

Introduction to Data Assimilation
or alternatively
Introduction to Estimation Theory

Ricardo Todling

Global Modeling and Assimilation Office, NASA/GSFC
E-mail: todling@gmao.gsfc.nasa.gov

University of Maryland

9 February 2007

Slides available online from:

<http://gmao.gsfc.nasa.gov/pubs/presentations/2007>

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

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 representative-ness 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 $p_{\mathbf{x},\mathbf{z}}(\mathbf{x}, \mathbf{y})$.

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

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

and Bayes rule for converting between conditional pdf's:

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

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:

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

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:

$$J = \frac{1}{2} \|\tilde{\mathbf{x}}\|_{\mathbf{E}} = \frac{1}{2} \tilde{\mathbf{x}}^T \mathbf{E} \tilde{\mathbf{x}}$$

(b) The uniform cost:

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

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:

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

Or, identifying the kernel as the conditional Bayes cost:

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

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

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

And noticing that p is a pdf, it follows that

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

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

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

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

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

or yet, we can minimize the conditional Bayes cost

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

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

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

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

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

or yet

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

- ▶ 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 $\mathcal{N}(\boldsymbol{\mu}_x, \mathbf{P}_x)$, then

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

Hence,

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

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

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

or equivalently,

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

- ▷ $\hat{\mathbf{x}}_{\text{ML}}$ 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:

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

Conditional mean:

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

Minimum variance estimate:

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

Maximum *a posteriori* probability estimate:

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

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

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

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

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

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

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

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

Linear transformations of Gaussian distributed variables result in Gaussian distributed variables (Ex. 2). Hence,

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

where $\boldsymbol{\mu}_{\mathbf{z}}$ and $\mathbf{P}_{\mathbf{z}}$ correspond to the mean and covariance of the random variable \mathbf{y} , respectively.

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

$$\boldsymbol{\mu}_z = \mathcal{E}\{\mathbf{H}\mathbf{x}\} + \mathcal{E}\{\mathbf{b}^o\} = \mathbf{H}\boldsymbol{\mu}$$

and also,

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

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

$$\mathbf{P}_z = \mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R}$$

Consequently,

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

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

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

Analogously to what we have just done above,

$$\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}} = \mathcal{E}\{\mathbf{H}\mathbf{x}|\mathbf{x}\} + \mathcal{E}\{\mathbf{b}^o|\mathbf{x}\} = \mathbf{H}\mathbf{x}$$

and

$$\begin{aligned} \mathbf{P}_{\mathbf{z}|\mathbf{x}} &= \mathcal{E}\{(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}})(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}})^T|\mathbf{x}\} \\ &= \mathcal{E}\{[(\mathbf{H}\mathbf{x} + \mathbf{b}^o) - \mathbf{H}\mathbf{x}][(\mathbf{H}\mathbf{x} + \mathbf{b}^o) - \mathbf{H}\mathbf{x}]^T|\mathbf{x}\} \\ &= \mathcal{E}\{\mathbf{b}^o\mathbf{b}^{oT}|\mathbf{x}\} \\ &= \mathcal{E}\{\mathbf{b}^o\mathbf{b}^{oT}\} \\ &= \mathbf{R}. \end{aligned}$$

Therefore,

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

which is the conditional probability of \mathbf{y} given \mathbf{x} .

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

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

where J is defined as,

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

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

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

where $\mathbf{P}_{\tilde{\mathbf{x}}}^{-1}$ is given by

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

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

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

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

We are now ready to derive the desired estimates.

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

$$\hat{\mathbf{x}}_{\text{MV}} = \hat{\mathbf{x}} = \int_{-\infty}^{\infty} \mathbf{x} p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{y}) d\mathbf{x} = \boldsymbol{\mu}$$

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

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

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

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

that is,

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

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:

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

Quick Recap

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

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

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

$$p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{y}) \propto \exp\left[-\frac{1}{2}(\mathbf{x} - \hat{\mathbf{x}})^T \mathbf{P}_{\hat{\mathbf{x}}}^{-1}(\mathbf{x} - \hat{\mathbf{x}})\right]$$

where

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

and

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

General Cost Function:

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

Estimation Results:

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

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

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

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

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:

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

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 $J_{\text{LSP}}(\mathbf{x})$ seen previously,

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

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

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

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 write $\mathbf{y} - \mathbf{h}(\mathbf{x})$ 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

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

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, they 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 essentially means that one can be converted into 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 such as 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. Difficulties in relaxing the assumptions behind these spectral space formulations has driven the reformulation of the covariances to operate in physical-space. Modern **3d-var** systems now minimize the cost function directly, but 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., $\mathbf{q}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_i)$.

