrift, 16, 795-818.

Hacking, I., 1975: The Emergence of Probability, Cambridge University Press.

Franklin, J., 2001: The Science of Conjecture: Evidence of Probability before Pascal, Johns Hopkins University Press.

Feinberg, S. E., 2006: Bayesian Analysis, 1, 1-40.

Ghil, M., S. E. Cohn, J. Tavantzis, K. Bube, & E. Isaacson, 1981: Applications of estimation theory to weather prediction. In L. Bengtsson, M. Ghil, and E. Kallen, eds., 1981, Dynamic Meteorology, Applied Mathematical Sciences, Springer-Verlag, New York.

Grewal, M. S. & A. P. Andrews, 2010: IEEE Control Systems Magazine, 69-78.

Hunt, J.C.R., 1998: Annu. Rev. Fluid Mech. 30:xiii–xxxvi.

Lanczos, C., 1949: The Variational Principle of Mechanics, Dover 4th ed.

Kalman, R. E, 1960: ASME-D, J. Basic Eng., 82, 35-45.

Kalman, R. E, & R. S. Bucy, 1961: ASME-D, J. Basic Eng., 83, 95-108.

Laplace, P.-S. (1774). “‘Memoire sur la Probabilite des Causes par les Evenements.” Memoires de Mathematique et de Physique Presentes a l’Academie Royale des Sciences, Par Divers Savans, & Luˆs dans ses Assemblees, 6:621–656.

Lewis, J.M. 2005: Mon. Wea. Rev., 133, 1865-1885.

Lewis, J.M., S. Lakshmivarahan, & S. K. Dhall, 2006: Dynamics Data Assimilation: A least squares approach. Cambridge University Press.

Maybeck, P. S., 1982: Stochastic models, estimation, and control. Vol. 2, Academic Press, NY

McGee & Schmidt, 1985: NASA TM 86847.

Sasaki, Y. (1958): J. Meteorol.Soc. Japan, 36, 1-12.

Schmidt, G. L. & J. D. McLean, 1962: NASA TN D-1208.

Talangrand, O., & P. Courtier, 1987: Q. J. R. Meterol. Soc., 113, 1311-1328.

Tarantola, A., 2005: Inverse Problem Theory. Elsevier, New York.

Vonbun, F.O. 1966: Astronautics & Aeronautics, 104-115.

Introduction to Data Assimilation (alternatively) Introduction to Estimation Theory

Ricardo Todling

NASA Global Modeling and Assimilation Office

First presented to GMAO group discussion c 2001

Outline

- ➢Concepts of Probabilistic Estimation
- ➢Example: Estimation of a Constant Vector
- ➢The Three-dimensional Variational Approach
- ➢The Four-dimensional Variational Approach
- ➢The Probabilistic Approach to Filtering
- ➢Simple Illustrations and Points to Remember

Main Objective

The main objective of this lecture is to present a summary of some of the methods most commonly used for state estimation.

What I hope to convey to you:

- ▷ The *probabilistic approach* allows for the proper description of most (if not all) methods currently employed in data assimilation.
- ▷ In practice, most methods used in atmospheric and oceanic data assimilation boil down to slightly different versions of *least-squares*.
- ▷ Good understanding of the example of "estimation of a constant vector" provides a solid basis for understanding many of the methods currently used.
- ▷ Much attention should be given to details:
  - off-line and on-line quality control
  - removal of both model and observation biases
  - proper use of observations; they should be used at right time and be given proper error characteristics
  - fields should be properly initialized
  - careful consideration of tangent linear and adjoint models issues
- ▷ Remember ... *adaptive procedures are robust*.

Bayesian Approach to Estimation

Central to probabilistic estimation is the concept of a joint probability distribution (pdf) of two processes $\mathbf{x}$ and $\mathbf{y}$, and denoted $p_{\mathrm{x,y}}(\mathbf{x},\mathbf{y})$.

Also, fundamental to Bayesian estimation is the definition of conditional probability distribution functions:

$$p_{\mathrm{x|y}}(\mathbf{x}|\mathbf{y}) = \frac{p_{\mathrm{xy}}(\mathbf{x},\mathbf{y})}{p_{\mathrm{y}}(\mathbf{y})}$$

and Bayes rule for converting between conditional pdf's:

$$p_{\mathrm{x|y}}(\mathbf{x}|\mathbf{y}) = \frac{p_{\mathrm{y|x}}(\mathbf{y}|\mathbf{x})p_{\mathrm{x}}(\mathbf{x})}{p_{\mathrm{y}}(\mathbf{y})}$$

The m-th conditional moment is defined as:

$$\mathcal{E}\{\mathbf{x}^m|\mathbf{y}\} \equiv \int_{-\infty}^{\infty} \mathbf{x}^m \, p_{\mathrm{x|y}}(\mathbf{x}|\mathbf{y})$$

with the first moment, the mean, $\boldsymbol{\mu}_{\mathrm{x|y}} = \mathcal{E}\{\mathbf{x}|\mathbf{y}\}$.

A typical conditional pdf is that of a normally distributed random variable $\mathbf{x}$ conditioned on $\mathbf{y}$

$$p_{\mathrm{x|y}}(\mathbf{x}|\mathbf{y}) = \frac{1}{(2\pi)^{n/2}|\mathbf{P}_{\mathrm{x|y}}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{\mathrm{x|y}})^T \mathbf{P}_{\mathrm{x|y}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathrm{x|y}})\right]$$

which is a n-dimensional Gaussian function.

Bayesian Approach to Estimation

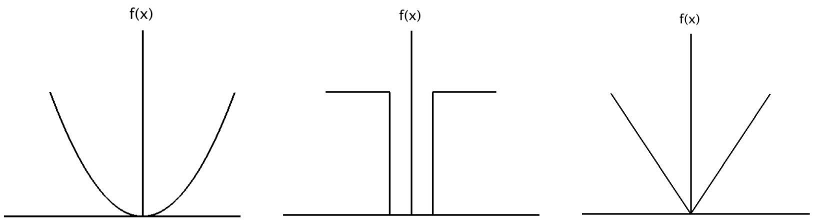
In the Bayesian approach to estimation we define a function expressing our confidence in the estimate. This function is referred to as the cost (or risk, or fit) function and it takes the general form:

$$\begin{aligned}\mathcal{J}(\hat{\mathbf{x}}) &\equiv \mathcal{E}\{J(\mathbf{x}-\hat{\mathbf{x}})\} \\ &= \int_{-\infty}^{\infty} J(\mathbf{x}-\hat{\mathbf{x}})\, p_{\mathrm{x}}(\mathbf{x})\, d\mathbf{x} \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} J(\mathbf{x}-\hat{\mathbf{x}})\, p_{\mathrm{xy}}(\mathbf{x},\mathbf{y})\, d\mathbf{y}\, d\mathbf{x}\end{aligned}$$

where

$\mathbf{x}$	true state vector
$\mathbf{y}$	observation vector
$\hat{\mathbf{x}}$	state estimate vector
$\tilde{\mathbf{x}} = \mathbf{x} - \hat{\mathbf{x}}$	error estimate vector
$J(\tilde{\mathbf{x}})$	measure of accuracy
$p_{\mathrm{x}}(\mathbf{x})$	marginal pdf of $\mathbf{x}$
$p_{\mathrm{xy}}(\mathbf{x},\mathbf{y})$	joint pdf between $\mathbf{x}$ and $\mathbf{y}$

Note: Not all function J's are satisfactory cost functions.



Quadratic error, uniform error and absolute-value error cost functions, for constant parameter.

A Few Examples of Cost Functions

(a) The quadratic cost:

$$J = \frac{1}{2}||\mathbf{x}-\hat{\mathbf{x}}||_{\mathbf{E}}^2 = \frac{1}{2}(\mathbf{x}-\hat{\mathbf{x}})^T\mathbf{E}(\mathbf{x}-\hat{\mathbf{x}})$$

(b) The absolute-value cost:

$$J = \frac{1}{2}||\mathbf{x}-\hat{\mathbf{x}}||_{\mathbf{E}} = \frac{1}{2}||\mathbf{E}(\mathbf{x}-\hat{\mathbf{x}})||$$

(c) The uniform cost:

$$J = \begin{cases} 0\,, & ||\mathbf{x}-\hat{\mathbf{x}}|| < \epsilon \\ 1/2\epsilon\,, & ||\mathbf{x}-\hat{\mathbf{x}}|| \geq \epsilon \end{cases}$$

(d) The Huber (1964) cost:

$$J = \begin{cases} ||\mathbf{x}-\hat{\mathbf{x}}||^2\,, & |\mathbf{x}-\hat{\mathbf{x}}| \leq \delta \\ \delta(||\mathbf{x}-\hat{\mathbf{x}}|| - \frac{1}{2}\delta)\,, & \text{otherwise} \end{cases}$$

