Assimilation Algorithms
Lecture 1: Basic Concepts

Sébastien Massart and Mike Fisher

ECMWF

27 March 2017

ECMWF

Outline

©ECMWF

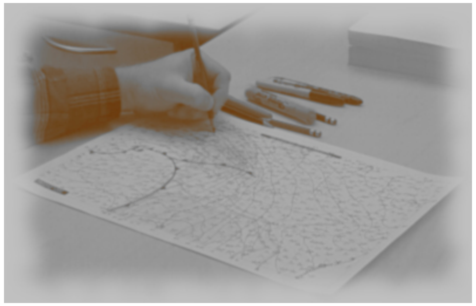
Outline

©ECMWF

Terminology

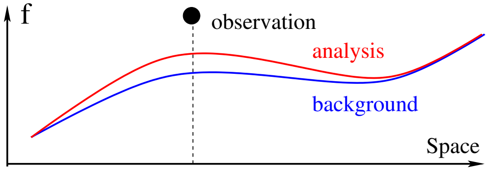
Analysis

- Analysis: The process of approximating the true state of a (geo)physical system at a given time using the available knowledge.
- For example:
  - Hand analysis of synoptic observations (1850 LeVerrier, Fitzroy).
  - Polynomial Interpolation (1950s Panofsky)



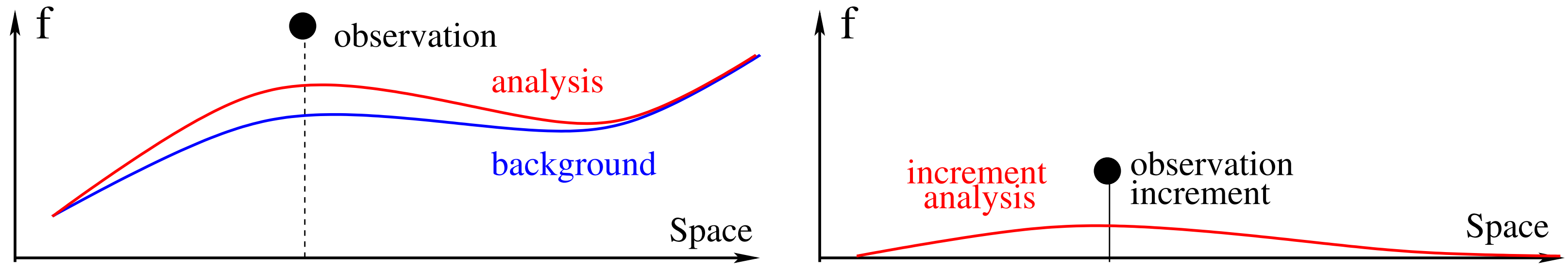
Background

- An important step forward was made by Gilchrist and Cressman (1954), who introduced the idea of using a previous numerical forecast to provide a preliminary estimate of the analysis.
- This prior estimate was called the background.



Optimal interpolation

- Bergthorsson and Döös (1955) took the idea of using a background field a step further by casting the analysis problem in terms of increments which were added to the background.
- The increments were weighted linear combinations of nearby observation increments (observation minus background), with the weights determined statistically.
- This idea of statistical combination of background and synoptic observations led ultimately to Optimal Interpolation.
- The use of statistics to merge model fields with observations is fundamental to all current methods of analysis.



Data Assimilation

- An important change of emphasis happened in the early 1970s with the introduction of primitive-equation models.
- Primitive equation models support inertia-gravity waves. This makes them much more fussy about their initial conditions than the filtered models that had been used hitherto.
- The analysis procedure became much more intimately linked with the model. The analysis had to produce an initial state that respected the model's dynamical balances.
- Unbalanced increments from the analysis procedure would be rejected as a result of geostrophic adjustment.
- Initialisation techniques (which suppress inertia-gravity waves) became important.

ⓒECMWF

Data Assimilation

The idea that the analysis procedure must present observational information to the model in a way in which it can be absorbed (i.e. not rejected by geostrophic adjustment) led to the coining of the term data assimilation.

