Introduction to Data Assimilation

or alternatively

Introduction to Estimation Theory

Ricardo Todling

Global Modeling and Assimilation Office, NASA/GSFC

Applications of Remote Sensed Observations in Data Assimilation
JCSDA & University of Maryland, 2007

Contact: todling@gmao.gsfc.nasa.gov

Outline

1. Objectives
2. Concepts of probabilistic estimation
3. Example: Estimation of a constant vector
4. Three-dimensional variational assimilation
5. Four-dimensional variational assimilation
6. The probabilistic approach to filtering
7. The probabilistic approach to smoothing
8. Illustrations
9. Closing Remarks

Illustration 1: Data Assimilation for Chaotic Dynamics

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:

$$(0, 0, 0)$$

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

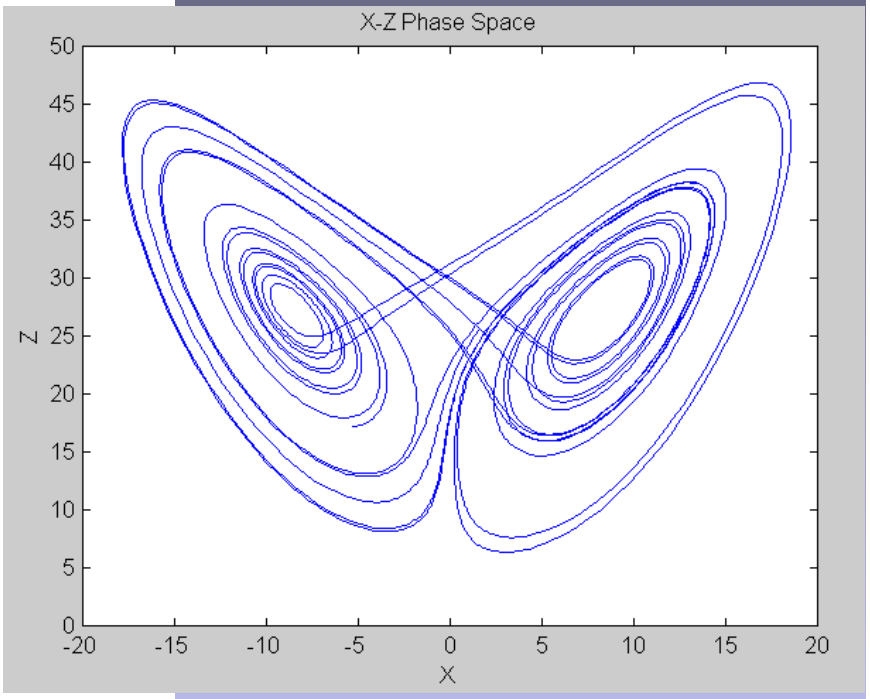
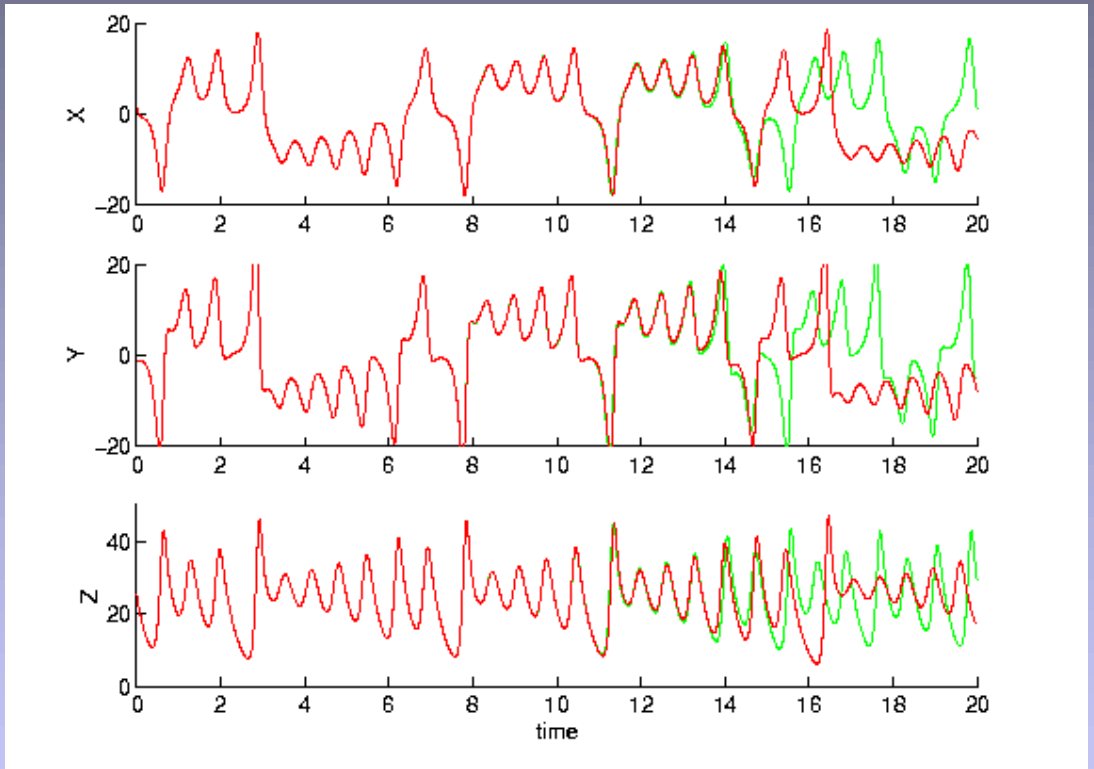


Illustration 1(cont.): Data Assimilation for Chaotic Dynamics

What does a tiny initial perturbation do to prediction?