A desirable property of an estimate is that it be unconditionally unbiased, that is,

$$\mathcal{E}\{\hat{\mathbf{x}}\} = \mathcal{E}\{\mathbf{x}\}$$

Sometimes the estimate is conditionally unbiased:

$$\mathcal{E}\{\hat{\mathbf{x}}|\mathbf{x}\} = \mathbf{x}$$

Estimating a Constant Vector from Noisy Observations

Estimators

Bayes rule for pdf's:

$$p_{\mathrm{x}|\mathrm{y}}(\mathbf{x}|\mathbf{y}) = \frac{p_{\mathrm{y}|\mathrm{x}}(\mathbf{y}|\mathbf{x})p_{\mathrm{x}}(\mathbf{x})}{p_{\mathrm{y}}(\mathbf{y})}$$

Conditional mean:

$$\mathcal{E}\{\mathbf{x}|\mathbf{y}\} \equiv \int_{-\infty}^{\infty} \mathbf{x}\, p_{\mathrm{x}|\mathrm{y}}(\mathbf{x}|\mathbf{y})$$

Minimum variance estimate:

$$\begin{aligned} \hat{\mathbf{x}}_{\mathsf{MV}}(\mathbf{y}) &= \int_{-\infty}^{\infty} \mathbf{x} p_{\mathrm{x}|\mathrm{y}}(\mathbf{x}|\mathbf{y})\, d\mathbf{x} \\ &= \mathcal{E}\{\mathbf{x}|\mathbf{y}\} \end{aligned}$$

Maximum *a posteriori* probability estimate:

$$\left.\frac{\partial p_{\mathrm{y}|\mathrm{x}}(\mathbf{y}|\mathbf{x})p_{\mathrm{x}}(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathrm{x}=\hat{\mathrm{x}}_{\mathsf{MAP}}} = \mathbf{0}$$

Maximum likelihood estimate (max *a priori* pdf):

$$\left.\frac{\partial p_{\mathrm{y}|\mathrm{x}}(\mathbf{y}|\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathrm{x}=\hat{\mathrm{x}}_{\mathsf{ML}}} = \mathbf{0}$$

Observer and Solutions

Observations: $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{b}^o$

Want to determine: $p_{\mathrm{x}|\mathrm{y}}(\mathbf{x}|\mathbf{y})$

when $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{P})$, and $\mathbf{b}^o \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$, we find:

$$p_{\mathrm{x}|\mathrm{y}}(\mathbf{x}|\mathbf{y}) \;\alpha\; \mathrm{exp}[-\frac{1}{2}(\mathbf{x}-\hat{\mathbf{x}})^T \mathbf{P}_{\tilde{\mathrm{x}}}^{-1}(\mathbf{x}-\hat{\mathbf{x}})]$$

where

$$\mathbf{P}_{\tilde{\mathrm{x}}}^{-1} = \mathbf{P}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\,,$$

and

$$\hat{\mathbf{x}} = \mathbf{P}_{\tilde{\mathrm{x}}}(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{y} + \mathbf{P}^{-1}\boldsymbol{\mu})$$

General Cost Function:

$$J(\mathbf{x}) = \frac{1}{2}(\boldsymbol{\mu}-\mathbf{x})^T\mathbf{P}^{-1}(\boldsymbol{\mu}-\mathbf{x}) + \frac{1}{2}(\mathbf{y}-\mathbf{H}\mathbf{x})^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x})$$

Estimation Results:

$$\hat{\mathbf{x}}_{\mathsf{MV}} = \hat{\mathbf{x}}_{\mathsf{MAP}} = \hat{\mathbf{x}}$$

$$\hat{\mathbf{x}}_{\mathsf{ML}} = \mathbf{P}_{\tilde{\mathrm{x}}}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{y}$$

$$\hat{\mathbf{x}}_{\mathsf{MV}}|_{\mathbf{P}^{-1}=\mathbf{0}} = \hat{\mathbf{x}}_{\mathsf{MAP}}|_{\mathbf{P}^{-1}=\mathbf{0}} = \hat{\mathbf{x}}_{\mathsf{ML}}$$

The Least Squares Connection

Case I: No prior information on $\mathbf{x}$ is available.

Minimization of the cost function

$$J_{\mathsf{LS}}(\hat{\mathbf{x}}) = \frac{1}{2}(\mathbf{y} - \mathbf{H}\hat{\mathbf{x}})^T\tilde{\mathbf{R}}^{-1}(\mathbf{y} - \mathbf{H}\hat{\mathbf{x}})$$

results in

$$\hat{\mathbf{x}}_{\mathsf{LS}} = (\mathbf{H}^T\tilde{\mathbf{R}}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\tilde{\mathbf{R}}^{-1}\mathbf{y}$$

which is identical to the ML (MV/MAP) estimate(s) if $\tilde{\mathbf{R}} = \mathbf{R}$. In general, however, the LS solution can be shown to always be less accurate than that of ML (MV/MAP).

Case II: Some information on $\mathbf{x}$ is available.

The cost function to be minimized is now

$$J_{\mathsf{LSP}}(\hat{\mathbf{x}}) = \frac{1}{2}(\boldsymbol{\mu} - \hat{\mathbf{x}})^T\tilde{\mathbf{P}}^{-1}(\boldsymbol{\mu} - \hat{\mathbf{x}}) + \frac{1}{2}(\mathbf{y} - \mathbf{H}\hat{\mathbf{x}})^T\tilde{\mathbf{R}}^{-1}(\mathbf{y} - \mathbf{H}\hat{\mathbf{x}})$$

with minimum achieved for

$$\hat{\mathbf{x}}_{\mathsf{LSP}} = (\tilde{\mathbf{P}}^{-1} + \mathbf{H}^T\tilde{\mathbf{R}}^{-1}\mathbf{H})^{-1}(\mathbf{H}^T\tilde{\mathbf{R}}^{-1}\mathbf{y} + \tilde{\mathbf{P}}^{-1}\boldsymbol{\mu})$$

which is identical to the MV/MAP estimate if $\tilde{\mathbf{R}} = \mathbf{R}$ and $\tilde{\mathbf{P}} = \mathbf{P}$. In general, however, the LSP solution can be shown to be always less accurate than that of MV/MAP.

Three-dimensional Variational Approach

The approach known in atmospheric data assimilation as 3d-var is essentially a least squares method that in the linear sense minimizes the cost function $J_{\mathsf{LSP}}(\mathbf{x})$ seen previously,

$$J_{\mathsf{LSP}}(\mathbf{x}) = \frac{1}{2}(\boldsymbol{\mu}-\mathbf{x})^T\tilde{\mathbf{P}}^{-1}(\boldsymbol{\mu}-\mathbf{x})+\frac{1}{2}(\mathbf{y}-\mathbf{H}\mathbf{x})^T\tilde{\mathbf{R}}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x})$$

The minimization is typically done at *synoptic* hours, with a frequency of 6 hours and using observations available within a 6-hr window around the synoptic time.

In practice, an atmospheric prediction model is assumed to provide the mean state estimate $\boldsymbol{\mu}$, that is,

$$\boldsymbol{\mu} \equiv \mathbf{x}^b = \mathbf{m}(\mathbf{x}_0)$$

where $\mathbf{x}^b$ is the forecast (background) at a given time after evolving the model $\mathbf{m}$ forward in time, starting from an initial condition $\mathbf{x}_0$ representing the best estimate of the state of the atmosphere at a previous time.

To describe 3d-var, the time indexes are not so relevant and are dropped for simplification. Moreover, the mapping between observations and the estimate is nonlinear and a slightly more general cost function is actually used

$$J_{\mathsf{3dvar}}(\mathbf{x}) = \frac{1}{2}(\mathbf{x}^b-\mathbf{x})^T\tilde{\mathbf{P}}^{-1}(\mathbf{x}^b-\mathbf{x})+\frac{1}{2}[\mathbf{y}-\mathbf{h}(\mathbf{x})]^T\tilde{\mathbf{R}}^{-1}[\mathbf{y}-\mathbf{h}(\mathbf{x})]$$

where $\mathbf{h}(\mathbf{x})$ is the nonlinear observation function (operator).

Three-dimensional Variational Approach

To minimize this cost function using feasible computational methods, one needs to transform the cost function back to a quadratic function. This can be done by linearizing the observation operator $\mathbf{h}(\mathbf{x})$ around the background state, that is,

$$\mathbf{h}(\mathbf{x}) \approx \mathbf{h}(\mathbf{x}^b) + \mathbf{H}(\mathbf{x}^b)\delta\mathbf{x}$$

with $\delta\mathbf{x} \equiv \mathbf{x} - \mathbf{x}^b$ and $\mathbf{H}(\mathbf{x}^b)$ now denotes the Jacobian of the observation operator, $\mathbf{h}(\mathbf{x})$,

$$\mathbf{H}(\mathbf{x}^b) \equiv \left.\frac{\partial \mathbf{h}(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{x}^b}$$

Hence, we can right $\mathbf{y} - \mathbf{h}(\mathbf{x})$ as

$$\begin{aligned} \mathbf{y} - \mathbf{h}(\mathbf{x}) &= \mathbf{y} - \mathbf{h}(\mathbf{x}^b) - \mathbf{h}(\mathbf{x}) + \mathbf{h}(\mathbf{x}^b) \\ &= \mathbf{d} - \mathbf{H}(\mathbf{x}^b)\delta\mathbf{x} \end{aligned}$$

Using this first order expansion of the observation operator the cost function becomes quadratic form again

$$J_{\text{3dvar}}(\delta\mathbf{x}) = \frac{1}{2}\delta\mathbf{x}^T\tilde{\mathbf{P}}^{-1}\delta\mathbf{x} + \frac{1}{2}[\mathbf{d} - \mathbf{H}(\mathbf{x}^b)\delta\mathbf{x}]^T\tilde{\mathbf{R}}^{-1}[\mathbf{d} - \mathbf{H}(\mathbf{x}^b)\delta\mathbf{x}]$$

and it defines the so-called incremental 3d-var problem, since the cost is now written as a function of the increment vector $\delta\mathbf{x}$.