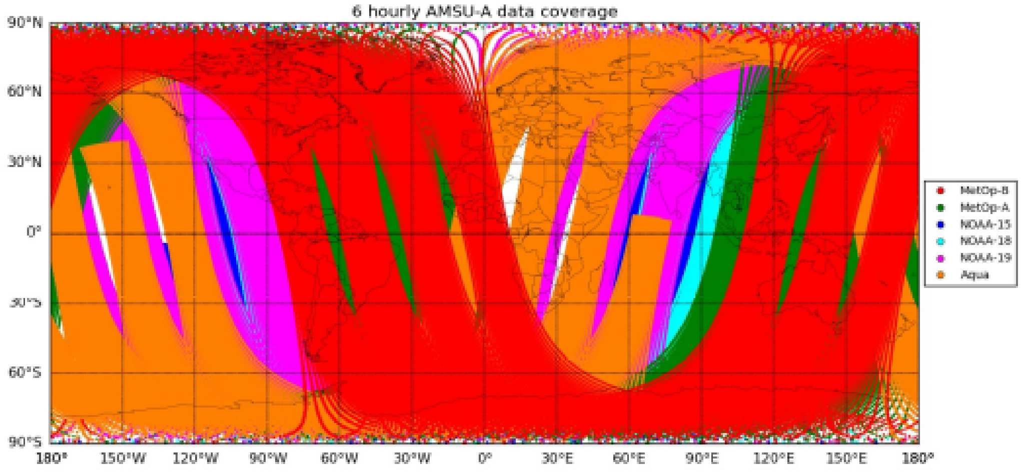
Google define:Assimilate

- ✗ To incorporate nutrients into the body after digestion
- ✗ To incorporate or absorb knowledge into the mind
- ✗ The social process of absorbing one cultural group into harmony with another
- ✗ The process by which the Borg integrate beings and cultures into their collective.
- ✗ The process of objectively adapting the model state to observations in a statistically optimal way taking into account model and observation errors

©ECMWF

Data Assimilation

- A final impetus towards the modern concept of data assimilation came from the increasing availability of asynoptic observations from satellite instruments.
- It was no longer sufficient to think of the analysis purely in terms of spatial interpolation of contemporaneous observations.
- The time dimension became important, and the model dynamics assumed the role of propagating observational information in time to allow a synoptic view of the state of the system to be generated from asynoptic data.



- Example of satellite data coverage in 6 hours (AMSU-A data).

©ECMWF

Outline

©ECMWF

Elementary Statistics

Problem

Suppose we want to estimate the temperature of this room, given:

- A prior estimate: T_b.
  - E.g., we measured the temperature an hour ago, and we have some idea (i.e. a model) of how the temperature varies as a function of time, the number of people in the room, whether the windows are open, etc.
- A thermometer: T_o.

Errors

- Denote the true temperature of the room by T_t.
- The errors in T_b and T_o are:

$$
\begin{aligned}
\varepsilon_b &= T_b - T_t \\
\varepsilon_o &= T_o - T_t
\end{aligned}
$$

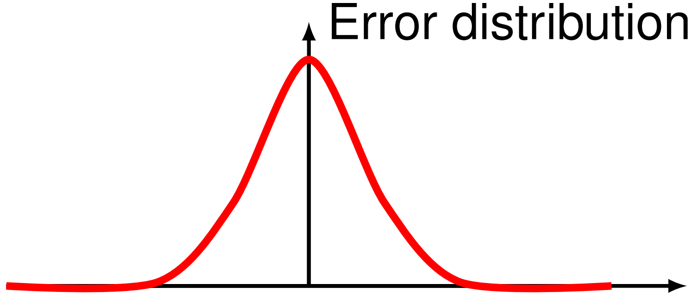
- ε_b and ε_o are random variables (or stochastic variables)

©ECMWF

Elementary Statistics

Hypotheses

- We will assume that the error statistics of T_b and T_o are known.



- We will assume that T_b and T_o have been adjusted (bias corrected) so that their mean errors are zero:

$$\overline{\varepsilon_b} = \overline{\varepsilon_o} = 0\,.$$

- There is usually no reason for ε_b and ε_o to be connected in any way:

$$\overline{\varepsilon_o \varepsilon_b} = 0\,.$$

- The quantity $\overline{\varepsilon_o \varepsilon_b}$ represents the covariance between the error of our prior estimate and the error of our thermometer measurement.

