Optimal Interpolation and 3DVAR (Three Dimensional Variational) Method

The optimal interpolation method will soon be discussed in detail by Dr. Carr in his lectures.

Here we present the optimal interpolation (OI) equations for vectors of observations and background fields, to lead us into the 3DVAR discussions. Many concepts and notations are also used by 3DVAR and later 4DVAR.

The OI equations were originally derived by Eliassen (1954, reproduced in Bengtsson et al, 1981). However, Gandin (1963), derived the multivariate OI equations independently and applied them to objective analysis in the former Soviet Union.

Gandin's work had a profound influence upon the research and operational community, and OI became the operational analysis scheme of choice during the 1980's and early 1990's.

In our discussion, we follow the general notation proposed by Ide et al (1997) for data assimilation methods.

It can be shown that 3D-Var is equivalent to the OI problem, except that the method to solve the problem is quite different and advantageous for operational systems.

Variable Notations

Consider the complete NWP operational problem of finding

\[ \mathbf{x}_a \] - an optimum analysis of a field of model variables,

given
\( x_b \) - a background field available at grid points, and
\( y_0 \) - a set of \( p \) observations available at irregularly spaced points \( r_i \).

\( x_a \) and \( x_b \) are vectors of length \( n \).

The unknown analysis and the known background can be 2D fields of a single variable like the temperature analysis \( T_a(x, y) \), or the 3D field of the initial conditions for all the model prognostic variables: \( x = (p_x, T, q, u, v) \).

These model variables are ordered by grid point and by variable, forming a single vector of length \( n \), the product of the number of points times the number of variables. The (unknown) "truth" \( x_t \), discretized at the model points, is also a vector of length \( n \).

A different variable \( y_o \) is for the observations than what we use of gridded variables.

The observed variables are, in general, different from the model variables by

a) being located in different points, and
b) by possibly being indirect measures of the model variables.

Examples of this are radar reflectivities and Doppler shifts, satellite radiances, and Global Positioning System (GPS) atmospheric refractivities.

For example, a simplified formula of the radar reflectivity formula given the mixing ratios of rain water, snow and hail species inside the cloud is

\[
Z = 10\{3 + \log_{10}[17.3(p q_r)^{7/4} + 38(p q_r + q_h)^{22}]\}
\]
where $Z$ is the reflectivity in dBz, $\rho$ is air density in kg/m$^3$, and $q_r$, $q_s$ and $q_h$ are the mixing ratios in g/kg. It is a formula that brings model variables $q_r$, $q_s$ and $q_h$ to the observed values $Z$, is therefore a forward operator, and it is a nonlinear function of q’s.

In the case of satellite radiances, it is the radiative transfer equations that relates the atmospheric temperature profile to the radiances that satellite sees at the top of the atmosphere.

References:


Optimal Analysis

The optimal analysis is equal to the background plus the innovation weighted by optimal weights which are determined so as to minimize the analysis error variance.

\[ x_a = x_b + W(y_o - H(x_b)) = x_b + Wd. \]

The analysis and the background are vectors of length \( n \) (the total number of grid points times the number of model variables), the weights are given by a matrix of dimension \( n \times p \).

\( H \) is the forward observational operator \( H \) that converts the background field into "observed first guesses."

\( H \) can be nonlinear (e.g., the radiative transfer equations that go from temperature and moisture vertical profiles to the satellite observed radiances).

The observation field \( y_o \) is a vector of length \( p \), the number of observations.

The vector \( d \), also of length \( p \), is the "innovation" or "observational increments" vector:

\[ d = y_o - H(x_b), \]

which is defined as the difference between the observation and the background mapped to the observational point via forward operator \( H \).
Remarks:

a) The weight matrix $W$ is also called the gain matrix $K$.

b) An error covariance matrix is obtained by multiplying a vector error $\varepsilon = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$ by its transpose $\varepsilon^T = [e_1 \ e_2 \ \ldots \ e_n]$, and averaging over many cases, to obtain the expected value:

$$P = \overline{\varepsilon\varepsilon^T} = \begin{bmatrix} e_1e_1 & e_1e_2 & \cdots & e_1e_n \\ e_2e_1 & e_2e_2 & \cdots & e_2e_n \\ \vdots & \vdots & \ddots & \vdots \\ e_ne_1 & e_ne_2 & \cdots & e_ne_n \end{bmatrix}$$

where the overbar represents the expected value (same as $E(\ )$).

Therefore the background error covariance matrix is a huge $n \times n$ matrix where $n$ can easily be $10^7$ for an operational NWP model!

A covariance matrix is symmetric and positive definite (i.e., $x^TAx > 0$ for all nonzero $x$).

The diagonal elements are the variances of the vector error components $\overline{e_ie_i} = \sigma_i^2$. 

5
If we normalize the covariance matrix, dividing each component by the product of the standard deviations, \( \frac{e_i e_j}{\sigma_i \sigma_j} = \text{corr}(e_i, e_j) = \rho_{ij} \), we obtain a correlation matrix

\[
C = \begin{bmatrix}
1 & \rho_{12} & \ldots & \rho_{1n} \\
\rho_{12} & 1 & \ldots & \rho_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
\rho_{1n} & \rho_{2n} & \ldots & 1
\end{bmatrix}
\]

and if \( D = \begin{bmatrix}
\sigma_1^2 & 0 & \ldots & 0 \\
0 & \sigma_2^2 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & \sigma_n^2
\end{bmatrix} \) is the diagonal matrix of the variances, then we can write

\[
P = D^{1/2} C D^{1/2}
\]

and \( D^{1/2} = \begin{bmatrix}
\sigma_1 & 0 & \ldots & 0 \\
0 & \sigma_2 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & \sigma_n
\end{bmatrix} \).

c) \( [AB]^T = B^T A^T \); \( [AB]^{-1} = B^{-1} A^{-1} \)
d) The general form of a quadratic function is $F(x) = \frac{1}{2}x^TAx + d^Tx + c$ where $A$ is a symmetric matrix, $d$ is a vector and $c$ a scalar.