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

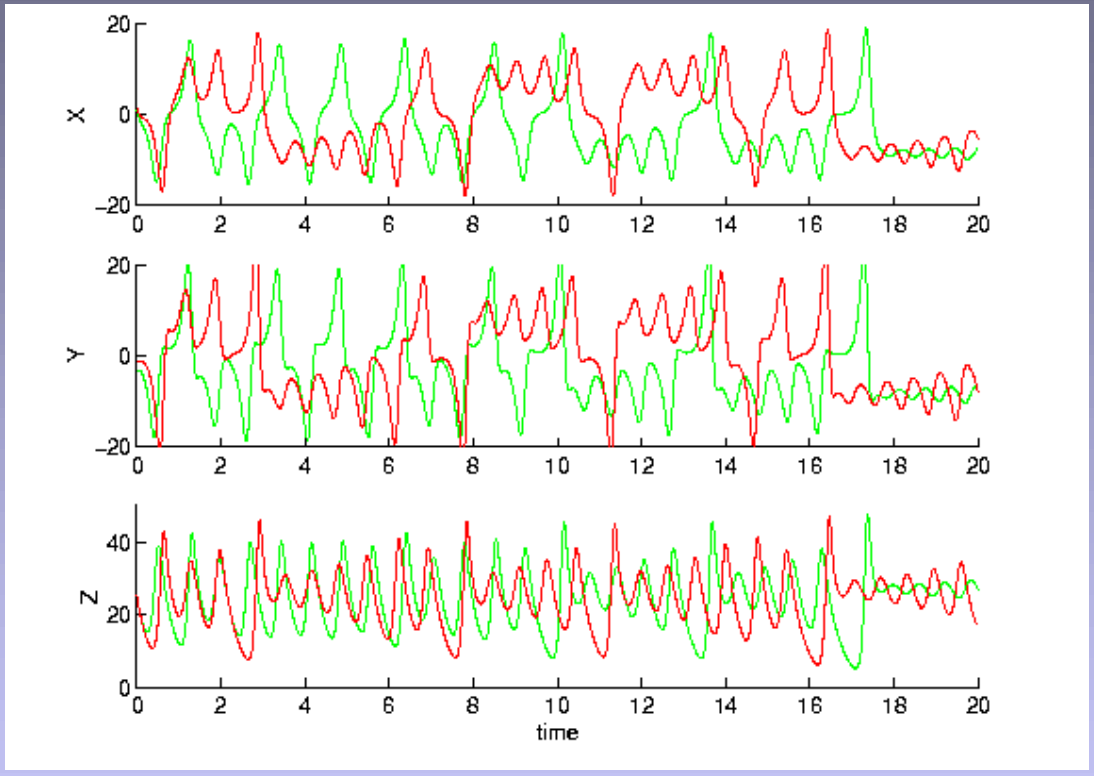


Answer: Cause some (chaotic) trouble!

Illustration 1(cont.): Data Assimilation for Chaotic Dynamics

What about a not-so-tiny initial perturbation?

$\sigma(0) = 1$

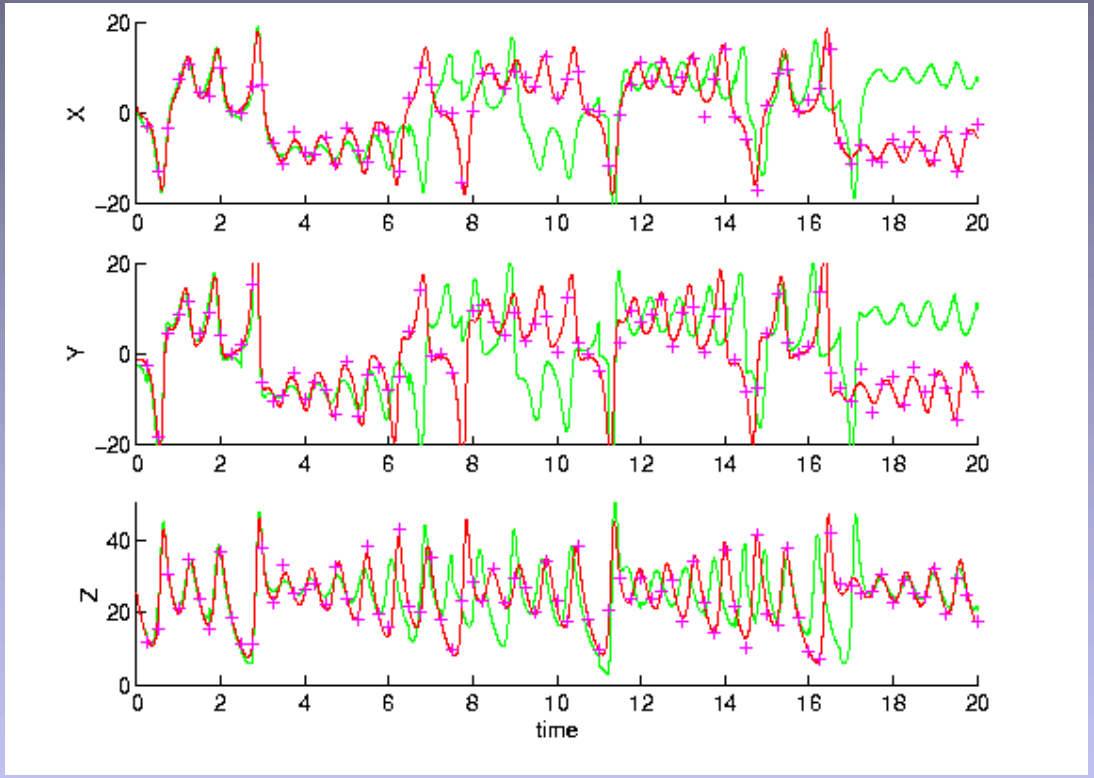


Answer: It causes a lot of trouble! The two runs started from initial conditions differing by about one percent in magnitude. You can think of the red line as being the true state evolution and the green line as being the predicted state. In this case, the prediction becomes useless very quickly. The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

Illustration 1(cont.): Data Assimilation for Chaotic Dynamics

Then, what does data assimilation do?

σ(obs) = 2



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).

1. Objectives

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 also be given to details:
  - off-line and on-line quality control
  - removal of both model and observation biases
  - proper usage of observations, that is, they should be used at right time, be given proper representativeness error characteristics
  - properly initialized fields
  - tangent linear and adjoint models issues
- Remember ... *adaptive procedures are robust*.