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

$$J_{4dvar} = J_{4dvar}(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_I) \text{ to } J_{4dvar} = J_{4dvar}(\delta\mathbf{x}_0, \mathbf{q}_1, \dots, \mathbf{q}_I).$$

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:

$$\delta\mathbf{x} \equiv [\delta\mathbf{x}_0^T \mathbf{q}_1^T \dots \mathbf{q}_I^T]^T \text{ and } \mathbf{d} \equiv [\mathbf{d}_0^T \mathbf{d}_1^T \dots \mathbf{d}_I^T]^T. \text{ Therefore (Ex. 8),}$$

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

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

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

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.

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.

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 λ . The main complication in this operation comes from having to calculate $\mathbf{GDG}^T\lambda$. To do so, we can notice that an element j of this term is given by (**Ex. 9**)

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

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

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

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

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

for $m = 1, 2, \dots, j$, 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, $\mathbf{Q} = \mathbf{0}$, the **4d-var** and **4d-PSAS** equations above dramatically simplify.

6. The Probabilistic Approach to Filtering

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

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

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

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

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

Whiteness of the observation sequence allows us to write

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

and therefore,

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

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, $\mathbf{x}_k^o = \mathbf{H}_k \mathbf{x}_k^t + \mathbf{b}_k^o$, 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:

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

Consequently we have

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

where we noticed that

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

and

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

Analogously, we have

$$p(\mathbf{x}_k^o | \mathbf{X}_{k-1}^o) = \frac{1}{(2\pi)^{m_k/2} |\mathbf{\Gamma}_k|^{1/2}} \exp \left[-\frac{1}{2} (\mathbf{x}_k^o - \mathbf{H}_k \mathbf{x}_{k|k-1}^f)^T \mathbf{\Gamma}_k^{-1} (\mathbf{x}_k^o - \mathbf{H}_k \mathbf{x}_{k|k-1}^f) \right]$$

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

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

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

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

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

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

and the conditional pdf of interest can be written as

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

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 $\hat{\mathbf{x}}_{MV}$ and $\mathbf{P}_{\hat{\mathbf{x}}}$ 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

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

Remarks

- ▷ The estimate $\mathbf{x}_{k|k}^a$ 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,

$$\mathbf{x}_{k|k}^a \equiv \mathcal{E}\{\mathbf{x}_k^t | \mathbf{X}_k^o\},$$

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

- ▷ Similar results can be obtained by minimizing the cost function

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

where $\delta\mathbf{x}_k \equiv \mathbf{x}_k^t - \mathbf{x}_{k|k-1}^f$, and $\mathbf{d}_k \equiv \mathbf{x}_k^o - \mathbf{H}_k \mathbf{x}_{k|k-1}^f$. In the meteorological literature $J_{3d\text{Var}}(\delta\mathbf{x}_k)$ 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 $\mathbf{P}_{k|k-1}^f$, the minimization problem above solves none other than a LSP problem, given some prior information.

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

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

7. The Probabilistic Approach to Smoothing

Smoothing is the problem of determining the state of a system given all the data available before, during, and after the time of the desired estimate. In this respect, the smoothing problem refers to the following conditional pdf

$$p(\mathbf{X}_k^t | \mathbf{X}_N^o) = p(\mathbf{x}_1^t, \dots, \mathbf{x}_{k-1}^t, \mathbf{x}_k^t | \mathbf{x}_1^o, \dots, \mathbf{x}_{N-1}^o, \mathbf{x}_N^o)$$

where $N \geq k$.

Remarks

- ▷ In general, an estimate obtained by maximizing the pdf

$$p(\mathbf{x}_k^t | \mathbf{x}_1^o, \dots, \mathbf{x}_{k-1}^o, \mathbf{x}_k^o)$$

will be distinct from one maximizing $p(\mathbf{X}_k^t | \mathbf{X}_N^o)$ above.

- ▷ However, in the linear, Gaussian, white noise case, with $N = k$, maximization of either one of the pdf's above amounts to the same solution, at the final time t_k .
- ▷ When the error (noise) statistics are Gaussian, the pdf $p(\mathbf{X}_k^t | \mathbf{X}_N^o)$ is also Gaussian and its maximization amounts to minimization of the following quadratic cost function:

$$J_N = \sum_{i=0}^N \|\mathbf{x}_i^o - \mathbf{H}_i \mathbf{x}_{i-1}\|_{\mathbf{R}_i^{-1}} + \sum_{i=0}^N \|\mathbf{x}_i - \mathbf{M}_{i,i-1} \mathbf{x}_{i-1}\|_{\mathbf{Q}_i^{-1}}$$

with respect to the entire trajectory $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_N$, and subjected to the ICs: $\mathbf{M}_{0,-1} \mathbf{x}_{-1} = \bar{\mathbf{x}}_0$ and $\mathbf{Q}_{-1} = \bar{\mathbf{P}}_0$.