To find the gradient of this scalar function $\nabla_x F = \frac{\partial F}{\partial x} = \left[ \begin{array}{c} \frac{\partial F}{\partial x_1} \\ \frac{\partial F}{\partial x_2} \\ \vdots \\ \frac{\partial F}{\partial x_n} \end{array} \right]$,

a column vector, we use the following properties of the gradient with respect to $x$:

\[ \nabla(d^Tx) = \nabla(x^Td) = d \] (since $\nabla_x x^T = I$, the identity matrix), or
\[ \nabla(d^Tx) = \nabla(d_1x_1 + d_2x_2 + \ldots d_n x_n) = \left[ \begin{array}{c} \frac{\partial (d_1x_1 + d_2x_2 + \ldots d_n x_n)}{\partial x_1} \\ \frac{\partial (d_1x_1 + d_2x_2 + \ldots d_n x_n)}{\partial x_2} \\ \vdots \\ \frac{\partial (d_1x_1 + d_2x_2 + \ldots d_n x_n)}{\partial x_n} \end{array} \right] = \left[ \begin{array}{c} d_1 \\ d_2 \\ \vdots \\ d_n \end{array} \right] = d. \]

And $\nabla(x^T Ax) = 2Ax$.

We can see this by taking, say, time derivative of $x$:
\[
\frac{d(x^T Ax)}{dt} = \dot{x}^T A x + x^T A \dot{x} = (\dot{x}^T A x)^T + x^T A \dot{x} = x^T A \dot{x} + x^T A \dot{x} = 2x^T A \dot{x}
\]

\[
d(x^T Ax) = 2x^T A dx = 2(Ax)^T dx
\]

Because the differential of function \( F(x) = F(x_1, x_2, \ldots, x_n) \) is

\[
dF = \frac{\partial F}{\partial x_1} dx_1 + \frac{\partial F}{\partial x_2} dx_2 + \ldots + \frac{\partial F}{\partial x_n} dx_n = \left[ \begin{array}{c}
\frac{\partial F}{\partial x_1} \\
\frac{\partial F}{\partial x_2} \\
\vdots \\
\frac{\partial F}{\partial x_n}
\end{array} \right] \begin{bmatrix}
dx_1 \\
dx_2 \\
\vdots \\
dx_n
\end{bmatrix} = \nabla F^T dx
\]

In our current case, \( F = x^T A x \), therefore

\[
\nabla (x^T A x) = 2Ax
\]

Therefore, \( F(x) = \frac{1}{2} x^T A x + d^T x + c \),

\[
\nabla F(x) = Ax + d, \quad \nabla^2 F(x) = A \quad \text{and} \quad \delta F = (\nabla F)^T \delta x
\]

The above will be used when derive the gradient of quadratic-form cost function with respective the analysis variable \( x \).
Statistical assumptions

We define the background error and the analysis error as vectors of length $n$:

$$
\varepsilon_b(x, y) = x_o(x, y) - x_i(x, y)
$$

$$
\varepsilon_a(x, y) = x_a(x, y) - x_i(x, y)
$$

The $p$ observations available at irregularly spaced points $y_o(r_i)$ have observational errors

$$
\varepsilon_{oi} = y_o(r_i) - y_o(r_i) = y_o(r_i) - H(x_i).
$$

We do not know the truth $x_i$, thus we do not know the errors of the available background and observations, but we can make a number of assumptions about their statistical properties.

The background and observations are assumed to be unbiased:

$$
E[\varepsilon_b(x, y)] = E[x_o(x, y)] - E[x_i(x, y)] = 0
$$

$$
E[\varepsilon_a(x, y)] = E[y_o(x)] - E[y_i(x)] = 0
$$

If the forecasts (background) and the observations are biased, in principle we can and should correct the bias before proceeding.
We define the **error covariance matrices** for the analysis, background and observations respectively:

\[
P_a = A = E\{\epsilon_a \epsilon_a^T\} \\
P_b = B = E\{\epsilon_b \epsilon_b^T\} \\
P_o = R = E\{\epsilon_o \epsilon_o^T\}
\]

The nonlinear observation operator \( H \) that transforms model variables into observed variables can be linearized as

\[
H(x + \delta x) = H(x) + H\delta x
\]

where \( H \) is a \( p \times n \) matrix with elements \( h_{ij} = \partial H_i / \partial x_j \) called the linear observation operator.

Let's see why \( H \) is defined as above. First, for a scalar function of multiple independent variables, such as

\[
y = y(x_1, x_2, \ldots, x_n) \quad \text{or written in a vector form} \quad y = y(x), \quad \text{using Tayler series expansion, we have}
\]

\[
y(x + \delta x) = y(x_1 + \delta x_1, x_2 + \delta x_2, \ldots, x_n + \delta x_n) \\
= y(x_1, x_2, \ldots, x_n) + \frac{\partial y}{\partial x_1} \delta x_1 + \frac{\partial y}{\partial x_2} \delta x_2 + \ldots + \frac{\partial y}{\partial x_n} \delta x_n + O(\delta x_1^2, \delta x_2^2, \ldots, \delta x_n^2)
\]

\[
= y(x) + \left[ \frac{\partial y}{\partial x_1} \quad \frac{\partial y}{\partial x_2} \quad \ldots \quad \frac{\partial y}{\partial x_n} \right] \delta x = y(x) + \nabla_y x^T \delta x
\]
Now the function we have a vector function $y = H(x)$ where $y$ is a vector of length $p$, the number of observations and $H$ is a vector representing $p$ number of functions. Writing out the vector function explicitly, we have

$$
\begin{bmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_p \\
\end{bmatrix} = 
\begin{bmatrix}
  H_1(x) \\
  H_2(x) \\
  \vdots \\
  H_p(x) \\
\end{bmatrix}
$$