2. Concepts of Probabilistic Estimation

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

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

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

and Bayes rule for converting between conditional pdf's:

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

In the light of conditional pdf's we can define the conditional mean:

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

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

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

which is a n-dimensional Gaussian function.

2.1 Cost Function

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:

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

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
$p_{\mathbf{x}}(\mathbf{x})$	marginal pdf of $\mathbf{x}$
$p_{\mathbf{xy}}(\mathbf{x},\mathbf{y})$	joint pdf between $\mathbf{x}$ and $\mathbf{y}$

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

2.2 Two Examples of Cost Functions

(a) The quadratic cost:

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

(b) The uniform cost:

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

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:

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

2.3 Minimum Variance Estimation

In this case we use the quadratic cost function to get:

$$\mathcal{J}_{\mathrm{MV}}(\hat{\mathbf{x}}) = \frac{1}{2}\int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\infty}(\mathbf{x}-\hat{\mathbf{x}})^T\mathbf{E}(\mathbf{x}-\hat{\mathbf{x}})\,p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\,d\mathbf{x}\right\}p_{\mathbf{y}}(\mathbf{y})\,d\mathbf{y}$$

Or, identifying the kernel as the conditional Bayes cost:

$$\mathcal{J}_{\mathrm{MV}}(\hat{\mathbf{x}}|\mathbf{y}) \equiv \frac{1}{2}\int_{-\infty}^{\infty}(\mathbf{x}-\hat{\mathbf{x}})^T\mathbf{E}(\mathbf{x}-\hat{\mathbf{x}})\,p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\,d\mathbf{x}$$

Minimization of the cost $\mathcal{J}_{\mathrm{MV}}(\hat{\mathbf{x}}|\mathbf{y})$ gives

$$\begin{aligned}\mathbf{0} &= \left.\frac{\partial\mathcal{J}_{\mathrm{MV}}(\hat{\mathbf{x}}|\mathbf{y})}{\partial\hat{\mathbf{x}}}\right|_{\hat{\mathbf{x}}=\hat{\mathbf{x}}_{\mathrm{MV}}}\\ &= -\,\mathbf{E}\left.\int_{-\infty}^{\infty}(\mathbf{x}-\hat{\mathbf{x}})\,p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\,d\mathbf{x}\right|_{\hat{\mathbf{x}}=\hat{\mathbf{x}}_{\mathrm{MV}}}\end{aligned}$$

And noticing that p is a pdf, it follows that

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

Conclusion: the estimate with minimum variance is the conditional mean.

- this estimate is unbiased
- this estimate is indeed the minimum of the cost function (Ex. 1)

2.4 Maximum a posteriori Probability Estimation

Using now the uniform cost function we have

$$\mathcal{J}_U(\hat{\mathrm{x}}) = \int_{-\infty}^{\infty} \frac{1}{2\epsilon} \left\{ 1 - \int_{\hat{\mathrm{x}}-\epsilon}^{\hat{\mathrm{x}}+\epsilon} p_{\mathrm{x}|\mathrm{y}}(\mathrm{x}|\mathrm{y})\, d\mathrm{x} \right\} p_{\mathrm{y}}(\mathrm{y})\, d\mathrm{y}$$

To minimize $\mathcal{J}_U$ with respect to $\hat{\mathrm{x}}$, the first term gives no relevant contribution, thus

$$\mathcal{J}_U(\hat{\mathrm{x}}) \sim -(1/2\epsilon) \int_{-\infty}^{\infty} \left\{ \int_{\hat{\mathrm{x}}-\epsilon}^{\hat{\mathrm{x}}+\epsilon} p_{\mathrm{x}|\mathrm{y}}(\mathrm{x}|\mathrm{y})\, d\mathrm{x} \right\} p_{\mathrm{y}}(\mathrm{y})\, d\mathrm{y}\,.$$

or yet, we can minimize the conditional Bayes cost

$$\mathcal{J}_U(\hat{\mathrm{x}}|\mathrm{y}) \equiv -(1/2\epsilon) \int_{\hat{\mathrm{x}}-\epsilon}^{\hat{\mathrm{x}}+\epsilon} p_{\mathrm{x}|\mathrm{y}}(\mathrm{x}|\mathrm{y})\, d\mathrm{x}$$

As $\epsilon \to 0$, the mean value theorem gives

$$\mathcal{J}_U(\hat{\mathrm{x}}|\mathrm{y}) = -p_{\mathrm{x}|\mathrm{y}}(\hat{\mathrm{x}}|\mathrm{y})$$

Conclusion: The maximum a posteriori estimate is obtained by maximizing the conditional pdf, that is,

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

or yet

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

- this estimate is NOT guaranteed to be unbiased

2.5 Maximum Likelihood Estimation

In ML estimation we assume the *a priori* information is unknown. Suppose for the moment that the *a priori* pdf is $\mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}}, \mathbf{P}_{\mathbf{x}})$, then

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

Hence,

$$\frac{\partial \ln p_{\mathbf{x}}(\mathbf{x})}{\partial \mathbf{x}} = -\mathbf{P}_{\mathbf{x}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})$$

Since that lack of information implies infinite variance, $\mathbf{P}_{\mathbf{x}} \to \infty$, or yet $\mathbf{P}_{\mathbf{x}}^{-1} \to \mathbf{0}$, the maximum likelihood estimate of $\mathbf{x}$ can be obtained by

$$\begin{aligned} \mathbf{0} &= \left[\frac{\partial \ln p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})}{\partial \mathbf{x}} + \frac{\partial \ln p_{\mathbf{x}}(\mathbf{x})}{\partial \mathbf{x}}\right]\Bigg|_{\mathbf{x}=\hat{\mathbf{x}}_{\mathrm{MAP}}} \\ &= \frac{\partial \ln p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})}{\partial \mathbf{x}}\Bigg|_{\mathbf{x}=\hat{\mathbf{x}}_{\mathrm{ML}}} \end{aligned}$$

or equivalently,

$$\boxed{\frac{\partial p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})}{\partial \mathbf{x}}\Bigg|_{\mathbf{x}=\hat{\mathbf{x}}_{\mathrm{ML}}} = \mathbf{0}}$$

- $\hat{\mathbf{x}}_{\mathrm{ML}}$ can be referred to as the most likely estimate
- This estimate is NOT guaranteed to be unbiased.
- The estimate obtained this way is NOT Bayesian.