By inspection of our "estimation of a constant" exercise we see that minimization of the incremental 3d-var problem leads to the solution

$$\delta\mathbf{x}^a = \tilde{\mathbf{P}}^a\mathbf{H}^T\tilde{\mathbf{R}}^{-1}\mathbf{d}$$

with $\tilde{\mathbf{P}}^a = (\tilde{\mathbf{P}}^{-1} + \mathbf{H}^T\tilde{\mathbf{R}}^{-1}\mathbf{H})^{-1}$.

Three-dimensional Variational Approach

Remarks

- The 3d-var solution provides a LSP solution to the problem given the uncertainties in the background and observation error covariances $\tilde{\mathbf{P}}$ and $\tilde{\mathbf{R}}$.

- Employing computational methods to minimize the cost function directly is referred to as the 3d-var approach; whereas calculating the estimate from the analytical solution has become known as the PSAS approach, for the Physical-space Statistical Analysis System.

- In the analytical (PSAS) approach one avoids the n dimensional matrix inversion, by solving an algebraically equivalent equation (Ex. 7):

  $$\delta\mathbf{x}^a = \tilde{\mathbf{P}}\mathbf{H}^T(\mathbf{H}\tilde{\mathbf{P}}\mathbf{H}^T + \tilde{\mathbf{R}})^{-1}\mathbf{d}$$

  which is known as the PSAS equation, and it involves the inversion of an $m < n$ dimensional matrix.

- In practice, even this observation-space inversion is not directly calculated. Instead, the equation above is split in two stages:

  $$\begin{aligned}(\mathbf{H}\tilde{\mathbf{P}}\mathbf{H}^T + \tilde{\mathbf{R}})\boldsymbol{\lambda} &= \mathbf{d}\\ \delta\mathbf{x}^a &= \tilde{\mathbf{P}}\mathbf{H}^T\boldsymbol{\lambda}\end{aligned}$$

  where the first equation is solved using an iterative method, such as a conjugate gradient method. Because of the size of these matrices, they are all handled as operators, meaning, the are not actual matrices but are function calls simulating the application of a matrix on to a vector.

Three-dimensional Variational Approach

Remarks (cont.)

▷ The interplay between the 3d-var and PSAS approaches is a statement of the fact that these approaches are dual of each other. This essential means that one can be converted in to the other and their solutions are equivalent (Ex. 8).

▷ But don't get confused. Addressing the problem from the analytical solution has nothing to do with the wording "physical-space" as in PSAS. Solving the problem from the analytical solution is detached from the way the background error covariance is formulated.

▷ The *a priori* (background) error covariance is a parameterized quantity based on assumptions related to balance relationships and possible structure of errors. Traditional implementations of the direct minimization 3d-var approach (e.g., NCEP's SSI) have modeled background error covariances in spectral space. Difficulty in relaxing the assumptions behind these spectral space formulations has driven the reformulation of the covariances so they operate in physical-space. Modern 3d-var systems now minimize the cost function directly, and formulate the covariance in physical space (e.g., the Grid-space Statistical Interpolation approach)

Three-dimensional Variational Approach

Remarks (cont.)

- As described here, 3d-var operates at a single time, that is, the solution of the minimization problem is sought at a given time. However, the observation vector $\mathbf{y}$ jams together observations from a 6-hr time interval. This means in particular that calculation of the residual vector $\mathbf{d} \equiv \mathbf{y} - \mathbf{h}(\mathbf{x})$ is not accurate since $\mathbf{x}$ is taken at the time of the solution (analysis).

- Work done at operational centers has demonstrated that an improvement in the solution of the problem can be obtained when using an approach called FGAT: first guess at appropriate time. In this approach the function $\mathbf{h}$ is augmented to accommodate backgrounds (first-guesses) at various times within the window of observations. Typically, in 3d-var systems, FGAT means taking $\mathbf{x}$ at -3, 0, and 3 hrs from the synoptic hour; or sometimes taking them on an hourly basis. In these cases, the function $\mathbf{h}(\mathbf{x})$ also accommodates a time interpolation procedure to calculate the $\mathbf{d}$ vectors at exactly the time of the observations.

Four-dimensional Variational Approach

The FGAT approach is a simple attempt to address the lack of a time dimension in 3d-var. The proper way to account for the time dimension is to redefine the cost function:

$$2J_{4dvar} = ||\mathbf{x}-\mathbf{x}_0||_{\mathbf{B}^{-1}}+\sum_{i=0}^{I}||\mathbf{y}_i-\mathbf{h}(\mathbf{x}_i)||_{\mathbf{R}_i^{-1}}+\sum_{i=1}^{I}||\mathbf{x}_i-\mathbf{m}(\mathbf{x}_{i-1})||_{\mathbf{Q}_i^{-1}}$$

where $||\mathbf{x}||_{\mathbf{A}} \equiv \mathbf{x}^T\mathbf{A}\mathbf{x}$, for an arbitrary n-vector $\mathbf{x}$ and an arbitrary $n \times n$-matrix $\mathbf{A}$.

The cost function above applies to a discrete time interval with a total of I time slots. The first term accommodates the uncertainty in the initial condition with the matrix $\mathbf{B}$ being the error covariance associated with this uncertainty; the second term accommodates the uncertainties in the states $\mathbf{x}_i$ with respect to the observations at all times t_i in the interval, weighted by the observation error covariances $\mathbf{R}_i$; and the last term accommodates for uncertainties in the states themselves, weighted by the model error covariances $\mathbf{Q}_i$. This last term takes care of the fact that the prediction model is assumed to be imperfect:

$$\mathbf{x}_i = \mathbf{m}(\mathbf{x}_{i-1}) + \mathbf{q}_i$$

with the sequence of $\mathbf{q}_i$ vectors assumed to be white in time and normal with mean zero and covariance $\mathbf{Q}_i$, i.e., $\mathbf{q}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_i)$.

Four-dimensional Variational Approach

Using the incremental approach we can re-write the cost function as

$$2J_{4dvar} = ||\delta\mathbf{x}_0||_{\mathbf{B}^{-1}} + \sum_{i=0}^{I} ||\mathbf{d}_i - \mathbf{H}_i\delta\mathbf{x}_i||_{\mathbf{R}_i^{-1}} + \sum_{i=1}^{I} ||\mathbf{q}_i||_{\mathbf{Q}_i^{-1}}$$

where here again, $\mathbf{H}_i$ is the Jacobian of $\mathbf{h}$. This transforms the dependence on the cost function from $J_{4dvar} = J_{4dvar}(\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_I)$ to $J_{4dvar} = J_{4dvar}(\delta\mathbf{x}_0, \mathbf{q}_1, \cdots, \mathbf{q}_I)$.

The simplest way to understand how 4d-var basically amounts to a gigantic LSP is by re-writing further the cost function based on the following augmented vectors: $\delta\mathbf{x} \equiv \left[\delta\mathbf{x}_0^T\mathbf{q}_1^T \cdots \mathbf{q}_I^T\right]^T$ and $\mathbf{d} \equiv \left[\mathbf{d}_0^T\mathbf{d}_1^T \cdots \mathbf{d}_I^T\right]^T$. Therefore (Ex. 9),

$$2J_{4dvar}(\delta\mathbf{x}) = \delta\mathbf{x}^T\mathbf{D}^{-1}\delta\mathbf{x} + (\mathbf{G}\delta\mathbf{x} - \mathbf{d})\mathbf{R}^{-1}(\mathbf{G}\delta\mathbf{x} - \mathbf{d})$$

where the *a priori* error covariance matrix becomes $\mathbf{D} \equiv diag(\mathbf{B}, \mathbf{Q}_1, \cdots, \mathbf{Q}_N)$, the observations error covariance becomes $\mathbf{R} \equiv diag(\mathbf{R}_1, \mathbf{R}_2, \cdots, \mathbf{R}_N)$ and the "observation" matrix becomes

$$\mathbf{G} \equiv \begin{pmatrix} \mathbf{H}_0 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{H}_1\mathbf{M}_{1,0} & \mathbf{H}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{H}_2\mathbf{M}_{2,0} & \mathbf{H}_2\mathbf{M}_{2,1} & \mathbf{H}_2 & \mathbf{0} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{H}_I\mathbf{M}_{I,0} & \mathbf{H}_I\mathbf{M}_{I,1} & \mathbf{H}_I\mathbf{M}_{I,2} & \cdots & \mathbf{H}_I \end{pmatrix}$$

where $\mathbf{M}_{i,i-1}$ is the Jacobian of the forward model

$$\mathbf{M}_{i,i-1}(\mathbf{x}_{i-1}^b) \equiv \left.\frac{\partial\mathbf{m}(\mathbf{x}_{i-1})}{\partial\mathbf{x}_{i-1}}\right|_{\mathbf{x}_{i-1}=\mathbf{x}_{i-1}^b}$$

is now part of the observation matrix.