therefore, using Taylor series expansion,

$$
y(x + \delta x) =
\begin{bmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_p \\
\end{bmatrix} =
\begin{bmatrix}
  H_1(x + \delta x) \\
  H_2(x + \delta x) \\
  \vdots \\
  H_p(x + \delta x) \\
\end{bmatrix} =
\begin{bmatrix}
  \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \ldots & \frac{\partial y_1}{\partial x_n} \\
  \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \ldots & \frac{\partial y_2}{\partial x_n} \\
  \vdots & \vdots & \ddots & \vdots \\
  \frac{\partial y_p}{\partial x_1} & \frac{\partial y_p}{\partial x_2} & \ldots & \frac{\partial y_p}{\partial x_n} \\
\end{bmatrix}
\begin{bmatrix}
  \delta x_1 \\
  \delta x_2 \\
  \vdots \\
  \delta x_n \\
\end{bmatrix} + O(\delta x^2) 
\approx H(x) + H\delta x
$$

and the elements of matrix $H$ are $h_{i,j} = \frac{\partial H_i}{\partial x_j}$, where $i = 1, 2, \ldots, p$, and $j = 1, 2, \ldots, n$. In practice, $H$ is often a sparse matrix because each element of $y$ is often not dependent on all elements of $x$. For example, if $H$ operator is a bilinear interpolation operator, only 4 instead of $n$ grid point values are involved, i.e., this particular $y$ element here is only a function of 4 elements of $x$. In the example of radar reflectivity, only three variables are involved instead of 10 or model variables (no u, v, w, T or p).
We also assume that the background (usually a model forecast) is a good approximation of the truth, so that the analysis and the observations are equal to the background values plus small increments. Therefore, the innovation vector $d$ defined earlier can be written as

$$d = y_o - H(x_b) = y_o - H[x_i + (x_b - x_i)]$$

$$= y_o - H(x_i) - H(x_b - x_i) = \epsilon_o - H\epsilon_b$$

The $H$ matrix (a $p \times n$ matrix) transforms vectors in model space (vectors of length $n$) into their corresponding values in observation space (vectors of length $p$). Its transpose or adjoint $H^T$ (an $n \times p$ matrix) transforms vectors in observation space (vectors of length $p$) to vectors in model space (vectors of length $n$).

The background error covariance $B$ (a matrix of size $n \times n$) and the observation error covariance $R$ (a matrix of size $p \times p$) are assumed to be known.

In addition, we assume that the observation and background errors are uncorrelated.

The analysis error covariance $P_a$ is not known, but we will minimize it through the optimal choice of weights.

**Derivation for analysis error covariance**

In order to be able to minimize the analysis error, we first derive the formula for analysis error covariance.

From the above, $d = y_o - H(x_b) = \epsilon_o - H\epsilon_b$ and $x_a = x_b + W(y_o - H(x_b)) = x_b + Wd$, and therefore the analysis error is

$$\epsilon_a = x_a - x_i = (x_b + Wd) - x_i = (x_b - x_i) + Wd = \epsilon_b + W(\epsilon_o - H\epsilon_b)$$
(so we wrote the analysis errors in terms of the background error and observation error)

We want to choose the weight matrix \( W \) (a.k.a. gain matrix \( K \)) so as to minimize the analysis error covariance (note that minimizing the analysis error covariance is the key of optimal statistical analysis):

\[
P_a = E\{ (x_a - x_i)(x_a - x_i)^T \} = E\{ [\varepsilon_b + W(\varepsilon_o - He_b)][\varepsilon_b + W(\varepsilon_o - He_b)]^T \}
\]

Expanding the above equation and using the rules for matrix multiplication from the earlier remarks, we get

\[
P_a = E\{ \varepsilon_o \varepsilon_o^T \} = E\{ [\varepsilon_b + W(\varepsilon_o - He_b)][\varepsilon_b^T + (\varepsilon_o - He_b)^T W^T ] \}
\]

\[
= E\{ \varepsilon_b \varepsilon_b^T + \varepsilon_b (\varepsilon_o - He_b)^T W^T + W(\varepsilon_o - He_b) \varepsilon_b^T + W(\varepsilon_o - He_b)(\varepsilon_o - He_b)^T W^T \}
\]

\[
= E\{ \varepsilon_b \varepsilon_b^T \} + E\{ 2\varepsilon_b (\varepsilon_o - He_b)^T W^T \} + E\{ W(\varepsilon_o - He_b)(\varepsilon_o - He_b)^T W^T \}
\]

In the above, we used the symmetry of \( W(\varepsilon_o - He_b)\varepsilon_b^T \).

The optimal weight matrix is obtained by differentiating each element of \( P_a \) with respect to each element of \( W \) and setting the result to zero.

This gives the following equation for the weights:

\[
E\{ \varepsilon_b (\varepsilon_o^T - \varepsilon_b^T H^T ) \} + WE\{ ((\varepsilon_o - He_b)(\varepsilon_o^T - \varepsilon_b^T H^T )\} = 0
\]

or, carrying out the matrix multiplications and using the rules for transpose of a product,

\[
E\{ \varepsilon_b \varepsilon_o^T \} - E\{ \varepsilon_b \varepsilon_b^T \}H^T + WE\{ \varepsilon_o \varepsilon_o^T \} - HE\{ \varepsilon_b \varepsilon_o^T \} - E\{ \varepsilon_b \varepsilon_b^T \}H^T + HE\{ \varepsilon_b \varepsilon_b^T \}H^T = 0
\]
Recall that we assumed that the *background errors are not correlated with the observational errors*, i.e., that their covariance is equal to zero. Replacing the definitions of background error covariance $\mathbf{B}$ and observational error covariance $\mathbf{R}$ into above we obtain an equation for the optimal weights, $\mathbf{W}(\mathbf{R} + \mathbf{HBH}^T) = \mathbf{BH}^T$, so that the optimal weight matrix is given by

$$\mathbf{W} = \mathbf{BH}^T (\mathbf{R} + \mathbf{HBH}^T)^{-1}. $$

Finally, we obtain the analysis error covariance:

$$\mathbf{P}_a = \mathbf{B} - \mathbf{BH}^T \mathbf{W}^T - \mathbf{WHB} + \mathbf{WRW}^T + \mathbf{WHB}^T \mathbf{W}^T$$

and substituting in $\mathbf{W}$ obtained above, we have

$$\mathbf{P}_a = \mathbf{B} - \mathbf{BH}^T \mathbf{W}^T - \mathbf{WHB} + \mathbf{W}(\mathbf{R} + \mathbf{HBH}^T)\mathbf{W}^T$$

$$= \mathbf{B} - \mathbf{BH}^T \mathbf{W}^T - \mathbf{WHB} + \mathbf{BH}^T \mathbf{W}^T$$

$$= (\mathbf{I} - \mathbf{WH})\mathbf{B}$$

For convenience, we repeat the basic equations of OI, and express in words their interpretation:

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{W}[\mathbf{y}_a - H(\mathbf{x}_b)] = \mathbf{x}_b + \mathbf{Wd} \quad (1)$$

$$\mathbf{W} = \mathbf{BH}^T (\mathbf{R} + \mathbf{HBH}^T)^{-1} \quad (2a)$$
We will see later when we derive the variational approach or 3D-Var that the weight matrix (2a) can be written in an alternative equivalent form as

\[ W = (B^{-1} + H^T R^{-1} H)^{-1} H^T R^{-1} \]  

(2b)

\[ P_a = (I_n - WH)B \]  

(3)

where the subindex \( n \) is a reminder that the identity matrix is in the analysis or model space.

The interpretation of these equations is:

Equation (1) says:  
\[ (x_a = x_b + W[y_o - H(x_b)] = x_b + Wd) \]

"The analysis is obtained by adding to the first guess (background) the product of the optimal weight (or gain) matrix and the innovation (difference between the observation and first guess).

The first guess of the observations is obtained by applying the observation operator \( H \) to the background vector."

Also, note that from earlier, \( H(x_b) = H(x_i) + H(x_b - x_i) = H(x_i) + H\epsilon_b \), where the matrix \( H \) is the linear tangent perturbation of \( H \).

Equation (2a) says:  
\[ W = BH^T (R + HBH^T)^{-1} \]

"The optimal weight (or gain) matrix is given by the background error covariance in the observation space (\( BH^T \) - actually the background error covariances between the grid points and observation points) multiplied by the inverse
of the total error covariance (sum of the background and the observation error covariances).” In the latter case, the background error is brought into the observation space by the $H$ operator and its adjoint (transpose), therefore $HBH^T$ present that background error covariances between background values at observation points.

Note that the larger the background error covariance compared to the observation error covariance, the larger the correction to the first guess.

**Equation (3) says:** \( P_a = (I_n - WH)B \)

"The error covariance of the analysis is given by the error covariance of the background, reduced by a matrix equal to the identity matrix \((n \times n)\) minus the optimal weight matrix".

Finally we derive an alternative formulation for the analysis error covariances showing the additive properties of the "precision " (if all the statistical assumptions hold true):

From equations $\varepsilon_a = x_a - x_i = x_b + Wd - x_i = \varepsilon_b + W(\varepsilon_o - He_b)$ and $W = (B^{-1} + H^T R^{-1} H)^{-1} H^T R^{-1}$ (obtained using the 3DVAR approach later), we can show that

$$\varepsilon_a = \varepsilon_b + [B^{-1} + H^T R^{-1} H]^{-1} H^T R^{-1} (\varepsilon_o - He_b)$$

$$= [B^{-1} + H^T R^{-1} H]^{-1} \{ \varepsilon_b [B^{-1} + H^T R^{-1} H] + H^T R^{-1} (\varepsilon_o - He_b) \}$$

$$= [B^{-1} + H^T R^{-1} H]^{-1} [B^{-1} \varepsilon_b + H^T R^{-1} \varepsilon_o]$$

If we compute again $P_a = E\{\varepsilon_a \varepsilon_a^T\}$ from the above, and make use of $E\{\varepsilon_b \varepsilon_o^T\} = 0$, $P_b = B = E\{\varepsilon_b \varepsilon_b^T\}$, $P_o = R = E\{\varepsilon_o \varepsilon_o^T\}$, we obtain

$$P_a^{-1} = B^{-1} + H^T R^{-1} H . \quad (4)$$
Here is proof (noting that $B^{-1} + H^T R^{-1} H$ is symmetric):

\[
P_a = E\{[B^{-1} \varepsilon_b + H^T R^{-1} \varepsilon_o][B^{-1} \varepsilon_b + H^T R^{-1} \varepsilon_o]^T\} \\
= E\{[B^{-1} \varepsilon_b + H^T R^{-1} \varepsilon_o][\varepsilon_b^T B^{-1} + \varepsilon_o^T R^{-1} H]\} \\
\frac{[B^{-1} + H^T R^{-1} H][B^{-1} + H^T R^{-1} H]^T}{[B^{-1} + H^T R^{-1} H][B^{-1} + H^T R^{-1} H]^T} \\
= B^{-1} E\{\varepsilon_b \varepsilon_b^T\} B^{-1} + H^T R^{-1} E\{\varepsilon_o \varepsilon_o^T\} R^{-1} H \\
\frac{[B^{-1} + H^T R^{-1} H][B^{-1} + H^T R^{-1} H]^T}{[B^{-1} + H^T R^{-1} H][B^{-1} + H^T R^{-1} H]^T} \\
= B^{-1} B^{-1} + H^T R^{-1} R^{-1} H \\
\frac{[B^{-1} + H^T R^{-1} H][B^{-1} + H^T R^{-1} H]^T}{[B^{-1} + H^T R^{-1} H][B^{-1} + H^T R^{-1} H]^T} \\
= 1 \\
\frac{[B^{-1} + H^T R^{-1} H]^T}{[B^{-1} + H^T R^{-1} H]^T}
\]

Equation (4) says:

"The analysis precision, defined as the inverse of the analysis error covariance, is the sum of the background precision and the observation precision projected into the model space."

Note that all these statements are dependent on the assumption that the statistical estimates of the errors are accurate. If

1) the observations and/or background error covariances are poorly known,
2) there are biases, or
3) the observations and background errors are correlated,
the analysis accuracy can be considerably worse than implied by (3) or (4).
Simple Examples

Let’s look at a few simple examples of I/O analysis.