Quick Recap

Bayes rule for pdf's:

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

Conditional mean:

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

Minimum variance estimate:

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

Maximum *a posteriori* probability estimate:

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

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

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

Back to Lecture 1: Estimating T from Irradiance

Let us go back to Lecture 1 by Dr. Kalnay and use her example of estimating the temperature T of a stone via "measurements" of total irradiant energy. That is, consider the case where the measurement equation is

$$y = \sigma T^4 + \epsilon$$

Assumptions: ϵ and the error in the knowledge on T are statistically independent; ϵ is Gaussian distributed, $\mathcal{N}(0, \sigma_o)$; and the error in knowledge of T is also Gaussian distributed, $\mathcal{N}(T_b, \sigma_T)$.

To estimate T with the Bayesian approach we need to construct the conditional probability

$$\begin{aligned} p(T|y) &= \frac{p(y|T)p(T)}{p(y)} \\ &= \frac{exp[-(y-\sigma T^4)^2/2\sigma_o^2]exp[-(T_b-T)^2/2\sigma_T^2)]}{2\pi\sigma_o\sigma_T p(y)} \end{aligned}$$

To derive the MAP and ML estimates of T we need to minimize the exponent (notice $p(y)$ is not relevant for this task), that is,

$$\frac{d}{dT}\left[\frac{1}{2\sigma_o^2}(y-\sigma T^4)^2 + \frac{1}{2\sigma_T^2}(T_b - T)^2\right]\Big|_{T=\hat{T}_{\mathrm{MAP}}} = 0$$

The MAP estimate can be obtained by determining the roots of a 7th order polynomial:

$$\hat{T}^7_{\text{MAP}} - \frac{y}{\sigma}\hat{T}^3_{\text{MAP}} + \frac{\sigma_o^2}{4\sigma^2\sigma_T^2}\hat{T}_{\text{MAP}} - \frac{\sigma_o^2}{4\sigma^2\sigma_T^2}T_b = 0$$

As we have seen, the ML estimate can be obtained by assuming no knowledge of prior information, that is, which means taking $\sigma_T \to \infty$. In this case, the ML estimate can be obtained from the roots of the polynomial:

$$\hat{T}^4_{\text{ML}} - \frac{y}{\sigma} = 0$$

Obtaininig the minimum variance estimate requires knowledge of the full conditional pdf $p(T|y)$ and to solve a tough integral.

Instead, we recall the alternative solution presented in Lecture 1, where an estimate is sought by linearizing the problem about a background (prior estimate), say, in this case T_b. This leads to the following solution, referred to here as LMAP:

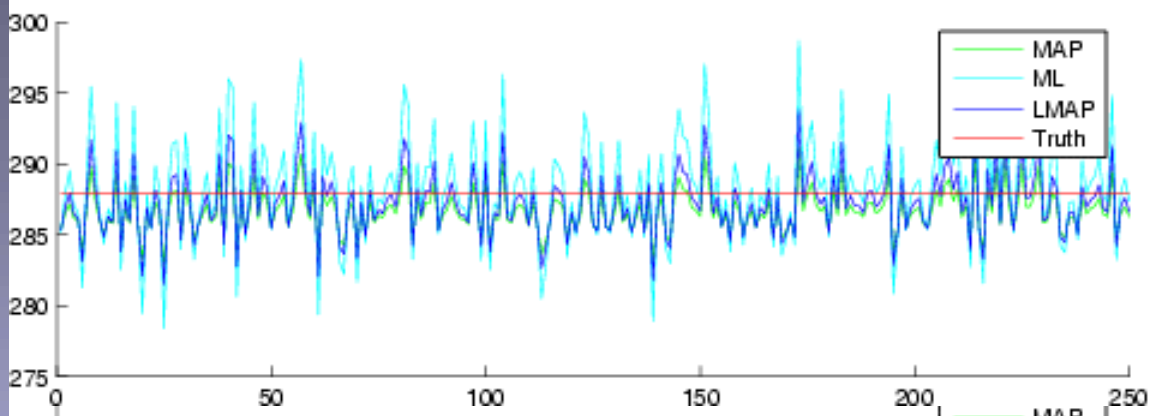
$$\hat{T}_{\text{LMAP}} = \left(\frac{4\sigma T_b^3}{\sigma_o^2} + \frac{1}{\sigma_T^2}\right)\frac{\sigma}{\sigma_o^2}(y - \sigma T_b^4)$$

Table 1: Parameters used to illustrate example of T estimation from irradiance

Parameters	Values
True temperature of stone (Earth surface)	288 K
Mean prior known temperature	288 K + error
Stefan-Boltzmann constant	5.6693×10^{-8} W m^{-2} K^{-4}
Irradiance observation error	5% of true irradiance