©ECMWF

Elementary Statistics

- We estimate the temperature of the room as a linear combination of T_b and T_o:

$$T_a = \alpha T_o + \beta T_b + \gamma$$

- Denote the error of our estimate as $\varepsilon_a = T_a - T_t$.
- We want the estimate to be unbiased: $\overline{\varepsilon_a} = 0$.
- We have:

$$T_a = T_t + \varepsilon_a = \alpha (T_t + \varepsilon_o) + \beta (T_t + \varepsilon_b) + \gamma$$

- Taking the mean and rearranging gives:

$$\overline{\varepsilon_a} = (\alpha + \beta - 1) T_t + \gamma$$

- Since this holds for any T_t, we must have
  - $\gamma = 0$, and
  - $\alpha + \beta - 1 = 0$.
- I.e. $T_a = \alpha T_o + (1 - \alpha) T_b$

©ECMWF

Elementary Statistics

- The general Linear Unbiased Estimate is:

$$T_a = \alpha T_o + (1-\alpha) T_b$$

- Now consider the error of this estimate.
- Subtracting T_t from both sides of the equation gives

$$\varepsilon_a = \alpha \varepsilon_o + (1-\alpha) \varepsilon_b$$

- The variance of the estimate is:

$$\overline{\varepsilon_a^2} = \alpha^2 \overline{\varepsilon_o^2} + 2\alpha(1-\alpha)\overline{\varepsilon_o \varepsilon_b} + (1-\alpha)^2 \overline{\varepsilon_b^2}$$

- With the previous hypothesis $\overline{\varepsilon_o \varepsilon_b} = 0$:

$$\overline{\varepsilon_a^2} = \alpha^2 \overline{\varepsilon_o^2} + (1-\alpha)^2 \overline{\varepsilon_b^2}$$

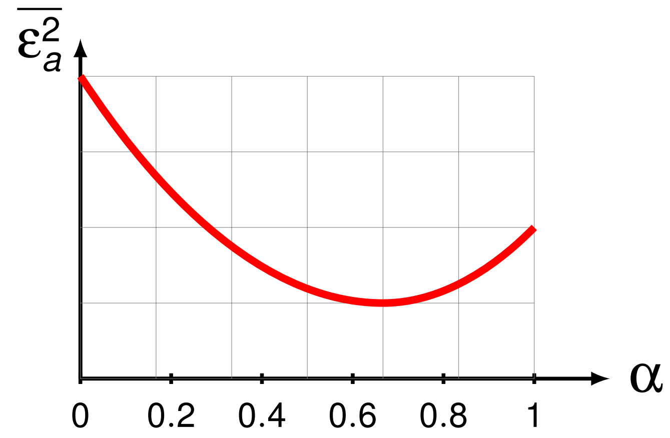
©ECMWF

Elementary Statistics

$$\overline{\varepsilon_a^2} = \alpha^2 \overline{\varepsilon_o^2} + (1-\alpha)^2 \overline{\varepsilon_b^2}$$

We can easily derive some properties of our estimate:

- ✘ $\frac{d\overline{\varepsilon_a^2}}{d\alpha} = 2\alpha\overline{\varepsilon_o^2} - 2(1-\alpha)\overline{\varepsilon_b^2}$
- ✘ For $\alpha = 0$, $\overline{\varepsilon_a^2} = \overline{\varepsilon_b^2}$ and $\frac{d\overline{\varepsilon_a^2}}{d\alpha} = -2\overline{\varepsilon_b^2} < 0$
- ✘ For $\alpha = 1$, $\overline{\varepsilon_a^2} = \overline{\varepsilon_o^2}$ and $\frac{d\overline{\varepsilon_a^2}}{d\alpha} = 2\overline{\varepsilon_o^2} > 0$



From this we can deduce:

- ✘ For $0 \leq \alpha \leq 1$, $\overline{\varepsilon_a^2} \leq \max(\overline{\varepsilon_b^2}, \overline{\varepsilon_o^2})$
- ✘ The minimum-variance estimate occurs for $\alpha \in (0,1)$.
- ✘ The minimum-variance estimate satisfies $\overline{\varepsilon_a^2} < \min(\overline{\varepsilon_b^2}, \overline{\varepsilon_o^2})$, which means it is lower than the variance of each piece of information.