**Example 1:** Consider a scalar problem. We have one observation $y$ which is located at the analysis point. We also have a background estimate $x_b$. In addition, we assume that we know the error statistics of $y$ and $x_b$:

\[
\begin{align*}
&< x_b - x_t > = 0 \\
&< (x_b - x_t)^2 > = \sigma_b^2 \\
&< (y - y_t)(x_b - x_t) > = 0 \\
&< y - y_t > = 0 \\
&< (y - y_t)^2 > = \sigma_r^2
\end{align*}
\]

Using the notion of O/I solution, $n = 1, \ p = 1, \ H = (1), \ R = (\sigma_r^2), \ B = (\sigma_b^2)$.

\[
(x_a) = x_b + \frac{(1)(\sigma_r^2)^{-1}[((y) - (1)(x_b))]}{(\sigma_b^2)^{-1} + (1)(\sigma_r^2)^{-1}(1)}
\]

\[
x_a = x_b + \frac{\sigma_b^2}{\sigma_b^2 + \sigma_r^2} (y - x_b)
\]

\[
(\sigma_a^2)^{-1} = (\sigma_b^2)^{-1} + (\sigma_r^2)^{-1}
\]

Or

\[
\frac{x_a}{\sigma_a^2} = \frac{x_b}{\sigma_b^2} + \frac{y}{\sigma_r^2}.
\]
We see that the analysis is a linear combination of the background and observation, with the weighting coefficients proportional to the inverse of variances. The variance of the analysis is smaller than that of both background and observations, i.e., $\sigma_a^2 < \min[\sigma_b^2, \sigma_r^2]$.

Now assume that the analysis is for temperature and we have

$$y = 22.0, \ R = (\sigma_t)^2 = 1.0$$

$$x_b = 20.5, \ B = (\sigma_b)^2 = 2.0$$

and

$$x_a = 21.5, \ A \approx 0.67 (\sigma_a \approx 0.81).$$

The following figure shows that background, observation and analysis probability distribution functions assuming the errors have Gaussian distributions.

It is clear that the analysis is between the background and observation, and is in this particular case closer to the observation because of its smaller expected errors.

It is also shown that the analysis has a higher probability and smaller expected errors compared to both the background and the observation.
Probability distribution function for analysis, background and observation
Example 2. Now, suppose we have one observation, $y$, located between two analysis points. We have background information on the two analysis points, denoted by $x_{b1}$ and $x_{b2}$, and we can linearly interpolate the background information to the observation location as

$$
Hx_b = \alpha x_{b1} + (1 - \alpha) x_{b2}
$$

where $0 \leq \alpha \leq 1$.

The assumed observation error is as given before,

$$
R = (\sigma_i)^2,
$$

while the background error covariance matrix now takes the form of

$$
B = \begin{pmatrix}
\sigma^2_{b1} & \sigma_{b12} \\
\sigma_{b12} & \sigma^2_{b2}
\end{pmatrix} = \sigma^2_{b} \begin{pmatrix}
1 & \eta \\
\eta & 1
\end{pmatrix}.
$$

Here we have assumed that

$$
b_{11} = b_{22} = \sigma^2_{b},
$$

$$
b_{21} = b_{12} = \sigma^2_{b} \eta
$$

with $\eta$ being the error correlation coefficient.

The O/I solution is then
\[
\begin{pmatrix}
x_{a1} \\
x_{a2}
\end{pmatrix} = \begin{pmatrix}
x_{b1} \\
x_{b2}
\end{pmatrix} + \sigma_b^2 \left( \alpha + \eta (1 - \alpha) \right) \frac{y - [\alpha x_{b1} + (1 - \alpha) x_{b2}]}{\left[ \alpha^2 + 2 \alpha (1 - \alpha) \eta + (1 - \alpha)^2 \right] \sigma_b^2 + \sigma_r^2}.
\]

(5)

Let’s consider three cases:

Case 1. The observation is collocated with analysis grid point 1 (\(\alpha=1\)) and the background errors are not correlated between points 1 and 2 (\(\eta=0\)). The above solution now reduces to

\[
\begin{pmatrix}
x_{a1} \\
x_{a2}
\end{pmatrix} = \begin{pmatrix}
x_{b1} \\
x_{b2}
\end{pmatrix} + \sigma_b^2 \frac{1}{\sigma_b^2 + \sigma_r^2} y - x_{b1}, \quad \Rightarrow
\]
\[
x_{a1} = x_{b1} + \sigma_b^2 \frac{y - x_{b1}}{\sigma_b^2 + \sigma_r^2},
\]
\[
x_{a2} = x_{b2} + 0.
\]

In this case, the solution at point 1 is identical to that in example 1. The solution at point 2 is equal to the background and no information from the observation is added there.

Case 2. The observation is collocated with analysis at grid 1 (\(\alpha=1\)) as in case 1. However, the background errors are correlated between point 1 and point 2 (\(\eta \neq 0\)). The solution now reduces to

\[
\begin{pmatrix}
x_{a1} \\
x_{a2}
\end{pmatrix} = \begin{pmatrix}
x_{b1} \\
x_{b2}
\end{pmatrix} + \sigma_b^2 \frac{1}{\eta \sigma_b^2 + \sigma_r^2} y - x_{b2}, \quad \Rightarrow
\]
\[
x_{a1} = x_{b1} + \sigma_b^2 \frac{y - x_{b1}}{\sigma_b^2 + \sigma_r^2},
\]
\[
x_{a2} = x_{b2} + \eta \sigma_b^2 \frac{y - x_{b1}}{\sigma_b^2 + \sigma_r^2}.
\]
In this case, the solution at point 1 is unchanged from case 1, but the solution at point 2 is equal to the background plus $\eta$ times the analysis increment added to point 1 – now we can see the role of background error correlation in spreading observational information (or more strictly the analysis increment).