Bottom line: Be careful with the assumptions

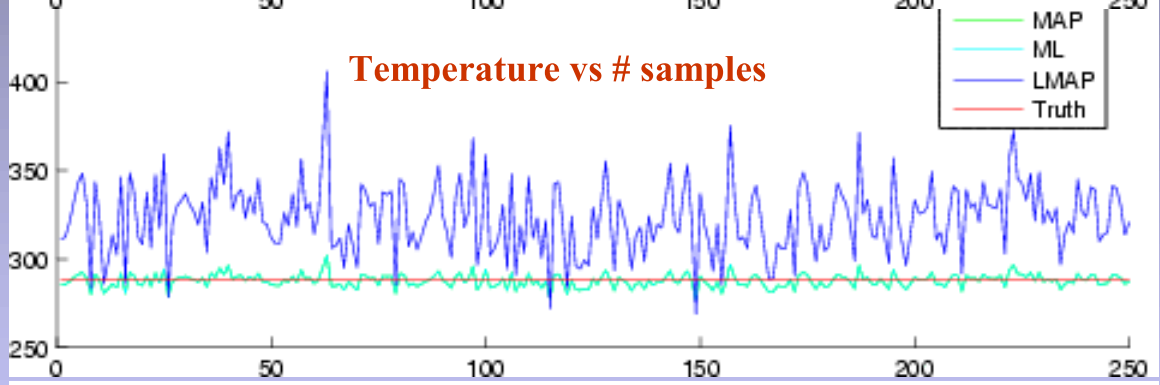
Somewhat Accurate Prior



All estimates seem reasonable

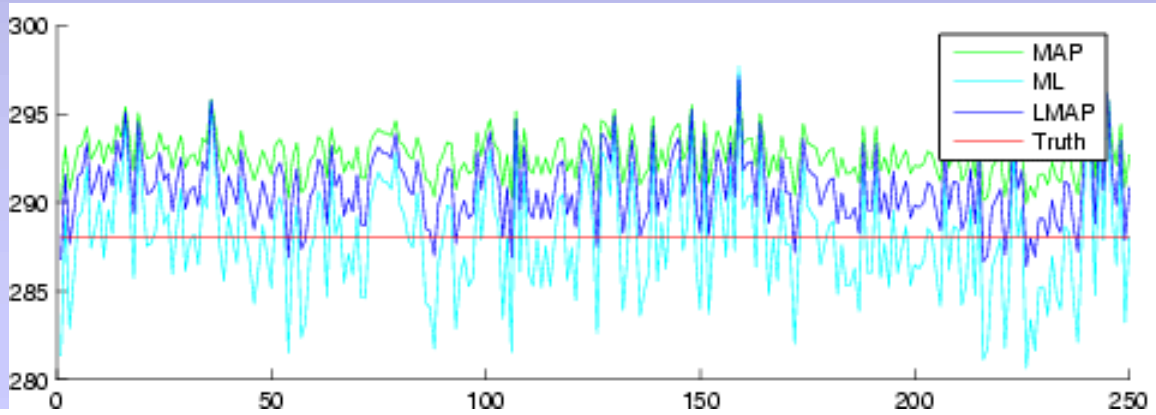
Very Inaccurate Prior

Temperature vs # samples



Linearized MAP is considerably biased.

Somewhat Inaccurate Prior



MAP is biased; Linearized MAP is also biased, but to a lesser extent.

3. Example: Estimation of a Constant Vector

Consider the time-constant observational process

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

where $\mathbf{x}$ is an n–vector, $\mathbf{y}$ and $\mathbf{b}^o$ are m–vectors, and $\mathbf{H}$ is an $m \times n$ matrix.

Assumptions: $\mathbf{x}$ and $\mathbf{b}^o$ are independent and Gaussian distributed, that is, $\mathbf{x} \sim \mathcal{N}(\mu, \mathbf{P})$, and $\mathbf{b}^o \sim \mathcal{N}(0, \mathbf{R})$.

Problem: What do the three estimates studied previously correspond to in this case?

For the MV estimate we need to determined the *a posteriori* pdf $p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})$ (Bayes rule):

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

consequently we need to determine each one of the pdf's in the expression above.

To begin with, we see that by definition

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

Now, to calculate the other pdf's recall that: linear transformations of Gaussian distributed variables result in Gaussian distributed variables (Ex. 2).

Hence,

$$p_y(y) = \frac{1}{(2\pi)^{m/2}|P_y|^{1/2}} \exp\left[-\frac{1}{2}(y-\mu_y)^T P_y^{-1}(y-\mu_y)\right]$$

where μ_y and P_y correspond to the mean and covariance of the random variable y, respectively.

Applying the ensemble average operator and using the definition of covariance:

$$\mu_y = \mathcal{E}\{Hx\} + \mathcal{E}\{b^o\} = H\mu$$

and also,

$$\begin{aligned} P_y &= \mathcal{E}\{(y-\mu_y)(y-\mu_y)^T\} \\ &= \mathcal{E}\{[(Hx+b^o)-H\mu)][(Hx+b^o)-H\mu)]^T\} \\ &= \mathcal{E}\{[(Hx-H\mu)-b^o][(Hx-H\mu)-b^o]^T\} \\ &= H\mathcal{E}\{(x-\mu)(x-\mu)^T\}H^T + \mathcal{E}\{b^o b^{oT}\} \\ &\quad + H\mathcal{E}\{(x-\mu)b^{oT}\} + \mathcal{E}\{b^o(x-\mu)^T\}H^T. \end{aligned}$$

Since we assume x and b^o to be independent $\mathcal{E}\{xb^{oT}\} = 0$, and since b^o is zero mean, it follows that

$$P_y = HPH^T + R$$

Consequently,

$$\begin{aligned} p_y(y) &= \frac{1}{(2\pi)^{m/2}|(HPH^T+R)|^{1/2}} \\ &\quad \times \exp\left[-\frac{1}{2}(y-H\mu)^T(HPH^T+R)^{-1}(y-H\mu)\right] \end{aligned}$$

It remains to determine the conditional pdf $p_{y|x}(y|x)$. This distribution is also Gaussian, and can be written as

$$p_{y|x}(y|x) = \frac{1}{(2\pi)^{m/2}|P_{y|x}|^{1/2}} \exp\left[-\frac{1}{2}(y-\mu_{y|x})^T P_{y|x}^{-1}(y-\mu_{y|x})\right]$$

Analogously to what we have just done above,

$$\mu_{y|x} = \mathcal{E}\{Hx|x\} + \mathcal{E}\{b^o|x\} = Hx$$

and

$$\begin{aligned} P_{y|x} &= \mathcal{E}\{(y-\mu_{y|x})(y-\mu_{y|x})^T|x\} \\ &= \mathcal{E}\{[(Hx+b^o)-Hx)][(Hx+b^o)-Hx)]^T|x\} \\ &= \mathcal{E}\{b^o b^{oT}|x\} \\ &= \mathcal{E}\{b^o b^{oT}\} \\ &= R. \end{aligned}$$

Therefore,

$$p_{y|x}(y|x) = \frac{1}{(2\pi)^{m/2}|R|^{1/2}} \exp\left[-\frac{1}{2}(y-Hx)^T R^{-1}(y-Hx)\right]$$

which is the conditional probability of y given x.