©ECMWF

Elementary Statistics

The minimum-variance estimate occurs when

$$\frac{d\overline{\varepsilon_a^2}}{d\alpha} = 2\alpha\overline{\varepsilon_o^2} - 2(1-\alpha)\overline{\varepsilon_b^2} = 0$$

$$\Rightarrow \qquad \alpha = \frac{\overline{\varepsilon_b^2}}{\overline{\varepsilon_b^2} + \overline{\varepsilon_o^2}}.$$

It is not difficult to show that the error variance of this minimum-variance estimate is:

$$\frac{1}{\overline{\varepsilon_a^2}} = \frac{1}{\overline{\varepsilon_b^2}} + \frac{1}{\overline{\varepsilon_o^2}},$$

and the analysis is:

$$\frac{1}{\overline{\varepsilon_a^2}} T_a = \frac{1}{\overline{\varepsilon_b^2}} T_b + \frac{1}{\overline{\varepsilon_o^2}} T_o \,.$$

©ECMWF

Outline

©ECMWF

Extension to Multiple Dimensions

- Now, let's turn our attention to the multi-dimensional case.
- Instead of a scalar prior estimate T_b, we now consider a vector $\mathbf{x}_b$.
- We can think of $\mathbf{x}_b$ as representing the entire state of a numerical model at some time.
- The elements of $\mathbf{x}_b$ might be grid-point values, spherical harmonic coefficients, etc., and some elements may represent temperatures, humidity, others wind components, etc.
- We refer to $\mathbf{x}_b$ as the background
- Similarly, we generalise the observation to a vector $\mathbf{y}$.
- $\mathbf{y}$ can contain a disparate collection of observations at different locations, and of different variables.

©ECMWF

Extension to Multiple Dimensions

- ✗ The major difference between the simple scalar example and the multi-dimensional case is that there is no longer a one-to-one correspondence between the elements of the observation vector and those of the background vector.

$\mathbf{x} \rightarrow$  $\leftarrow \mathbf{y}$

- ✗ It is no longer trivial to compare observations and background.
- ✗ Observations are not necessarily located at model gridpoints
- ✗ The observed variables (e.g. radiances) may not correspond directly with any of the variables of the model.
- ✗ To overcome this problem, we must assume that our model is a more-or-less complete representation of reality, so that we can always determine “model equivalents” of the observations.

©ECMWF

Extension to Multiple Dimensions

- ✗ We formalise this by assuming the existence of an observation operator, $\mathcal{H}$.
- ✗ Given a model-space vector, $\mathbf{x}$, the vector $\mathcal{H}(\mathbf{x})$ can be compared directly with $\mathbf{y}$, and represents the "model equivalent" of $\mathbf{y}$.

$$\mathbf{x} \xrightarrow{\mathcal{H}(\cdot)} \mathcal{H}(\mathbf{x}) \rightarrow$$

$$\leftarrow \mathbf{y}$$

- ✗ For now, we will assume that $\mathcal{H}$ is perfect. I.e. it does not introduce any error, so that:

$$\mathcal{H}(\mathbf{x}_t) = \mathbf{y}_t$$

where $\mathbf{x}_t$ is the true state, and $\mathbf{y}_t$ contains the true values of the observed quantities.

©ECMWF

Extension to Multiple Dimensions

- As we did in the scalar case, we will look for an analysis that is a linear combination of the available information:

$$\mathbf{x}_a = \mathbf{F}\mathbf{x}_b + \mathbf{K}\mathbf{y} + \mathbf{c}$$

  where $\mathbf{F}$ and $\mathbf{K}$ are matrices, and where $\mathbf{c}$ is a vector.
- If $\mathcal{H}$ is linear, we can proceed as in the scalar case and look for a linear unbiased estimate.
- In the more general case of nonlinear $\mathcal{H}$, we will require that error-free inputs ($\mathbf{x}_b = \mathbf{x}_t$ and $\mathbf{y} = \mathbf{y}_t$) produce an error-free analysis ($\mathbf{x}_a = \mathbf{x}_t$):

$$\mathbf{x}_t = \mathbf{F}\mathbf{x}_t + \mathbf{K}\mathcal{H}(\mathbf{x}_t) + \mathbf{c}$$

- Since this applies for any $\mathbf{x}_t$, we must have $\mathbf{c} = 0$ and

$$\mathbf{F} \equiv \mathbf{I} - \mathbf{K}\mathcal{H}(\cdot)$$

- Our analysis equation is thus:

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{K}\left(\mathbf{y} - \mathcal{H}\left(\mathbf{x}_b\right)\right)$$

©ECMWF

Extension to Multiple Dimensions

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{K}\left(\mathbf{y} - \mathcal{H}(\mathbf{x}_b)\right)$$

- Remember that in the scalar case, we had

$$\begin{aligned} T_a &= \alpha T_o + (1-\alpha) T_b \\ &= T_b + \alpha (T_o - T_b) \end{aligned}$$

- We see that the matrix **K** plays a role equivalent to that of the coefficient α.
- **K** is called the gain matrix.
- It determines the weight given to the observation increment
- It handles the transformation of information defined in "observation space" to the space of model variables.