- ▷ Minimization of the cost function J_N solves the **fixed-interval smoother**. In some sense, this is the problem that **4Dvar** attempts to solve.

There are different ways of solving the smoother problem sequentially. The **fixed-lag Kalman smoother** is a particularly attractive formulation.

To exemplify consider the case of seeking an improved state estimate at t_{k-1} given observations up to time t_k . The relevant pdf is

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

When all pdf's are Gaussian we can show that the maximum probability is obtained by minimizing

$$\begin{aligned} J_{lag=1} &= (\mathbf{x}_k^o - \mathbf{H}_k \mathbf{M}_{k,k-1} \mathbf{x}_{k-1}^t)^T \tilde{\mathbf{R}}_k^{-1} (\mathbf{x}_k^o - \mathbf{H}_k \mathbf{M}_{k,k-1} \mathbf{x}_{k-1}^t) \\ &+ (\mathbf{x}_{k-1|k-1}^a - \mathbf{x}_{k-1}^t)^T (\mathbf{P}_{k-1|k-1}^a)^{-1} (\mathbf{x}_{k-1|k-1}^a - \mathbf{x}_{k-1}^t) \end{aligned}$$

where $\tilde{\mathbf{R}}_k \equiv (\mathbf{H}_k \mathbf{Q}_{k-1} \mathbf{H}_k^T + \mathbf{R}_k)$, and the optimal solution is found to be

$$\begin{aligned} \mathbf{x}_{k-1|k}^a &= \mathcal{E}\{\mathbf{x}_{k-1}^t | \mathbf{X}_k^o\} \\ &= \mathbf{x}_{k-1|k-1}^a + \mathbf{P}_{k-1|k-1}^a \mathbf{M}_{k,k-1}^T \mathbf{H}_k \Gamma_k^{-1} (\mathbf{x}_k^o - \mathbf{H}_k \mathbf{x}_{k|k-1}^f). \end{aligned}$$

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/smoothing, 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.

Exercises

1. What is the condition on the weighting matrix \mathbf{E} for the minimum variance estimate to be indeed a minimum of the cost function?
2. Show that the linear transformation of a normally distributed vector is also normally distributed. That is, show that for a given normally distributed vector \mathbf{x} , with mean $\mu_{\mathbf{x}}$ and covariance $\mathbf{R}_{\mathbf{x}}$, the linear transformation

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$$

produces a normally distributed vector \mathbf{y} with mean $\mu_{\mathbf{y}} = \mathbf{A}\mu_{\mathbf{x}} + \mathbf{b}$ and covariance $\mathbf{R}_{\mathbf{y}} = \mathbf{A}\mathbf{R}_{\mathbf{x}}\mathbf{A}^T$.

3. Consider a simple scalar, Gaussian, case and show that the minimum variance estimate

$$x_{MV} = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x \exp \left[-\frac{(x - \mu)^2}{2\sigma^2} \right] dx = \mu$$

is indeed the mean value μ . (Hint: make use of the result $\int_{-\infty}^{\infty} \exp(-x^2/2) dx = \sqrt{2\pi}$).

4. Show that the maximum *a posteriori* probability estimate for the “estimation of a constant” case is indeed an unbiased estimate.

Exercises (cont.)

5. Show that the maximum likelihood estimate for the “estimation of a constant” case is a biased estimate.
6. Show that the solution of the **3d-var** minimization problem

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

can be written as

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

Hint: Make use of the Sherman-Morrison-Woodbury formula (c.f., Golub & Van Loan 1989, p. 51),
 $(\mathbf{A} + \mathbf{U} \mathbf{V}^T)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1} \mathbf{U} (\mathbf{I} + \mathbf{V}^T \mathbf{A}^{-1} \mathbf{U})^{-1} \mathbf{V}^T \mathbf{A}^{-1}$

7. Show that minimizing the cost function defined by

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

is equivalent to solving to first equation in the **PSAS** approach.

Exercises (cont.)

8. Show that using the augmented vectors \mathbf{x} and \mathbf{d} defined in the 4d-var section, the original cost function

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

can be written as

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

with the matrices \mathbf{D} , \mathbf{G} , and \mathbf{R} as defined in that same section. What are the explicit dimensions of each one of these matrices?

9. Unfold the first term in the first 4d-PSAS equation, that is, $(\mathbf{G}\mathbf{D}\mathbf{G}^T\boldsymbol{\lambda})$ to show that its j -th element has the form displayed in the text.
10. Show that the backward and forward sweeper steps are indeed an equivalent form of calculating the expression in Ex. 9 above.