Combining the previous results in Bayes rule for pdf's:

$$p_{x|y}(x|y) = \frac{|HPH^T+R|^{1/2}}{(2\pi)^{n/2}|P|^{1/2}|R|^{1/2}} \exp[-J]$$

where J is defined as,

$$\begin{aligned} J(x) \equiv\ & (y-Hx)^T R^{-1}(y-Hx) + (x-\mu)^T P^{-1}(x-\mu) \\ & -(y-H\mu)^T(HPH^T+R)^{-1}(y-H\mu) \end{aligned}$$

This quantity J can also be written in the following more compact form:

$$J(x) = \frac{1}{2}(x-\hat{x})^T P_{\tilde{x}}^{-1}(x-\hat{x})$$

where $P_{\tilde{x}}^{-1}$ is given by

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

the vector $\hat{x}$ is given by

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

and the reason for using the subscript $\tilde{x}$ for the matrix $P_{\tilde{x}}$, indicating a relationship with the estimation error, will soon become clear.

We are now ready to derived the desired estimates.

The maximum *a posteriori* probability estimate is the one that maximizes $p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})$, and is easily identified to be (Ex. 4)

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

The minimum variance estimate is given by the conditional mean of the *a posteriori* pdf (Ex. 3), that is,

$$\hat{\mathbf{x}}_{\mathsf{MV}} = \int_{-\infty}^{\infty} \mathbf{x} p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\, d\mathbf{x} = \hat{\mathbf{x}}$$

where the last equality can be derived after some algebra.

The maximum likelihood estimate can be determined by maximizing the pdf $p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})$, that is,

$$\mathbf{0} = \left.\frac{\partial p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\hat{\mathbf{x}}_{\mathsf{ML}}} = \mathbf{H}^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}_{\mathsf{ML}})$$

that is,

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

which is, in principle, distinct from the estimates obtained above (Ex. 5).

The MV and MAP estimates can be reduced to the ML estimate by taking $\mathbf{P}^{-1} = \mathbf{0}$, that is, when no statistical information on $\mathbf{x}$ is available:

$$\hat{\mathbf{x}}_{\mathsf{MV}}|_{\mathbf{P}^{-1}=\mathbf{0}} = \hat{\mathbf{x}}_{\mathsf{MAP}}|_{\mathbf{P}^{-1}=\mathbf{0}} = (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{y} = \hat{\mathbf{x}}_{\mathsf{ML}}$$

Quick Recap

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

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

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

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

where

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

and

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

General Cost Function:

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

Estimation Results:

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

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

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

The Least-Squares (LS) Connection

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

Minimization of the cost function

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

results in

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

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

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

with minimum achieved for

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

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.

Remarks

- All estimates above result in a *linear combination* of the observations.

- The MAP estimate can be obtained by minimizing the alternative cost function

$$J_{\mathsf{MAP}}(\mathbf{x}) \equiv \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{P}^{-1}(\mathbf{x}-\boldsymbol{\mu})+\frac{1}{2}(\mathbf{y}-\mathbf{H}\mathbf{x})^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x})\,,$$

  which amounts to noticing that the pdf $p_{\mathbf{z}}(\mathbf{y})$ does not play any role in the maximization of the *a posteriori* pdf.

- Similarly, the ML estimate can be obtained by minimizing the following cost function:

$$J_{\mathsf{ML}}(\mathbf{x}) \equiv \frac{1}{2}(\mathbf{y}-\mathbf{H}\mathbf{x})^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x})\,,$$

  and corresponding estimate is biased.

- In general there is no guarantee these three estimates coincide. In the case just considered they only coincide after knowledge on the prior is ignored in the MV and MAP results.

4. 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 $J_{\mathrm{LSP}}(\mathbf{x})$ seen previously,

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

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

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

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

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

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. 6):

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

  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.

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. 7).

- 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., NCEP's Grid-space Statistical Interpolation Approach; and ECMWF's 3d-var - as derived from its current 4d-Var).

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.

5. 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., $\mathbf{q}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_i)$.

Quick Recap

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

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

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

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

where

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

and

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

General Cost Function:

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

Estimation Results:

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

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

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

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. 8),

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

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.

Quick Recap

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

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

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

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

where

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

and

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

General Cost Function:

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

Estimation Results:

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

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

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

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:

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

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

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. 9)