Four-dimensional Variational Approach

Formally, we can infer the solution of the minimization of this gigantic cost function by referring back to our "estimation of a constant" exercise, i.e., at the minimum the solution is give by

$$\delta\mathbf{x}^a = (\mathbf{D}^{-1} + \mathbf{G}^T\mathbf{R}^{-1}\mathbf{G})^{-1}\mathbf{G}^T\mathbf{R}^{-1}\mathbf{d}$$

Similarly to 3dvar, when the solution to 4d-var is being sought by directly minimizing the cost function we need its gradient to be available

$$\nabla_{\delta\mathbf{x}} J = \mathbf{D}^{-1}\delta\mathbf{x} + \mathbf{G}^T\mathbf{R}^{-1}(\mathbf{G}\delta\mathbf{x} - \mathbf{d})$$

since practical minimization algorithms are gradient-based, e.g., the conjugate gradient method.

Alternatively, we can use the algebraically equivalent expression

$$\delta\mathbf{x}^a = \mathbf{D}\mathbf{G}^T(\mathbf{G}\mathbf{D}\mathbf{G}^T + \mathbf{R})^{-1}\mathbf{d}$$

which is analogous to the PSAS equation, but since it now involves the fourth dimension of time it is known here as the 4d-PSAS equation. Just as in the 3d case, a practical approach to solve the 4d-PSAS equation splits the equation in two steps:

$$\begin{aligned}(\mathbf{G}\mathbf{D}\mathbf{G}^T + \mathbf{R})\boldsymbol{\lambda} &= \mathbf{d}\\ \delta\mathbf{x}^a &= \mathbf{D}\mathbf{G}^T\boldsymbol{\lambda}\end{aligned}$$

where here the vectors $\delta\mathbf{x}^a$, $\boldsymbol{\lambda}$, and $\mathbf{d}$ are all four-dimensional.

Four-dimensional Variational Approach

Remarks

- ▷ To solve the first 4D-PSAS equation we must have a smart way of applying the gigantic matrix on the left-hand-side to the vector $\boldsymbol{\lambda}$. The main complication in this operation comes from having to calculate $\mathbf{GDG}^T\boldsymbol{\lambda}$. To do so, we can notice that an element j of this term is given by (Ex. 10)

$$
\begin{aligned}
(\mathbf{GDG}^T\boldsymbol{\lambda})_j &= \mathbf{H}_j\mathbf{M}_{j,0}\mathbf{B}\sum_{i=1}^{I}\mathbf{M}_{i,0}^T\mathbf{H}_i^T\boldsymbol{\lambda}_i \\
&+ \mathbf{H}_j\sum_{m=1}^{j}\mathbf{M}_{j,m}\mathbf{Q}_m\sum_{i=m}^{I}\mathbf{M}_{i,m}^T\mathbf{H}_i^T\boldsymbol{\lambda}_i
\end{aligned}
$$

  These calculations can be broken down in to a backward integration of the equation

$$\mathbf{f}_i = \mathbf{M}_{i+1,i}^T\mathbf{f}_{i+1} + \mathbf{H}_i^T\boldsymbol{\lambda}_i$$

  for $i = I-1, I-2, \cdots, 0$, with $\mathbf{f}_I \equiv \mathbf{H}_I^T\boldsymbol{\lambda}_I$; followed by a forward integration

$$\mathbf{g}_m = \mathbf{M}_{j,m-1}\mathbf{g}_{m-1} + \mathbf{Q}_m\mathbf{f}_m$$

  for $m = 1, 2, \cdots, j$, and with $\mathbf{g}_0 \equiv \mathbf{B}\mathbf{f}_0$. This sequence of operations is known as the sweeper method and specifically constitute the so called augmented representer approach to the practical solution to calculating the 4d-PSAS equation (Ex. 11).

- ▷ In the perfect model case, $\mathbf{Q} = \mathbf{0}$, the 4d-var and 4d-PSAS equations above dramatically simplify.

Probabilistic Approach to Filtering

Let us indicate by $\mathbf{Y}_k^o = \{\mathbf{y}_1^o, \cdots, \mathbf{y}_{k-1}^o, \mathbf{y}_k^o\}$, the set of all observations up to and including time t_k. Similarly, let us indicate by $\mathbf{X}_k^t = \{\mathbf{x}_1^t, \cdots, \mathbf{x}_{k-1}^t, \mathbf{x}_k^t\}$ the set of all true states of the underlying system up to time t_k.

Knowledge of the pdf of the true state over the entire time period given all observations over the same period would allow us to calculate an estimate of the trajectory of the system over the time period. Therefore, calculation of the following pdf

$$p(\mathbf{X}_k^t | \mathbf{Y}_k^o)$$

is desirable. But, before seeking a system trajectory estimate, let us seek an estimate of the state of the system only at time t_k. For that, the relevant pdf is

$$\begin{aligned}
p(\mathbf{x}_k^t | \mathbf{Y}_k^o) &= p(\mathbf{x}_k^t | \mathbf{y}_k^o, \mathbf{Y}_{k-1}^o) \\
&= \frac{p(\mathbf{x}_k^t, \mathbf{y}_k^o, \mathbf{Y}_{k-1}^o)}{p(\mathbf{y}_k^o, \mathbf{Y}_{k-1}^o)} \\
&= \frac{p(\mathbf{y}_k^o | \mathbf{x}_k^t, \mathbf{Y}_{k-1}^o) p(\mathbf{x}_k^t, \mathbf{Y}_{k-1}^o)}{p(\mathbf{y}_k^o, \mathbf{Y}_{k-1}^o)} \\
&= \frac{p(\mathbf{y}_k^o | \mathbf{x}_k^t, \mathbf{Y}_{k-1}^o) p(\mathbf{x}_k^t | \mathbf{Y}_{k-1}^o) p(\mathbf{Y}_{k-1}^o)}{p(\mathbf{y}_k^o | \mathbf{Y}_{k-1}^o) p(\mathbf{Y}_{k-1}^o)} \\
&= \frac{p(\mathbf{y}_k^o | \mathbf{x}_k^t, \mathbf{Y}_{k-1}^o) p(\mathbf{x}_k^t | \mathbf{Y}_{k-1}^o)}{p(\mathbf{y}_k^o | \mathbf{Y}_{k-1}^o)} .
\end{aligned}$$

This relates the transition probability of interest with pdf's that can be calculated more promptly.

Probabilistic Approach to Filtering

Whiteness of the observation sequence allows us to write

$$p(\mathbf{y}_k^o|\mathbf{x}_k^t, \mathbf{Y}_{k-1}^o) = p(\mathbf{y}_k^o|\mathbf{x}_k^t)$$

and therefore,

$$p(\mathbf{x}_k^t|\mathbf{Y}_k^o) = \frac{p(\mathbf{y}_k^o|\mathbf{x}_k^t)p(\mathbf{x}_k^t|\mathbf{Y}_{k-1}^o)}{p(\mathbf{y}_k^o|\mathbf{Y}_{k-1}^o)}$$

It remains for us to determine each one of the transition probability densities in this expression.

Assumption: all pdf's (processes) are Gaussian and the observation process is linear, that is, $\mathbf{y}_k^o = \mathbf{H}_k\mathbf{x}_k^t + \mathbf{b}_k^o$, with $\mathbf{b}_k^o \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_k)$.