**Case 3.** The observation is located inbetween of the analysis points ($\alpha \neq 1$) but the background errors are not correlated between points 1 and 2 ($\eta=0$). Now the solution becomes

$$
\begin{bmatrix}
  x_{a1} \\
  x_{a2}
\end{bmatrix} = \begin{bmatrix}
  x_{b1} \\
  x_{b2}
\end{bmatrix} + \sigma_b \begin{bmatrix}
  \alpha \\
  1-\alpha
\end{bmatrix} \left[ \frac{y - [\alpha x_{b1} + (1-\alpha) x_{b2}]}{\alpha^2 + (1-\alpha)^2 \sigma_b^2 + \sigma_r^2} \right].
$$

In this case, the analysis increments for point 1 and point 2 are proportional to $\alpha$ and $1-\alpha$, respectively.

**Final Comments:**

from the full solution (5), we can see that both the observation operator and error correlation have made contributions. However, when generalize the solution from two analysis points to $n$ points, the linear interpolation operator will only influence the analysis points around the observation, while the error correlations may spread information to all analysis points as long as the error correlation between the point is non-zero.
Approximations with Practical O/I Implementation

- The O/I analysis is given by
  \[ x_a = x_b + W[y_a - H(x_b)] \]
  \[ W = BH^T (R + HBH^T)^{-1} \]
  
- The actual implementation requires simplifications in the computation of the weight \( W \).

- The equation for \( x_a \) can be regarded as a list of scalar analysis equations, one per model variable in the vector \( x \).

- For each model variable the analysis increment is given by the corresponding line of \( W \) times the vector of background departures \( (y - H(x_b)) \).

- The fundamental hypothesis in OI is: *For each model variable, only a few observations are important in determining the analysis increment.*

Based on this assumption, the problem of matrix product and inversion is reduced by including only a smaller number of observations for the analysis at a given grid point. The following two figures show two data selection strategies.
Figure 9. One OI data selection strategy is to assume that each analysis point is only sensitive to observations located in a small vicinity. Therefore, the observations used to perform the analysis at two neighbouring points \( x_1 \) or \( x_2 \) may be different, so that the analysis field will generally not be continuous in space. The cost of the analysis increases with the size of the selection domains.
Figure 10. A slightly more sophisticated and more expensive OI data selection is to use, for all the points in an analysis box (black rectangle), all observations located in a bigger selection box (dashed rectangle), so that most of the observations selected in two neighbouring analysis boxes are identical.
The actual implemented can be as follows:

1) For each model variable $x_i$, select a small number of $p_i$ observations using empirical selection criteria.

2) Form the corresponding list of background departures $(y - H(x_b))_i$,

3) Form the $p_i$ background error covariances between the model variable $x_i$ and the model state interpolated at the $p_i$ observation points (i.e. the relevant $p_i$ coefficients of the $i$-th line of $BH^T$), and

4) Form the $p_i \times p_i$ background and observation error covariance submatrices formed by the restrictions of $HBH^T$ and $R$ to the selected observations.

5) Invert the positive definite matrix formed by the restriction of $(HBH^T + R)$ to the selected observations,

6) Multiply it by the $i$-th line of $BH^T$ to get the necessary line of $W$.

It is possible to save some computer time on the matrix inversion by solving directly a symmetric positive linear system, since we know in advance the vector of departures to which the inverse matrix will be applied. Also, if the same set of observations is used to analyze several model variables, then the same matrix inverse (or factorization) can be reused.
Models of Error Covariances

Correct specification of background and observation error covariances is crucial – they determine the relative weight of background and observations.

- Variances are essential for determining the magnitude of errors therefore the relative weight
- Covariance determines how observation information is spread in model space (when the model resolution does not match that of observation)

Observation error variance – include instrument errors and representativeness errors.
- Systematic observation biases should be removed before using the data

Observation error correlation/covariance – often assumed zero, i.e., measurements are assumed uncorrelated.
- Observation error correlation can show up when
  - Sets of observations are taken by the same platform, e.g., radar, rawinsonde, aircraft, satellite
  - Preprocessing that introduce systematic errors
  - Representative errors of close-by observations
  - Error of the forward operator, e.g., interpolator that contains similar errors
- The presence of (positive) observation error correlation reduces the weight given to the average of the observations – reasonable because these observations are alike
- Observation error correlations are difficult to estimate and account for. In practice, efforts are made to minimize them through reducing bias, by
  - Avoiding unnecessary preprocessing
- By thinning dense data (denser than grid resolution)
- By improving model and observation operators (model plays the role of forward operator in the case of 4DVAR)

- After these are done, it is safer to assume the observation correlation to be zero, i.e., the observation error covariance matrix $R$ is diagonal.

**Background error variances** – they are usually estimates of the error variances in the forecast used to produce $x_b$.

- This is a difficult problem, because they are never observed directly – they can only be estimated in a statistical sense.

- If the analysis is of good quality (i.e. if there are a lot of observations) an estimate can be provided by the variance of the differences between the forecast and a verifying analysis.

- If the observations can be assumed to be uncorrelated, much better averaged background error variances can be obtained by using the observational method (or the Hollingworth-Lonnberg method – Tellus 1986).

- However, in a system like the atmosphere the actual background errors are expected to depend a lot on the weather situation, and ideally the background errors should be flow-dependent. This can be achieved by the Kalman filter, by 4D-Var to some extent, or by some empirical laws of error growth based on physical grounds.

- If background error variances are badly specified, it will lead to too large or too small analysis increments.

- In least-squares analysis algorithms, only the relative magnitude of the background and observation error variances is important.
Background error correlations – they are essential because

- **Information spreading.**
  
  o **In data-sparse areas**, the shape of the analysis increment is completely determined by the covariance structures.

  o The correlations in $B$ will perform the spatial spreading of information from the observation points to a finite domain surrounding it.

- **Information smoothing.**
  
  o **In data-dense areas**, the amount of smoothing of the observed information is governed by the correlations in $B$, which can be understood by noting that the left most term in $W$ is $B$.

  o The smoothing of the increments is important in ensuring that the analysis contains scales which are statistically compatible with the smoothness properties of the physical fields.

    ▪ For instance, when analysing stratospheric or anticyclonic air masses, it is desirable to smooth the increments a lot in the horizontal in order to average and spread efficiently the measurements.