Extension to Multiple Dimensions

- The next step in deriving the analysis equation is to describe the statistical properties of the analysis errors.
- We define

$$
\begin{aligned}
\varepsilon_a &= \mathbf{x}_a - \mathbf{x}_t \\
\varepsilon_b &= \mathbf{x}_b - \mathbf{x}_t \\
\varepsilon_o &= \mathbf{y} - \mathbf{y}_t
\end{aligned}
$$

- We will assume that the errors are small, so that

$$\mathcal{H}(\mathbf{x}_b) = \mathcal{H}(\mathbf{x}_t) + \mathbf{H}\varepsilon_b + O(\varepsilon_b^2)$$

where **H** is the Jacobian of $\mathcal{H}$ (if **H** is nonlinear).

©ECMWF

Extension to Multiple Dimensions

- Substituting the expressions for the errors into our analysis equation, and using $\mathcal{H}(\mathbf{x}_t) = \mathbf{y}_t$, gives (to first order):

$$\varepsilon_a = \varepsilon_b + \mathbf{K}\left(\varepsilon_o - \mathbf{H}\varepsilon_b\right)$$

- As in the scalar example, we will assume that the mean errors have been removed, so that $\overline{\varepsilon_b} = \overline{\varepsilon_o} = 0$. We see that this implies that $\overline{\varepsilon_a} = 0$.
- In the scalar example, we derived the variance of the analysis error, and defined our optimal analysis to minimise this variance.
- In the multi-dimensional case, we must deal with covariances.

©ECMWF

Covariance

- The covariance between two variables x_i and x_j is defined as

$$\mathrm{cov}(x_i, x_j) = \overline{(x_i - \overline{x_i})(x_j - \overline{x_j})}$$

- Given a vector $\mathbf{x} = (x_1, x_2, \cdots, x_N)^{\mathrm{T}}$, we can arrange the covariances into a covariance matrix, $\mathbf{C}$, such that $C_{ij} = \mathrm{cov}(x_i, x_j)$.
- Equivalently:

$$\mathbf{C} = \overline{(\mathbf{x} - \overline{\mathbf{x}})(\mathbf{x} - \overline{\mathbf{x}})^{\mathrm{T}}}$$

- Covariance matrices are symmetric and positive definite
  - symmetric: $\mathbf{C}^T = \mathbf{C}$
  - positive definite: $\mathbf{z}^T\mathbf{C}\,\mathbf{z}$ is positive for every non-zero vector $\mathbf{z}$

©ECMWF

Extension to Multiple Dimensions

- The analysis error is:

$$\begin{aligned}\varepsilon_a &= \varepsilon_b + \mathbf{K}(\varepsilon_o - \mathbf{H}\varepsilon_b) \\ &= (\mathbf{I} - \mathbf{KH})\varepsilon_b + \mathbf{K}\varepsilon_o\end{aligned}$$

- Forming the analysis error covariance matrix gives:

$$\begin{aligned}\overline{\varepsilon_a\varepsilon_a^{\mathrm{T}}} &= \overline{[(\mathbf{I} - \mathbf{KH})\varepsilon_b + \mathbf{K}\varepsilon_o][(\mathbf{I} - \mathbf{KH})\varepsilon_b + \mathbf{K}\varepsilon_o]^{\mathrm{T}}} \\ &= \quad (\mathbf{I} - \mathbf{KH})\overline{\varepsilon_b\varepsilon_b^{\mathrm{T}}}(\mathbf{I} - \mathbf{KH})^{\mathrm{T}} + (\mathbf{I} - \mathbf{KH})\overline{\varepsilon_b\varepsilon_o^{\mathrm{T}}}\mathbf{K}^{\mathrm{T}} \\ &\quad + \mathbf{K}\overline{\varepsilon_o\varepsilon_b^{\mathrm{T}}}(\mathbf{I} - \mathbf{KH})^{\mathrm{T}} + \mathbf{K}\overline{\varepsilon_o\varepsilon_o^{\mathrm{T}}}\mathbf{K}^{\mathrm{T}}\end{aligned}$$