In this case, an immediate relationship between the variables above and those from the example of estimating a constant vector can be drawn:

- $\mathbf{y} \rightarrow \mathbf{y}_k^o$
- $\mathbf{x} \rightarrow \mathbf{x}_k^t$
- $p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) \rightarrow p(\mathbf{y}_k^o|\mathbf{x}_k^t)$
- $p_{\mathbf{x}}(\mathbf{x}) \rightarrow p(\mathbf{x}_k^t|\mathbf{Y}_{k-1}^o)$
- $p_{\mathbf{y}}(\mathbf{y}) \rightarrow p(\mathbf{y}_k^o|\mathbf{Y}_{k-1}^o)$

Probabilistic Approach to Filtering

Consequently we have

$$\begin{aligned} p(\mathbf{y}_k^o|\mathbf{x}_k^t) &= \frac{1}{(2\pi)^{m_k/2}|\mathbf{R}_k|^{1/2}} \\ &\quad \mathsf{exp}\left[-\frac{1}{2}(\mathbf{y}_k^o - \mathbf{H}_k\mathbf{x}_k^t)^T\mathbf{R}_k^{-1}(\mathbf{y}_k^o - \mathbf{H}_k\mathbf{x}_k^t)\right] \end{aligned}$$

where we noticed that

$$\mathcal{E}\{\mathbf{y}_k^o|\mathbf{x}_k^t\} = \mathcal{E}\{(\mathbf{H}_k\mathbf{x}_k^t + \mathbf{b}_k^o)|\mathbf{x}_k^t\} = \mathbf{H}_k\mathbf{x}_k^t$$

and

$$\begin{aligned} cov\{\mathbf{y}_k^o, \mathbf{y}_k^o|\mathbf{x}_k^t\} &\equiv \mathcal{E}\{[\mathbf{y}_k^o - \mathcal{E}\{\mathbf{y}_k^o|\mathbf{x}_k^t\}][\mathbf{y}_k^o - \mathcal{E}\{\mathbf{y}_k^o|\mathbf{x}_k^t\}]^T|\mathbf{x}_k^t\} \\ &= \mathbf{R}_k \end{aligned}$$

Analogously, we have

$$\begin{aligned} p(\mathbf{y}_k^o|\mathbf{Y}_{k-1}^o) &= \frac{1}{(2\pi)^{m_k/2}|\mathbf{\Gamma}_k|^{1/2}} \\ &\quad \mathsf{exp}\left[-\frac{1}{2}(\mathbf{y}_k^o - \mathbf{H}_k\mathbf{x}_{k|k-1}^f)^T\mathbf{\Gamma}_k^{-1}(\mathbf{y}_k^o - \mathbf{H}_k\mathbf{x}_{k|k-1}^f)\right] \end{aligned}$$

where we define $\mathbf{x}_{k|k-1}^f$ and the $m_k \times m_k$ matrix $\mathbf{\Gamma}_k$ as

$$\mathbf{x}_{k|k-1}^f \equiv \mathcal{E}\{\mathbf{x}_k^t|\mathbf{Y}_{k-1}^o\}\,, \quad \mathbf{\Gamma}_k \equiv \mathbf{H}_k\mathbf{P}_k^f\mathbf{H}_k^T + \mathbf{R}_k$$

with the $n \times n$ matrix $\mathbf{P}_k^f$ defined as

$$\mathbf{P}_{k|k-1}^f \equiv \mathcal{E}\{[\mathbf{x}_k^t - \mathbf{x}_k^f][\mathbf{x}_k^t - \mathbf{x}_k^f]^T|\mathbf{Y}_{k-1}^o\}$$

Probabilistic Approach to Filtering

To fully determine the *a posteriori* conditional pdf $p(\mathbf{x}_k^t|\mathbf{Y}_k^o)$, it remains to find the *a priori* conditional pdf $p(\mathbf{x}_k^t|\mathbf{Y}_{k-1}^o)$. Since we assumed all pdf's to be Gaussian, the from the definitions of $\mathbf{x}_k^f$ and $\mathbf{P}_k^f$ above we have $p(\mathbf{x}_k^t|\mathbf{Y}_{k-1}^o) \sim \mathcal{N}(\mathbf{x}_{k|k-1}^f, \mathbf{P}_{k|k-1})$, that is,

$$\begin{aligned} p(\mathbf{x}_k^t|\mathbf{Y}_{k-1}^o) &= \frac{1}{(2\pi)^{n/2}|\mathbf{P}_k^f|^{1/2}} \\ &\quad \exp\left[-\frac{1}{2}(\mathbf{x}_k^t - \mathbf{x}_{k|k-1}^f)^T(\mathbf{P}_{k|k-1}^f)^{-1}(\mathbf{x}_k^t - \mathbf{x}_{k|k-1}^f)\right] \end{aligned}$$

and the conditional pdf of interest can be written as

$$p(\mathbf{x}_k^t|\mathbf{Y}_k^o) = \frac{1}{(2\pi)^{n/2}|\mathbf{P}_{k|k}^a|^{1/2}} \exp\left(-\frac{1}{2}J\right)$$

where

$$J = (\mathbf{x}_{k|k}^a - \mathbf{x}_k^t)^T(\mathbf{P}_{k|k}^a)^{-1}(\mathbf{x}_{k|k}^a - \mathbf{x}_k^t)$$

is the cost function, with $\mathbf{x}_{k|k}^a$ minimizing it.

We can now identify the quantities $\hat{\mathbf{x}}_{\mathsf{MV}}$ and $\mathbf{P}_{\tilde{\mathbf{x}}}$ of the problem of estimating a constant vector with $\mathbf{x}_k^a$ and $\mathbf{P}_k^a$, respectively. Consequently, it follows from this correspondence that

$$\begin{aligned} \mathbf{x}_{k|k}^a &= \mathbf{x}_{k|k-1}^f + \mathbf{P}_{k|k-1}^f\mathbf{H}_k^T\mathbf{\Gamma}_k^{-1}(\mathbf{y}_k^o - \mathbf{H}_k\mathbf{x}_{k|k-1}^f) \\ (\mathbf{P}_{k|k}^a)^{-1} &= (\mathbf{P}_{k|k-1}^f)^{-1} + \mathbf{H}_k^T\mathbf{R}_k^{-1}\mathbf{H}_k \end{aligned}$$

Probabilistic Approach to Filtering

Remarks

- The estimate $\mathbf{x}^a_{k|k}$ maximizing the *a posteriori* pdf is the MAP estimate.

- Moreover, since the resulting *a posteriori* pdf is Gaussian, this estimate is also the conditional mean, that is,

$$\mathbf{x}^a_{k|k} \equiv \mathcal{E}\{\mathbf{x}^t_k | \mathbf{Y}^o_k\} \,,$$

  and therefore it is the MV estimate which is what the Kalman filter obtains.

- Similar results can be obtained by minimizing the cost function

$$J_{\text{3dVar}}(\delta\mathbf{x}_k) \equiv \delta\mathbf{x}_k^T (\mathbf{P}^f_{k|k-1})^{-1} \delta\mathbf{x}_k + (\mathbf{d}_k - \mathbf{H}_k \delta\mathbf{x}_k)^T \mathbf{R}_k^{-1} (\mathbf{d}_k - \mathbf{H}_k \delta\mathbf{x}_k)$$

  where $\delta\mathbf{x}_k \equiv \mathbf{x}^t_k - \mathbf{x}^f_{k|k-1}$, and $\mathbf{d}_k \equiv \mathbf{y}^o_k - \mathbf{H}_k \mathbf{x}^f_{k|k-1}$. In the meteorological literature $J_{\text{3dVar}}(\delta\mathbf{x}_k)$ is referred to as the incremental three-dimensional variational (3dvar) analysis cost function.

- Since in practice we have only rough estimates of the observations and forecast error covariance matrices $\mathbf{R}_k$ and $\mathbf{P}^f_{k|k-1}$, the minimization problem above solves none other than a LSP problem, given some prior information.

Probabilistic Approach to Filtering

Remarks (cont.)

- So far we have made no assumptions about the process $\mathbf{x}_k^t$ other than its conditional pdf $p(\mathbf{x}_k^t|\mathbf{X}_{k-1}^o)$ being Gaussian. However, if we want to be able to calculate an estimate of the state one time ahead, that is at t_{k+1}, using the knowledge gather up to time t_k we must consider the pdf

$$\begin{aligned} p(\mathbf{x}_{k+1}^t, \mathbf{x}_k^t|\mathbf{X}_k^o) &= p(\mathbf{x}_{k+1}^t|\mathbf{x}_k^t, \mathbf{X}_k^o)p(\mathbf{x}_k^t|\mathbf{X}_k^o) \\ &= p(\mathbf{x}_{k+1}^t|\mathbf{x}_k^t)p(\mathbf{x}_k^t|\mathbf{X}_k^o) \end{aligned}$$

  which refers to the yet unspecified transition pdf $p(\mathbf{x}_{k+1}^t|\mathbf{x}_k^t)$ and therefore we must know more about the process $\mathbf{x}_k^t$.