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

  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

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

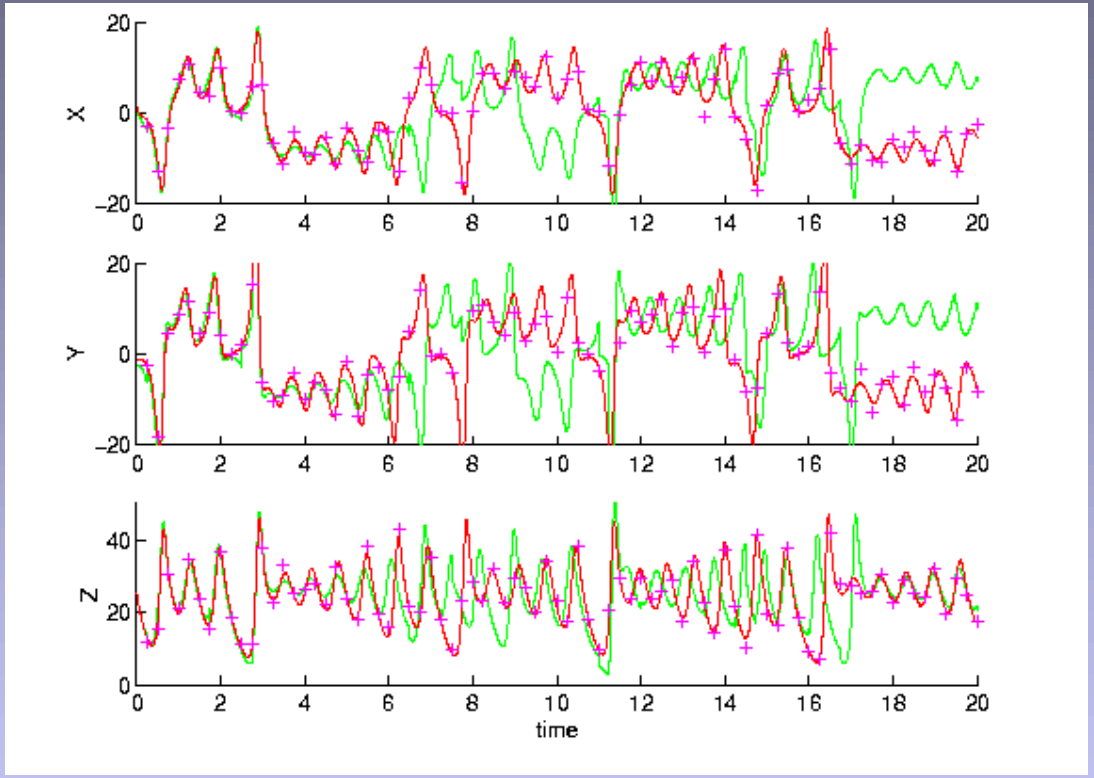
  for $m = 1, 2, \cdots, j$, and with $\mathbf{g}_0 \equiv \mathbf{B}\mathbf{f}_0$. This sequence of operations is known as the sweeper method and specifically constitute the so called augmented representer approach to the practical solution to calculating the 4d-PSAS equation (Ex. 10).

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

Illustration 1(cont.): Data Assimilation for Chaotic Dynamics

Then, what does data assimilation do?

σ(obs) = 2

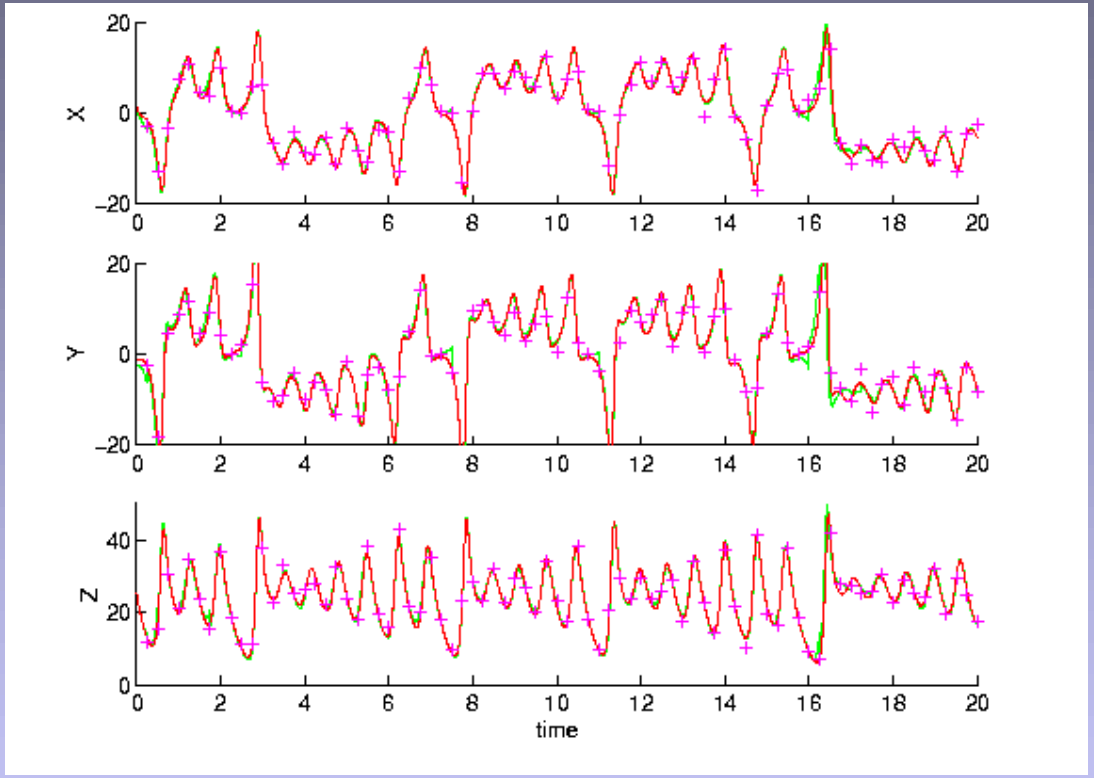


Answer: It improves our ability to estimate the true state and make relatively reasonable short- to medium-range predictions. However, depending on the data assimilation scheme, the estimate may diverge after a while. The red line represents the true state while the green line represents the estimate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

Illustration 1(cont.): Data Assimilation for Chaotic Dynamics

What if the data assimilation scheme is improved?