- Assuming that the background and observation errors are uncorrelated (i.e. $\overline{\varepsilon_o\varepsilon_b^{\mathrm{T}}} = \overline{\varepsilon_b\varepsilon_o^{\mathrm{T}}} = 0$), we find:

$$\overline{\varepsilon_a\varepsilon_a^{\mathrm{T}}} = (\mathbf{I} - \mathbf{KH})\overline{\varepsilon_b\varepsilon_b^{\mathrm{T}}}(\mathbf{I} - \mathbf{KH})^{\mathrm{T}} + \mathbf{K}\overline{\varepsilon_o\varepsilon_o^{\mathrm{T}}}\mathbf{K}^{\mathrm{T}}$$

©ECMWF

Extension to Multiple Dimensions

$$\overline{\varepsilon_a \varepsilon_a^{\mathrm{T}}} = (\mathbf{I} - \mathbf{KH}) \overline{\varepsilon_b \varepsilon_b^{\mathrm{T}}} (\mathbf{I} - \mathbf{KH})^{\mathrm{T}} + \mathbf{K} \overline{\varepsilon_o \varepsilon_o^{\mathrm{T}}} \mathbf{K}^{\mathrm{T}}$$

- ✗ This expression is the equivalent of the expression we obtained for the error of the scalar analysis:

$$\overline{\varepsilon_a^2} = (1 - \alpha)^2 \overline{\varepsilon_b^2} + \alpha^2 \overline{\varepsilon_o^2}$$

- ✗ Again, we see that **K** plays essentially the same role in the multi-dimensional analysis as α plays in the scalar case.
- ✗ In the scalar case, we chose α to minimise the variance of the analysis error.
- ✗ What do we mean by the minimum-variance analysis in the multi-dimensional case?

©ECMWF

Extension to Multiple Dimensions

- Note that the diagonal elements of a covariance matrix are variances $C_{ii} = \mathrm{cov}(x_i, x_i) = \overline{(x_i - \overline{x_i})^2}$.
- Hence, we can define the minimum-variance analysis as the analysis that minimises the sum of the diagonal elements of the analysis error covariance matrix.
- The sum of the diagonal elements of a matrix is called the trace.
- In the scalar case, we found the minimum-variance analysis by setting $\frac{d\overline{\varepsilon_a^2}}{d\alpha}$ to zero.
- In the multidimensional case, we are going to set

$$\frac{\partial \mathrm{trace}(\overline{\varepsilon_a \varepsilon_a^{\mathrm{T}}})}{\partial \mathbf{K}} = \mathbf{0}$$

- Note: $\frac{\partial \mathrm{trace}(\overline{\varepsilon_a \varepsilon_a^{\mathrm{T}}})}{\partial \mathbf{K}}$ is the matrix whose ij^{th} element is $\frac{\partial \mathrm{trace}(\overline{\varepsilon_a \varepsilon_a^{\mathrm{T}}})}{\partial K_{ij}}$.

©ECMWF

Extension to Multiple Dimensions

- We have: $\overline{\varepsilon_a \varepsilon_a^{\mathrm{T}}} = (\mathbf{I} - \mathbf{KH}) \overline{\varepsilon_b \varepsilon_b^{\mathrm{T}}} (\mathbf{I} - \mathbf{KH})^{\mathrm{T}} + \mathbf{K} \overline{\varepsilon_o \varepsilon_o^{\mathrm{T}}} \mathbf{K}^{\mathrm{T}}$.
- The following matrix identities come to our rescue:

$$\frac{\partial \mathrm{trace}(\mathbf{KAK}^{\mathrm{T}})}{\partial \mathbf{K}} = \mathbf{K}(\mathbf{A} + \mathbf{A}^{\mathrm{T}})$$

$$\frac{\partial \mathrm{trace}(\mathbf{KA})}{\partial \mathbf{K}} = \mathbf{A}^{\mathrm{T}} \qquad \frac{\partial \mathrm{trace}(\mathbf{AK}^{\mathrm{T}})}{\partial \mathbf{K}} = \mathbf{A}$$

- Applying these to $\partial \mathrm{trace}(\overline{\varepsilon_a \varepsilon_a^{\mathrm{T}}}) / \partial \mathbf{K}$ gives:

$$\frac{\partial \mathrm{trace}(\overline{\varepsilon_a \varepsilon_a^{\mathrm{T}}})}{\partial \mathbf{K}} = 2\mathbf{K} \left[\mathbf{H} \overline{\varepsilon_b \varepsilon_b^{\mathrm{T}}} \mathbf{H}^{\mathrm{T}} + \overline{\varepsilon_o \varepsilon_o^{\mathrm{T}}} \right] - 2 \overline{\varepsilon_b \varepsilon_b^{\mathrm{T}}} \mathbf{H}^{\mathrm{T}} = \mathbf{0}$$