- When the process $\mathbf{x}_k^t$ is linear the calculations are simple. That is, the system

$$\mathbf{x}_{k+1}^t = \mathbf{M}_{k+1,k}\mathbf{x}_k^t + \mathbf{b}_{k+1}^t$$

  with $\mathbf{b}_{k+1}^t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_{k+1})$ results in a Gaussian transition pdf (for an initial Gaussian pdf $p(\mathbf{x}_0^t)$):

$$p(\mathbf{x}_{k+1}^t|\mathbf{x}_k^t) \sim \mathcal{N}(\mathbf{M}_{k+1,k}\mathbf{x}_k^t, \mathbf{Q}_{k+1}) \,.$$

- For linear dynamical process above it follows that

$$\begin{aligned} \mathbf{x}_{k+1}^f &= = \mathbf{M}_{k+1,k}\mathcal{E}\{\mathbf{x}_{k+1}^t|\mathbf{Y}_k^o\} + \mathcal{E}\{\mathbf{b}_{k+1}^t|\mathbf{Y}_k^o\} \\ &= \mathbf{M}_{k+1,k}\mathbf{x}_{k|k}^a \\ \mathbf{P}_{k+1|k}^f &= cov\{\mathbf{x}_{k+1}^t, \mathbf{x}_{k+1}^t|\mathbf{Y}_k^o\} \\ &= \mathbf{M}_{k+1,k}\mathbf{P}_{k|k}^a\mathbf{M}_{k+1,k}^T + \mathbf{Q}_{k+1} \end{aligned}$$

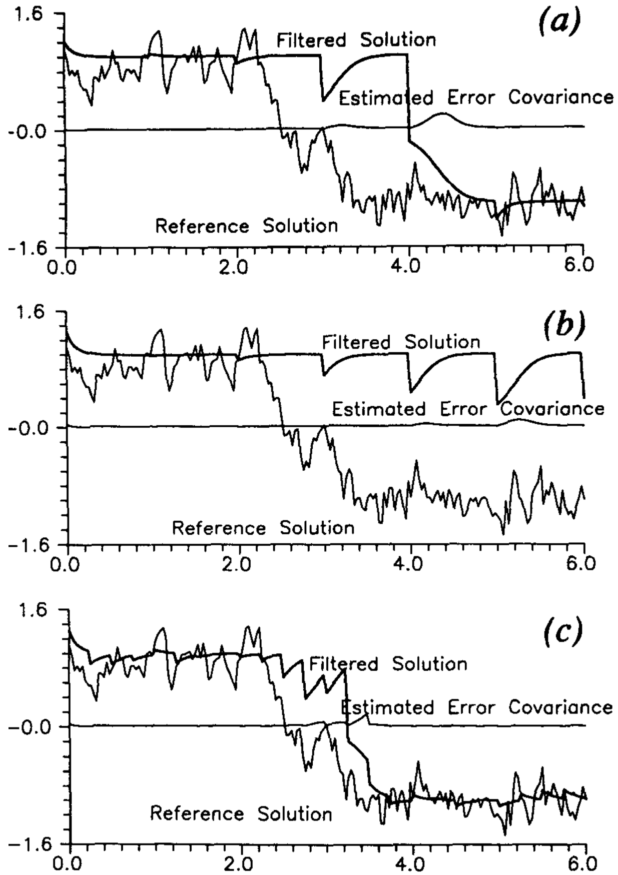
Simple Illustrations
&
Points to Remember

Kalman Filter for Highly Nonlinear Dynamics

Stochastically forced double-well potential

$$dx = f(x)dt + \sigma db$$

$$\dot{x} = f(x) \equiv -4x(x^2 - 1)$$



From Miller et al. (1994)

Robert N. Miller

Dynamical System: Lorenz (1963)

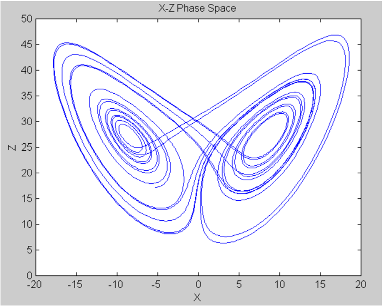
$$\begin{aligned}\dot{x} &= \sigma(y - x)\\ \dot{y} &= \rho x - y - xz\\ \dot{z} &= xy - \beta z\end{aligned}$$

Chaotic for the following parameters:

$$\sigma = 10 \quad \rho = 28 \quad \beta = 8/3$$

Unstable equilibrium points:

$$(0,0,0)$$

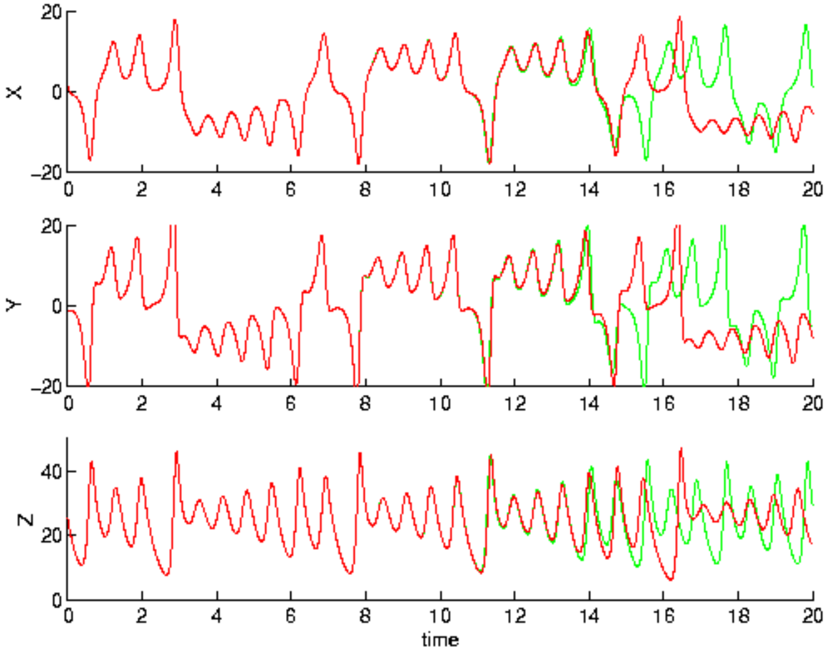


$$(\pm\sqrt{\beta(\rho-1)}, \pm\sqrt{\beta(\rho-1)}, -1)$$

Diverging Solutions from Highly Nonlinear Dynamics

What does a tiny initial perturbation do to prediction?

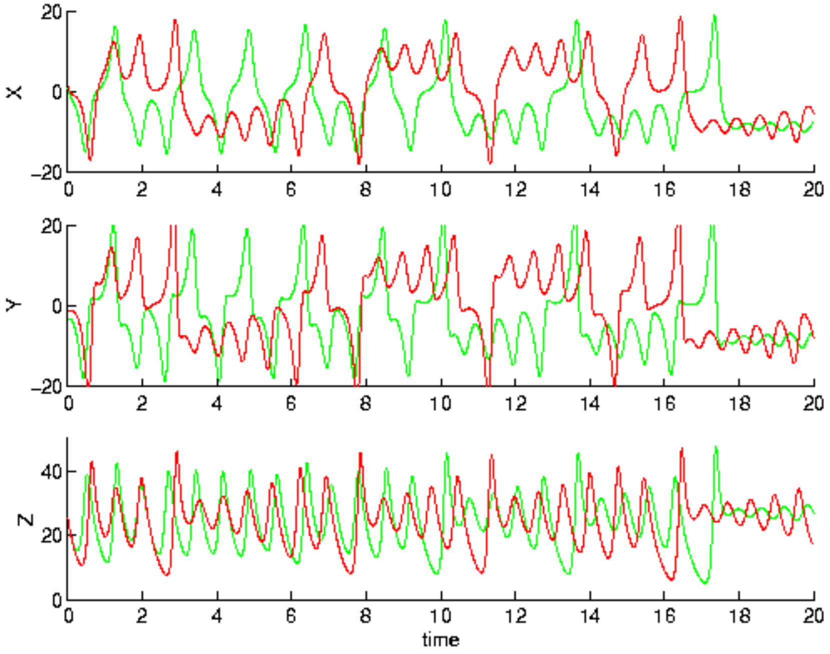
$\sigma(0) = 10^{-6}$



Answer: Cause some (chaotic) trouble!

What about a not-so-tiny initial perturbation?

$\sigma(0) = 1$



Answer: It causes a lot of trouble! The two runs started from initial conditions differing by about a few percent in magnitude. You can think of the red lines as being the true state evolution and the green lines as being the predicted state. In this case, the prediction becomes useless very quickly. The solution to this problem is to assimilate observations.

The Extended Kalman Filter for Highly Nonlinear Dynamics

Back to Miller et al (1994)

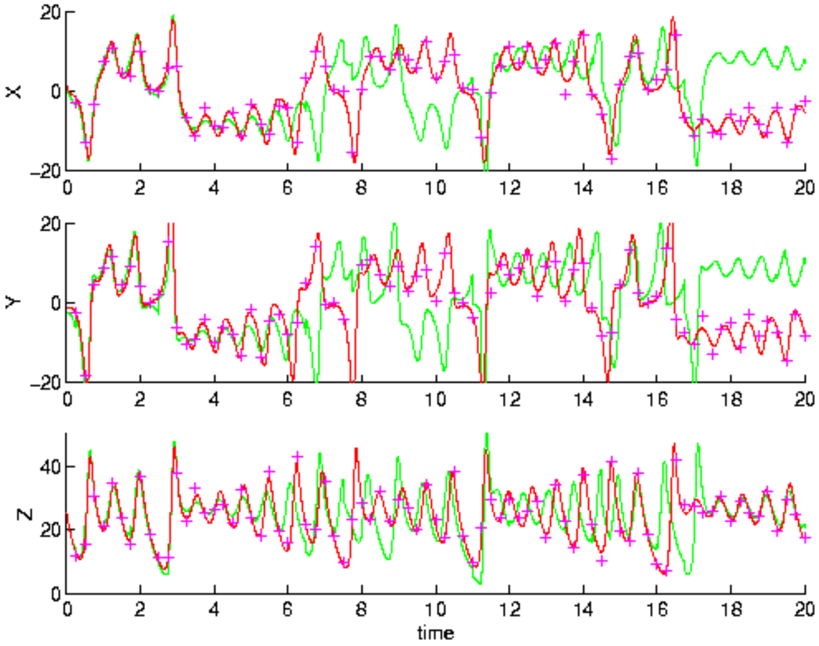
Then, what does data assimilation do?