σ(obs) = 2

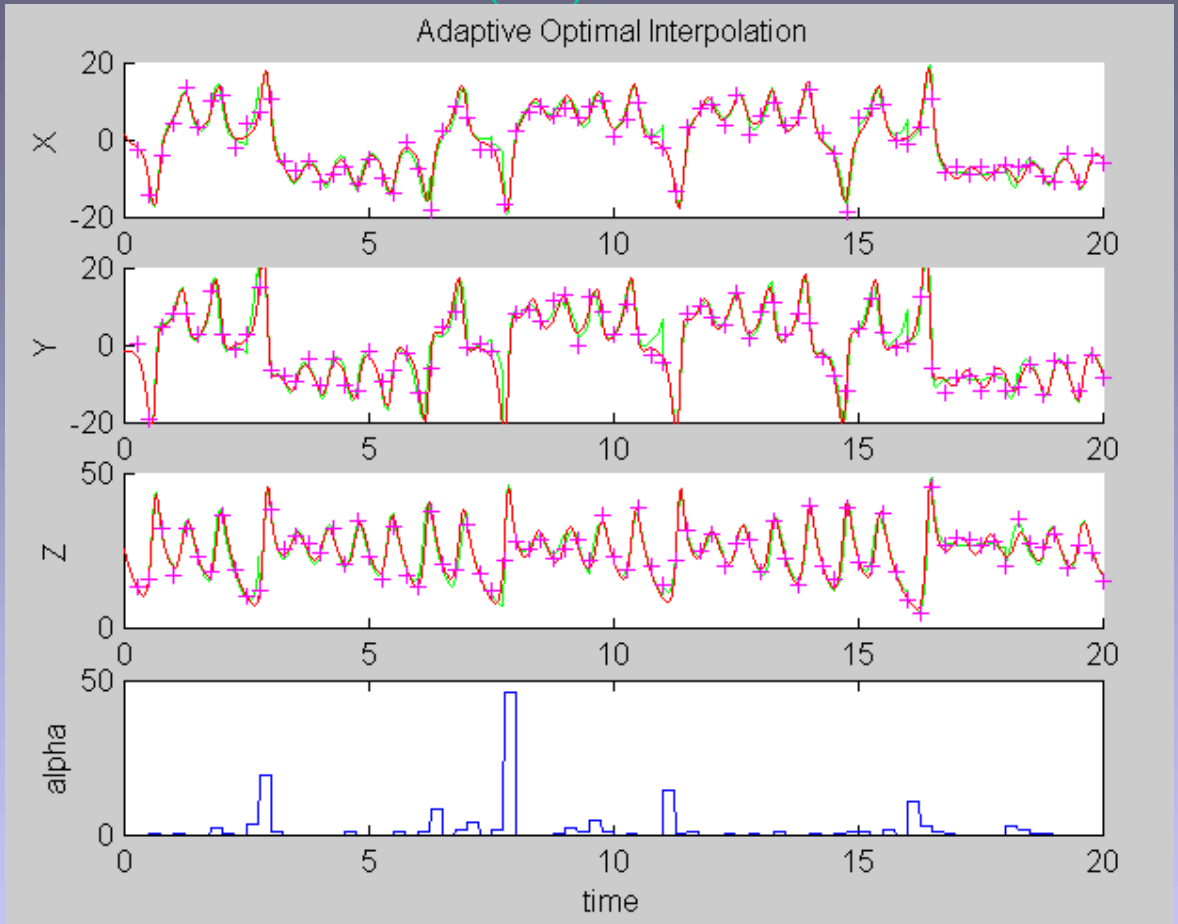


Answer: We get great results! Here we estimate the error due to linearization in the EKF via a Monte Carlo procedure proposed by Miller et al. (1994). We calculate the model error covariance off-line and then add that to the on-line EKF assimilation procedure. This is a rather good solution to prevent the EKF divergence due to misrepresentation of nonlinearities. However, this is a bit impractical for large data assimilation systems.

Illustration 1(cont.): Data Assimilation for Chaotic Dynamics

How does a simplified assimilation scheme perform?

σ(obs) = 2



Answer: Quite well! 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 this estimated parameter is displayed in the lower panel above.

8. Closing Remarks

- Most of the methods to solve inverse problems are either Least-Squares-based or bear a close relationship to Least-Squares.

- Beginners in the field should learn well Least-Squares, what it means, and how it relates to methods such as the Kalman filter/smoother, and 3d/4d variational procedures.

- Iterative methods for solving matrix-vector problems are often employed when calculating Least-Squares-like solutions to estimation problems. So, learn well conjugate-gradient, Newton-methods, etc.

Short (very biased) Reference List

Anderson, B.D.O., & J.B. Moore, 1979: *Optimal Filtering*. Prentice-Hall, 357 pp.

Cohn, S.E., 1997: An introduction to estimation theory. M. Ghil, K. Ide, A. Bennett, P. Courtier, M. Kimoto, N. Nagata, & N. Sato (Eds.): *Data Assimilation in Meteorology and Oceanography: Theory and Practice*, Universal Academic Press, 147-178.

Courtier, P., 1997: Dual formulation of four-dimensional variational assimilation. Q. J. Roy. Meteor. Soc., Part B, 123, 2449-2461.

Daley, R., 1991: *Atmospheric Data Analysis*. Cambridge University Press, 457 pp.

Dee, D.P., 1995: On-line estimation of error covariance parameters for atmospheric data assimilation. Mon. Wea. Rev., **123**, 1128-1145.

Dee, D.P., & R. Todling, 2000: Data assimilation in the presence of forecast bias: the GEOS moisture analysis. *Mon. Wea. Rev*., 128, 3268-3282.

Ghil, M., & P. Malanotte-Rizzoli, 1991: Data assimilation in meteorology and oceanography. *Advances Geophys*., Vol. 33, Academic Press, 141-266.

Jazwinski, A.H., 1970: *Stochastic Processes and Filtering Theory*. Academic Press, 376 pp.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 341pp.

Tarantola, A., 1994: *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation*. Elsevier, 613 pp.

Todling, R., 1999: *Class Notes on Estimation Theory and Atmospheric Data Assimilation*. NASA/DAO Office Note 99-01, 187 pp [Avail. Online: http://gmao.gsfc.nasa.gov/pubs/on/archive/on_1999.php].

Todling, R., & S.E. Cohn, 1994: Suboptimal schemes for atmospheric data assimilation based on the Kalman filter. Mon. Wea. Rev., **122**, 2530-2557.

Todling, R., S.E. Cohn, & N.S. Sivakumaran, 1998: Suboptimal schemes for retrospective data assimilation based on the fixed-lag Kalman smoother. Mon. Wea. Rev., **126**, 2274-2286.

Wunsch, C., 1996: *The Ocean Circulation Inverse Problem*. Cambridge University Press, 442 pp.

Wunsch, C., 2006: *Discrete Inverse and State Estimation Problems*. Cambridge University Press, 371 pp.