- Hence: $\mathbf{K} = \overline{\varepsilon_b \varepsilon_b^{\mathrm{T}}} \mathbf{H}^{\mathrm{T}} \left[\mathbf{H} \overline{\varepsilon_b \varepsilon_b^{\mathrm{T}}} \mathbf{H}^{\mathrm{T}} + \overline{\varepsilon_o \varepsilon_o^{\mathrm{T}}} \right]^{-1}$.

©ECMWF

Extension to Multiple Dimensions

$$\mathbf{K} = \overline{\varepsilon_b \varepsilon_b^{\mathrm{T}}} \mathbf{H}^{\mathrm{T}} \left[\mathbf{H} \overline{\varepsilon_b \varepsilon_b^{\mathrm{T}}} \mathbf{H}^{\mathrm{T}} + \overline{\varepsilon_o \varepsilon_o^{\mathrm{T}}} \right]^{-1}$$

- This optimal gain matrix is called the Kalman Gain Matrix.
- Note the similarity with the optimal gain we derived for the scalar analysis: $\alpha = \overline{\varepsilon_b^2} / (\overline{\varepsilon_b^2} + \overline{\varepsilon_o^2})$.
- The variance of analysis error for the optimal scalar problem was:

$$\frac{1}{\overline{\varepsilon_a^2}} = \frac{1}{\overline{\varepsilon_b^2}} + \frac{1}{\overline{\varepsilon_o^2}}$$

- The equivalent expression for the multi-dimensional case is:

$$\left[\overline{\varepsilon_a \varepsilon_a^{\mathrm{T}}} \right]^{-1} = \left[\overline{\varepsilon_b \varepsilon_b^{\mathrm{T}}} \right]^{-1} + \mathbf{H}^{\mathrm{T}} \left[\overline{\varepsilon_o \varepsilon_o^{\mathrm{T}}} \right]^{-1} \mathbf{H}$$

©ECMWF

Notation

- The notation we have used for covariance matrices can get a bit cumbersome.
- The standard notation is:

$$
\begin{aligned}
\mathbf{P}^a &\equiv \overline{\varepsilon_a \varepsilon_a^{\mathrm{T}}} \\
\mathbf{P}^b &\equiv \overline{\varepsilon_b \varepsilon_b^{\mathrm{T}}} \\
\mathbf{R} &\equiv \overline{\varepsilon_o \varepsilon_o^{\mathrm{T}}}
\end{aligned}
$$

- In many analysis schemes, the true covariance matrix of background error, $\mathbf{P}^b$, is not known, or is too large to be used.
- In this case, we use an approximate background error covariance matrix. This approximate matrix is denoted by **B**.

©ECMWF

Alternative Expression for the Kalman Gain

Finally, we derive an alternative expression for the Kalman gain:

$$\mathbf{K} = \mathbf{P}^b\mathbf{H}^{\mathrm{T}} \left[\mathbf{H}\mathbf{P}^b\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right]^{-1}$$

Multiplying both sides by $\left[\mathbf{P}^{b^{-1}} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right]$ gives:

$$\begin{aligned}\left[\mathbf{P}^{b^{-1}} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right]\mathbf{K} &= \left[\mathbf{H}^{\mathrm{T}} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\mathbf{P}^b\mathbf{H}^{\mathrm{T}}\right]\left[\mathbf{H}\mathbf{P}^b\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right]^{-1} \\ &= \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\left[\mathbf{R} + \mathbf{H}\mathbf{P}^b\mathbf{H}^{\mathrm{T}}\right]\left[\mathbf{H}\mathbf{P}^b\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right]^{-1} \\ &= \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\end{aligned}$$

Hence:

$$\mathbf{K} = \left[\mathbf{P}^{b^{-1}} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right]^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}$$

- ✘ Expression 1: need the inverse of a matrix of dimension size(**R**)
- ✘ Expression 2: need the inverse of a matrix of dimension size($\mathbf{P}^b$)