σ(obs) = 2

Red: Truth

Green: Estimate

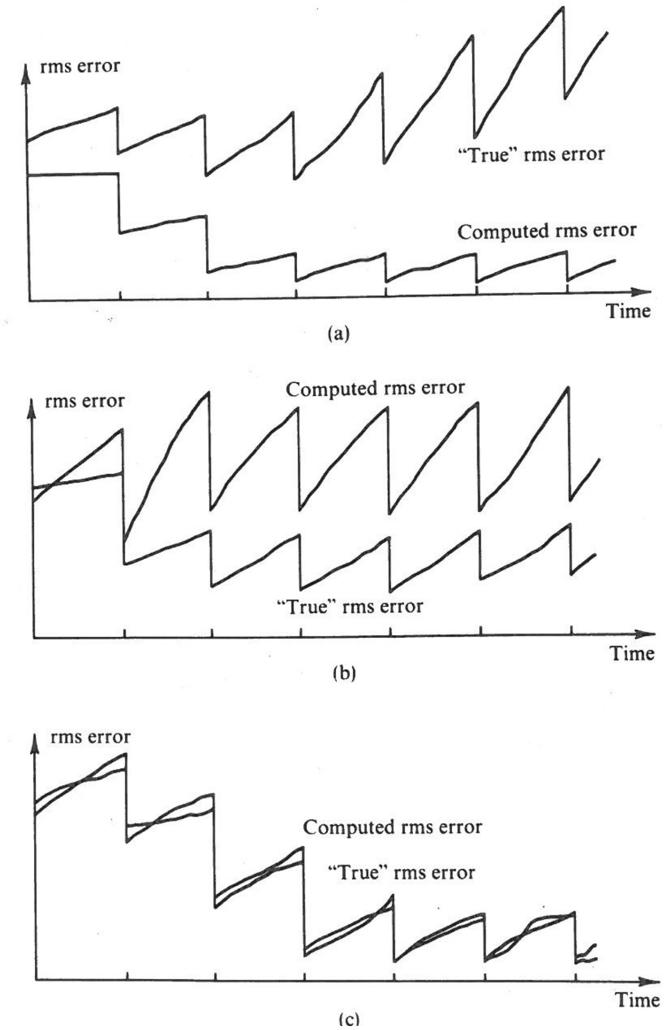
Pluses: Observations



Answer: It improves our ability to estimate the true state and make relatively reasonable short- to medium-range predictions. However, depending on the data assimilation scheme, the estimate may diverge after a while. The red line represents the true state while the green line represents the estimate (assimilation), the crosses are the observations; the data assimilation scheme is the extended Kalman filter (EKF).

Some Traps to Avoid

Filter error estimates are reliable indicators of performance!



From Maybeck (1981 also 1982)

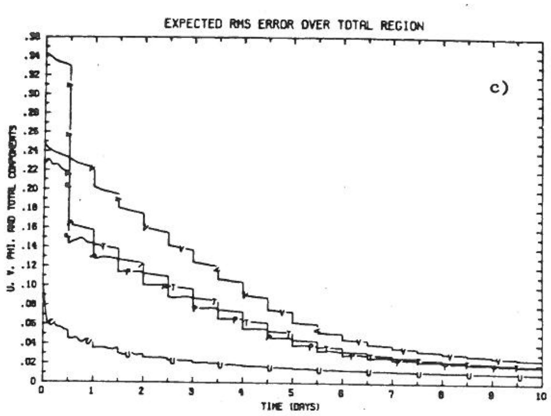
Lesson: *Ideally* the specified (computed) error covariances should be as close as possible to the true error covariance (bottom) – this is what we all aim when trying to tune the error statistics in our systems.

Under estimation of errors is rather undesirable as it is bound to lead to filter divergence (top).

In general, *slightly over estimation* of error keeps the filter from "believing" too much on its own estimates – thus preventing divergence (mid-plot shows an exaggerated version of this – that in this case still diverge).

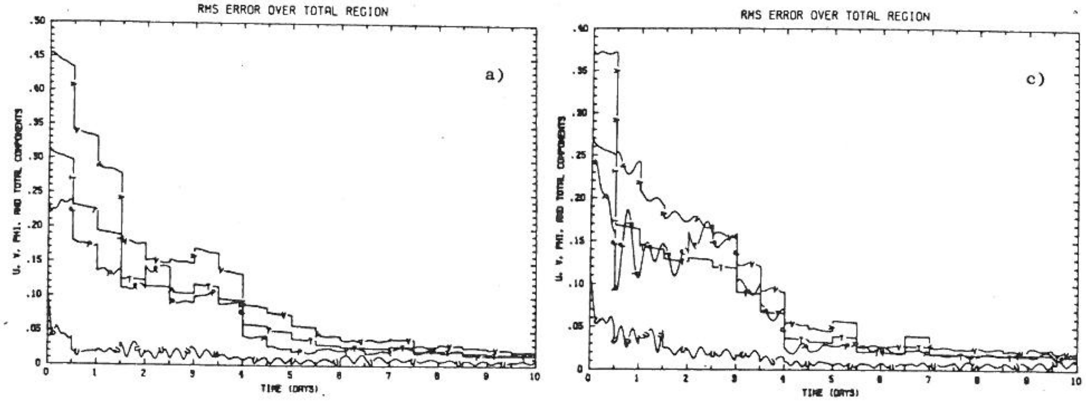
Some Traps to Avoid

Do observations always improve estimate through the analysis step?



Lesson: No. Only in the expected mean sense, *and* in the *optimal* (BLUE) circumstance this is the case. Recall also that in practice we only have a single realization of nature to work from.

Realization 1



Realization 2

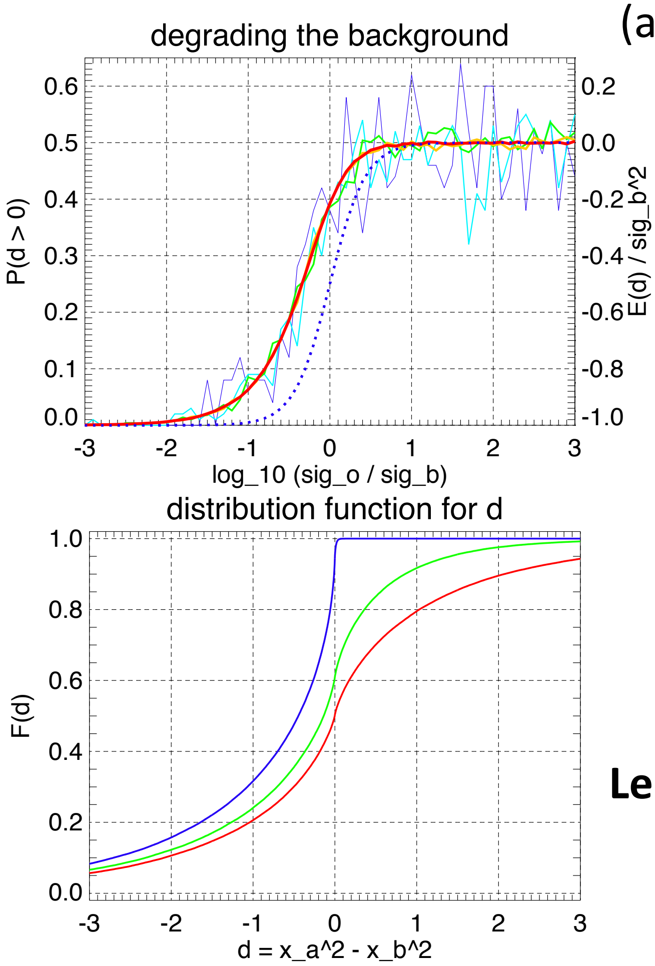
From Ghil et al. (1981)

BLUE: The linear Kalman filter is sometimes referred to as the (b)est (l)inear (u)nbiased (e)stimate – for linear problems under assumed error statistics.

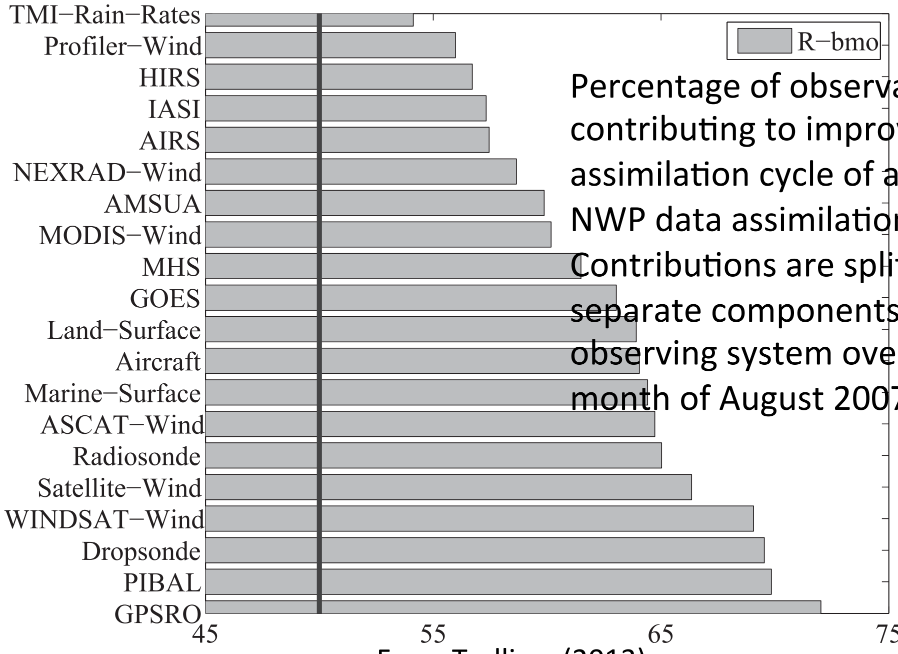
Some Traps to Avoid

Do observations always improve estimate through the analysis step?