    ▪ When doing a low-level analysis in frontal, coastal or mountainous areas, or near temperature inversions, it is desirable on the contrary to limit the extent of the increments so as not to produce an unphysically smooth analysis. This has to be reflected in the specification of background error correlations.

  o Both of the above address only the issue of correlations among same variables, or **auto-correlations**.
• **Balance properties** – correlation across variables, or **cross-correlations**

  o  There are often more degrees of freedom in a model than in reality (i.e., not all model variables are free from each other). For instance, the large-scale atmosphere is usually hydrostatic and is almost geostrophic – these relationships introduce **balances among the fields**

  o  These balance properties show up as correlations in the background errors

  o  Because of these correlations / balances, observations can be used more effectively, i.e., observing one model variable yields information about all variables that are balanced with it.

    ▪  For example, a low-level wind observation allows one to correct the surface pressure field by assuming some amount of geostrophy.

    ▪  When combined with the spatial smoothing of increments this can lead to a considerable impact on the quality of the analysis, e.g. a properly spread observation of geopotential height can produce a complete three-dimensional correction to the geostrophic wind field (see Figure ).

    ▪  The relative amplitude of the increments in terms of the various model fields will depend directly on the specified amount of correlation as well as on the assumed error variance in all the concerned parameters.

  o  Accurate estimation and use of background error (cross-) correlations can do the magic of ‘retrieving’ quantities not directly observed, a thing that ensemble Kalman filter attempts to do in, e.g., the assimilation of radar data.

  o  Accurate background errors are flow-dependent
Figure 7. Example of horizontal structure functions commonly used in meteorology: the horizontal autocorrelation of height (or pressure) has an isotropic, gaussian-like shape as a function of distance (right panel). In turn, geostrophy implies that wind will be cross-correlated with height at distances where the gradient of height correlation is maximum. Hence, an isolated height observation will generate an isotropic height “bump” with a rotating wind increment in the shape of a ring.
The observational (or Hollingworth–Lonnberg) method for estimating background error

- This method relies on the use of background departures \( y - H(x_b) \) in an observing network that is dense and large enough to provide information on many scales, and that can be assumed to consist of uncorrelated and discrete observations.

- The principle (illustrated in Fig. 8) is to calculate an histogram of background departure covariances, stratified against separation (for instance).

- At zero separation the histogram provides averaged information about the background and observation errors, at nonzero separation it gives the averaged background error correlation.

![Diagram](image)

Figure 8. Schematic representation the observational method. The (observation – background) covariance statistics for a given assimilation system are stratified against distance, and the intercept at the origin of the histogram provides an estimate of the average background and observation error variances for these particular assimilation and observation systems.
In most systems the background error covariances should go to zero for very large separations. If this is not the case, it is usually the sign of biases in the background and/or in the observations and the method may not work correctly (Hollingsworth and Lonnberg 1986.).

Reference:

The NMC method for estimating background errors

- The so-called "NMC method" (Parrish and Derber, 1992) estimates the forecast error covariance according to

$$\mathbf{B} = \alpha E\{[x_f(48\text{hr}) - x_f(24\text{hr})][x_f(48\text{hr}) - x_f(24\text{hr})]^T\},$$

i.e., the structure of the forecast or background error covariance is estimated as the average over many (e.g., 50) differences between two short-range model forecasts verifying at the same time. The magnitude of the covariance is then appropriately scaled.

- In this approximation, rather than estimating the structure of the forecast error covariance from differences with observations, the model-forecast differences themselves provide a multivariate forecast difference covariance.

- The forecast covariance strictly speaking is the covariance of the forecast differences and is only a proxy of the structure of forecast errors. Nevertheless, it has been shown to produce better results than the observational method. An important reason is that the rawinsonde network (used by the previous observational method) does not have enough density to allow a proper estimate of the global structures.

- The ‘NMC method’ has been in use at most operational centers because of its simplicity and comparative effectiveness.

- Being based on many past forecasts (over e.g., 1-2 months), the estimate is at best season dependent, however.

Reference:
The Modeling of background correlations

- The full $\mathbf{B}$ matrix is usually too big to be specified explicitly. The variances are the $n$ diagonal terms of $\mathbf{B}$, which are usually specified completely.
  - The off-diagonal terms are more difficult to specify. They must generate a symmetric positive definite matrix.
  - Additionally $\mathbf{B}$ is often required to have some physical properties which are required to be reflected in the analysis:
    - the correlations must be smooth in physical space, on sensible scales,
    - the correlations should go to zero for very large separations if it is believed that observations should only have a local effect on the increments,
    - the correlations should not exhibit physically unjustifiable variations according to direction or location,
    - the most fundamental balance properties, like geostrophy, must be reasonably well enforced.
    - the correlations should not lead to unreasonable effective background error variances for any parameter that is observed, used in the subsequent model forecast, or output to the users as an analysis product.

- The complexity and subtlety of these requirements mean that the specification of background error covariances is a problem similar to physical parameterization. Some of the more popular techniques are listed below.
  - Correlation models can be specified independently from variance fields, under the condition that the
scales of variation of the variances are much larger than the correlation scales (remember $P = D^{1/2}CD^{1/2}$).

- Vertical autocorrelation matrices for each variable are usually small enough to be specified explicitly.

- Horizontal autocorrelations cannot be specified explicitly, but they can be reduced to sparse matrices by assuming that they are homogeneous and isotropic to some extent.

- Three-dimensional multivariate correlation models can be built by carefully combining separability, homogeneity and independency hypotheses like: zero correlations in the vertical for distinct spectral wavenumbers, homogeneity of the vertical correlations in the horizontal and/or horizontal correlations in the vertical, property of the correlations being products of horizontal and vertical correlations. Numerically they imply that the correlation matrix is sparse because it is made of block matrices which are themselves block-diagonal.

- Balance constraints can be enforced by transforming the model variables into suitably defined complementary spaces of balanced and unbalanced variables. The latter are supposed to have smaller background error covariances than the former, meaning that they will contribute less to the increment structures.

- The geostrophic balance constraint can be enforced using the classical $f$-plane or $\beta$-plane balance equations, or projections onto subspaces spanned by so-called Rossby and Gravity normal modes.