©ECMWF

Outline

©ECMWF

Optimal Interpolation

- Optimal Interpolation is a statistical data assimilation method based on the multi-dimensional analysis equations we have just derived.
- The method was used operationally at ECMWF from 1979 until 1996, when it was replaced by 3D-Var.
- The basic idea is to split the global analysis into a number of boxes which can be analysed independently:

$$\mathbf{x}_a^{(i)} = \mathbf{x}_b^{(i)} + \mathbf{K}^{(i)} \left[\mathbf{y}^{(i)} - \mathcal{H}^{(i)}(\mathbf{x}_b) \right]$$

where

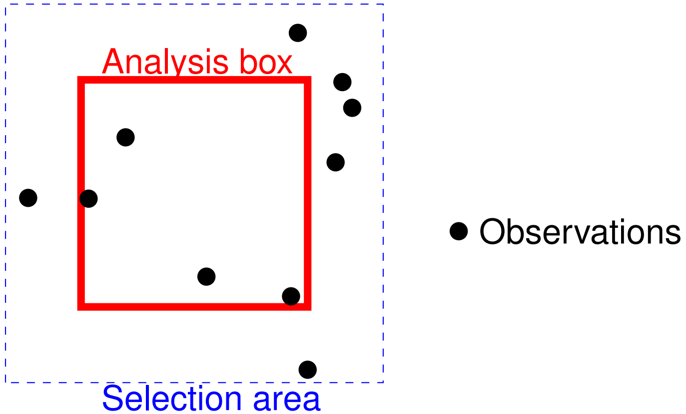
$$\mathbf{x}_a = \begin{pmatrix} \mathbf{x}_a^{(1)} \\ \mathbf{x}_a^{(2)} \\ \vdots \\ \mathbf{x}_a^{(M)} \end{pmatrix} \qquad \mathbf{x}_b = \begin{pmatrix} \mathbf{x}_b^{(1)} \\ \mathbf{x}_b^{(2)} \\ \vdots \\ \mathbf{x}_b^{(M)} \end{pmatrix} \qquad \mathbf{K} = \begin{pmatrix} \mathbf{K}^{(1)} \\ \mathbf{K}^{(2)} \\ \vdots \\ \mathbf{K}^{(M)} \end{pmatrix}$$

©ECMWF

Optimal Interpolation

$$\mathbf{x}_a^{(i)} = \mathbf{x}_b^{(i)} + \mathbf{K}^{(i)} \left(\mathbf{y}^{(i)} - \mathcal{H}^{(i)}(\mathbf{x}_b) \right)$$

- In principle, we should use *all* available observations to calculate the analysis for each box. However, this is too expensive.
- To produce a computationally-feasible algorithm, Optimal Interpolation (OI) restricts the observations used for each box to those observations which lie in a surrounding selection area:



©ECMWF

Optimal Interpolation

- The gain matrix used for each box is:

$$\mathbf{K}^{(i)} = \left(\mathbf{P}^b\mathbf{H}^{\mathrm{T}}\right)^{(i)} \left[\left(\mathbf{H}\mathbf{P}^b\mathbf{H}^{\mathrm{T}}\right)^{(i)} + \mathbf{R}^{(i)}\right]^{-1}$$

- Now, the dimension of the matrix $\left[\left(\mathbf{H}\mathbf{P}^b\mathbf{H}^{\mathrm{T}}\right)^{(i)} + \mathbf{R}^{(i)}\right]$ is equal to the number of observations in the selection box.
- Selecting observations reduces the size of this matrix, making it feasible to use direct solution methods to invert it.
- Note that to implement Optimal Interpolation, we have to specify $\left(\mathbf{P}^b\mathbf{H}^{\mathrm{T}}\right)^{(i)}$ and $\left(\mathbf{H}\mathbf{P}^b\mathbf{H}^{\mathrm{T}}\right)^{(i)}$. This effectively limits us to very simple observation operators, corresponding to simple interpolations.
- This, together with the artifacts introduced by observation selection, was one of the main reasons for abandoning Optimal Interpolation in favour of 3D-Var.

©ECMWF

Outline

©ECMWF

Summary

- We derived the linear analysis equation for a simple scalar example.
- We showed that a particular choice of the weight α given to the observation resulted in an optimal minimum-variance analysis.
- We repeated the derivation for the multi-dimensional case. This required the introduction of the observation operator.
- The derivation for the multi-dimensional case closely parallelled the scalar derivation.
- The expressions for the gain matrix and analysis error covariance matrix were recognisably similar to the corresponding scalar expressions.
- Finally, we considered the practical implementation of the analysis equation, in an Optimal Interpolation data assimilation scheme.

©ECMWF