(another illustration)



From Ehrendorfer (2007)



Percentage of observations contributing to improve the assimilation cycle of a real NWP data assimilation system. Contributions are split into separate components of observing system over the month of August 2007.

From Todling (2013)

Lesson: As long as there are uncertainty in the observations and in our models there will always be a considerable fraction of the data that will deteriorate our estimate. Attempts to eliminate observations that seem to offend the estimate can at best work locally.

Some Traps to Avoid

Time averaging provides good means of getting handle on statistics!

Stochastic harmonic oscillator system

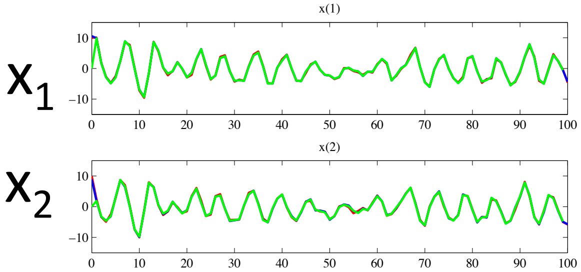
$$
\begin{aligned}
\mathbf{x}_k &= \begin{bmatrix} 0 & 1 \\ -1 & -0.8 \end{bmatrix} \mathbf{x}_{k-1} + \mathbf{q}_k \,, \\
\mathbf{y}^o_k &= \mathbf{H}\mathbf{x}_k + \boldsymbol{\epsilon}_k \,,
\end{aligned}
$$

$\mathbf{x}_0 = [10\ \ 10]^T$

$\mathbf{H}_{k|k-1} = \mathbf{I}$, with $\mathbf{R}_k = 0.1\mathbf{I}$.

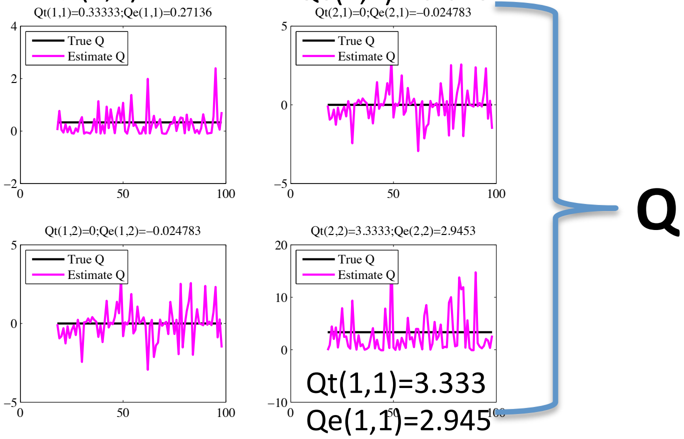
$\mathbf{Q} = diag(1/3, 10/3)$.

Single realization



Qt(1,1)=0.333
Qe(1,1)=0.271

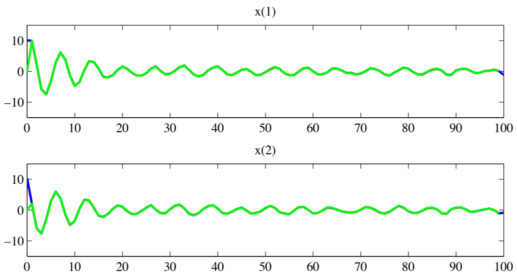
Qt(1,1)=0.
Qe(1,1)=-0.025



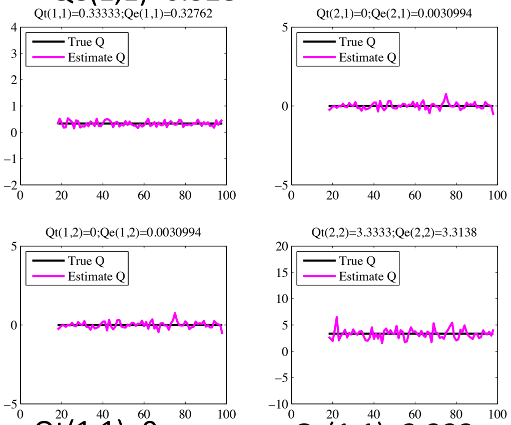
Qt(1,1)=3.333
Qe(1,1)=2.945

Lesson: Not necessarily. Time averaging filter statistics has the tendency to provide underestimates of variances, for example.

30-sample Monte Carlo



Qt(1,1)=0.333
Qe(1,1)=0.328



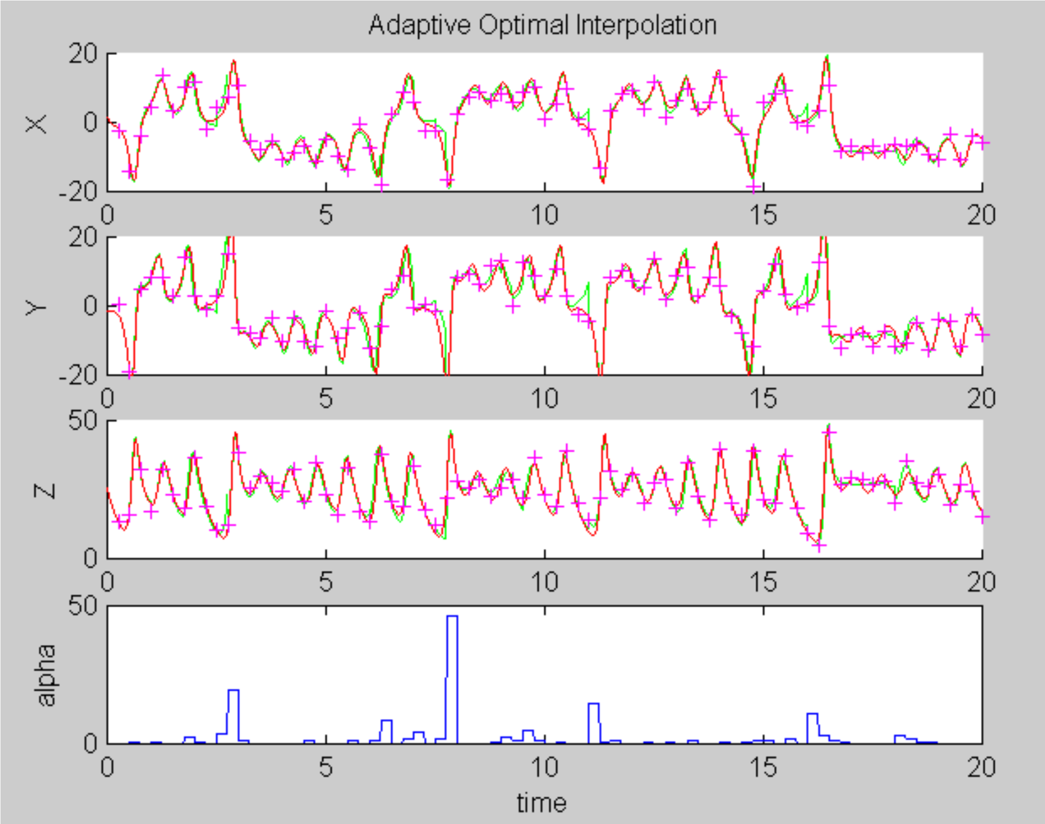
Qt(1,1)=0
Qe(1,1)=0.003

Qt(1,1)=3.333
Qe(1,1)=3.314

A twist on Maybeck (1982; Vol. 2, Ch. 8); see Todling (2014; QJRMS)

Something to Keep in Mind

Robust estimation can be achieved with adaptive procedures



The assimilation scheme here is an adaptive optimal interpolation. In this case, the propagated error covariance (the costly part of the EKF) is replaced by a constant forecast error covariance matrix scaled by a single parameter that gets to be adaptively estimated on the basis of the observation-minus-forecast residuals (see Dee 1995). The time series of the estimated parameter is displayed in the lower panel above.

Closing Remarks

- Solid understand of the three estimates (MV, MAP, ML) examined here gives a broad perspective on estimation problems.

- Most methods employed in practice fall under the LS-type category.

- Adaptive procedures are typically the most robust – viz. modern hybrid ensemble-variational approaches.