- More general kinds of balance properties can be expressed using linear regression operators calibrated on actual background error fields, if no analytical formulation is available.
Comment on O/I (and 3DVAR) versus other schemes

- Perhaps the most important advantage of statistical interpolation schemes such as Optimal Interpolation and 3D-Var over empirical schemes such as successive correction method (SCM), is that the correlation between observational increments is taken into account.

- With SCM, the weights of the observational increments depend only on their distance to the grid point. Therefore, if a number of observations are "bunched up" in one quadrant, with just a single observation in a different quadrant, then all the observations will be given similar weight. In Optimal Interpolation (or 3D Var), by contrast, the isolated observational increment will be given more weight in the analysis than observations that are close together and therefore less independent.

- When several observations are too close together, then the OI solution becomes an ill-posed problem. In those cases, it is common to compute a "super-observation" combining the close individual observations. This has the advantage of removing the ill posedness, while at the same time reducing by averaging the random errors of the individual observations. The super-observation should be a weighted average that takes into account the relative observation errors of the original close observations.
The role of observation operator $H$

Earlier, we said that

- The $H$ matrix (a $p \times n$ matrix) transforms vectors in model space (e.g., $x$, which is a vector of length $n$) into their corresponding values in observation space (vectors of length $p$).

- The transpose or adjoint of $H$, $H^T$, (an $n \times p$ matrix) transforms vectors in observation space (e.g., $y$, a vector of length $p$) to vectors in model space (vectors of length $n$).

The observation operator and its adjoint can also operate on the error covariance matrices, $B$ and $R$, and when they do so, they have similar effect as on vectors.

For example, we indicated earlier that $BH^T$ is the background error covariances between the grid points and observation points, therefore $H^T$ plays the role of ‘taking background error from grid point to observational points, partially though because the product still represents correlations between grid and observation points’.

The background error is completely brought into the observation space by the $H$ operator and its adjoint (transpose), so that $HBH^T$ represents background error covariances between observation points.

**Illustration of the point**

Suppose we have three observations, $y_1$, $y_2$ and $y_3$ taken between two grid points with background values of $x_1$ and $x_2$:

```
  x1 -----> y1 -------> y2 -------> y3 -----> x2
```
The forward operator is simply the linear interpolator, so that

\[
H_x = \begin{bmatrix}
\alpha_1 & 1 - \alpha_1 \\
\alpha_2 & 1 - \alpha_2 \\
\alpha_3 & 1 - \alpha_3
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix}
\equiv
\begin{bmatrix}
\alpha_1 & \beta_1 \\
\alpha_2 & \beta_2 \\
\alpha_3 & \beta_3
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix}.
\]

The background error covariance matrix \( B \) is

\[
B = \begin{bmatrix}
b_{11} & b_{12} \\
b_{21} & b_{22}
\end{bmatrix}
\]

therefore

\[
BH^T = \begin{bmatrix}
b_{11} & b_{12} \\
b_{21} & b_{22}
\end{bmatrix}
\begin{bmatrix}
\alpha_1 & \alpha_2 & \alpha_3 \\
\beta_1 & \beta_2 & \beta_3
\end{bmatrix}
= \begin{bmatrix}
b_{11}\alpha_1 + b_{12}\beta_1 & b_{11}\alpha_2 + b_{12}\beta_2 & b_{11}\alpha_3 + b_{12}\beta_3 \\
b_{21}\alpha_1 + b_{22}\beta_1 & b_{21}\alpha_2 + b_{22}\beta_2 & b_{21}\alpha_3 + b_{22}\beta_3
\end{bmatrix}.
\]

Indeed, the background error covariances have been ‘interpolated’ by the observation operator (interpolator in this case).

For example, the first element of \( BH^T \), \( b_{11}\alpha_1 + b_{12}\beta_1 \), represents the interpolation of the background error covariance between \( x_1 \) point with itself (\( b_{11} \)), and the background error covariance between \( x_1 \) point and \( x_2 \) point (\( b_{12} \)) to the \( y_1 \) observation point, using interpolation coefficients \( \alpha_1 \) and \( \beta_1 \). Let’s denote the matrix
\[
\text{BH}^T = \begin{bmatrix}
  c_{11} & c_{12} & c_{13} \\
  c_{21} & c_{22} & c_{23}
\end{bmatrix}.
\]

The first index of \(c\) denotes grid point location and second index indicates the observation point.

Applying \(H\) operator again to \(\text{BH}^T\) takes the ‘other grid point end’ of covariances also to the observation points, so that we are left with covariances between observation points, but still of background errors.

\[
\text{HBH}^T = \begin{bmatrix}
  \alpha_1 & \beta_1 \\
  \alpha_2 & \beta_2 \\
  \alpha_3 & \beta_3
\end{bmatrix} \begin{bmatrix}
  c_{11} & c_{12} & c_{13} \\
  c_{21} & c_{22} & c_{23}
\end{bmatrix}.
\]

\[
= \begin{bmatrix}
  \alpha_1 c_{11} + \beta_1 c_{21} & \alpha_1 c_{12} + \beta_1 c_{22} & \alpha_1 c_{13} + \beta_1 c_{23} \\
  \alpha_2 c_{11} + \beta_2 c_{21} & \alpha_2 c_{12} + \beta_2 c_{22} & \alpha_2 c_{13} + \beta_2 c_{23} \\
  \alpha_3 c_{11} + \beta_3 c_{21} & \alpha_3 c_{12} + \beta_3 c_{22} & \alpha_3 c_{13} + \beta_3 c_{23}
\end{bmatrix} = \begin{bmatrix}
  d_{11} & d_{12} & d_{13} \\
  d_{21} & d_{22} & d_{23} \\
  d_{31} & d_{32} & d_{33}
\end{bmatrix}.
\]

Here, covariances \(c_{ij}\) are interpolated again, from grid point (indicated by the first index) to the observational point. For example, \(d_{12}\) represents background error covariance between \(y_1\) and \(y_2\) point, and is equal to the interpolated value (using weight \(\alpha_1\) and \(\beta_1\)) of the covariance between \(x_1\) point and \(y_1\) point and the covariance between \(x_2\) point and \(y_1\) point.