METR 5303 – Lecture #4

Biased observations

In the least squares derivation in the previous lecture, we assumed unbiased observations. Here we will show how to account for bias, that is, when $\langle \varepsilon_n \rangle \neq 0$, the obs. are said to be biased, and thus there will be a bias in the analysis s_a .

Fortunately, if the bias is known, it is easy to produce unbiased estimates.

We start with eq. (17) from the previous derivation and replace s_n with $s_n - \langle \varepsilon_n \rangle$, i.e., correct the nth observation s_n with its known bias,

[Note that this redefines the error variance σ_n^2 to be $\sigma_n^2 = \langle \varepsilon_n^2 \rangle - \langle \varepsilon_n \rangle^2$

So, eq. (17), $I = \frac{1}{2}\sum_{n=1}^{N} \sigma_n^{-2}(s_a - s_n)^2$, is

$$I_b = \frac{1}{2}\sum_{n=1}^{N} \sigma_n^{-2}\left[s_a - (s_n - \langle \varepsilon_n \rangle)\right]^2 \qquad (23)$$

Minimization of I_b w.r.t. s_a yields

$$s_a = \frac{\sum_{n=1}^{N} \sigma_n^{-2}(s_n - \langle \varepsilon_n \rangle)}{\sum_{n=1}^{N} \sigma_n^{-2}} \qquad (24)$$

To check this result, we should examine the bias of the analysis estimate to see if it is zero.

Since the bias of the analysis estimate ε_a is defined by $< \varepsilon_a > = < s_a - s >$, then we can use eq. (24) to write

$$\langle \varepsilon_a \rangle = \left\langle \frac{\sum_{n=1}^{N} \sigma_n^{-2} (s_n - \langle \varepsilon_n \rangle)}{\sum_{n=1}^{N} \sigma_n^{-2}} \right\rangle - \langle s \rangle \qquad (25)$$

It is a **class exercise** to show that RHS of (25) is equal to 0.

Therefore, if we know the bias, we can always create an unbiased analysis estimate by first apply bias correction to the measurements.

Because of this, we shall usually assume unbiased errors in the future to keep the analysis simpler.

Finally, we note that we can write our expression for I (eq. 17), $I = \frac{1}{2}\sum_{n=1}^{N} \sigma_n^{-2} (s_a - s_n)^2$, in a simpler form, that is,

$$I_b = \sum_{n=1}^{N} w_n d_n^2 \,, \qquad (26)$$

where $d_n = s_a - s_n$ is the residual of the n^{th} observation

and $w_n = \sigma_n^{-2} / 2$ is the *a priori* (specified) weight.
These w_n are often called Gauss precision moduli and they are inversely proportional to the observational error variance - thus small errors yield larger weights.

Class exercise

Consider the $N = 2$ case for the equations developed for I, the quantity to be minimized (eq. 17); s_a, the analysis value (eq. 18); and $\langle \varepsilon_a^2 \rangle$, the expected error variance of the analysis (eq. 20). Denote the two observations of s as s_o and s_b, with corresponding error variances of σ_o^2 and σ_b^2 respectively. Show that these three equations become:

$$I = \frac{(s_a - s_o)^2}{2\sigma_o^2} + \frac{(s_a - s_b)^2}{2\sigma_b^2}, \qquad (27)$$

$$s_a = \frac{\sigma_o^{-2} s_o + \sigma_b^{-2} s_b}{\sigma_o^{-2} + \sigma_b^{-2}} = s_b + \frac{\sigma_b^2}{\sigma_o^2 + \sigma_b^2}\ s_o - s_b\ , \qquad (28)$$

$$\left\langle \varepsilon_a^2 \right\rangle = \sigma_b^2 - \frac{\sigma_b^4}{\sigma_o^2 + \sigma_b^2} = \frac{\sigma_b^2 \sigma_o^2}{\sigma_o^2 + \sigma_b^2} = \left[\sigma_o^{-2} + \sigma_b^{-2}\right]^{-1}. \qquad (29)$$

These equations represent the scalar or zero-dimensional application of least squares estimation. Although very simple, they foreshadow very accurately the results we will see from much more sophisticated analysis schemes. On the next page, we will start considering vector versions of least squares – although only in the spatial dimensions.

Spatial applications of least squares

Now we define a dependent state variable (e.g., temperature) $f(\mathbf{r})$, where $\mathbf{r} = (x, y, z)$ is a 3-D spatial coordinate.

We first do the $N = 1$ case - one observation at each point.

Denote $f_o(\mathbf{r}_k)$ as an observation of f at station $\mathbf{r}_k$ with expected obs. error variance of $< \varepsilon_o^2 (\mathbf{r}_k) >$.

Also assume K observations whose errors are normally distributed, unbiased, and spatially uncorrelated. This last assumption means that the covariance

$$< \varepsilon_o (\mathbf{r}_k)\, \varepsilon_o (\mathbf{r}_l) > \; = \; 0, \quad \text{for } k \neq l \qquad [\; k = l \text{ gives us } < \varepsilon_o^2 (\mathbf{r}_k) >]$$

Finally, define $f_A(\mathbf{r})$ as the analyzed field of $f(\mathbf{r})$.

The quantity to be minimized in this case, written in the notation of eq. (26), is

$$\mathrm{I} = \sum_{k=1}^{K} w_k d_k^2 = \frac{1}{2} \sum_{k=1}^{K} < \varepsilon_o^2(\mathbf{r}_k) >^{-1} [\, f_o(\mathbf{r}_k) - f_A(\mathbf{r}_k)]^2 \,. \tag{30}$$

Minimization of (30) w.r.t. each unknown analysis value $f_A(\mathbf{r}_k)$ yields

$$f_A(\mathbf{r}_k) = f_o(\mathbf{r}_k) \tag{31}$$

That is, since there is only a single observation per station, the most probable value is that observation. Note that (31) is the true solution only if we have error-free observations.

Also note that (30) is usually minimized w.r.t. some constraint; e.g. – a polynomial fit (because observations are likely not at the grid point locations).

We now introduce matrix notation. Equation (30) is now

$$\mathrm{I} = \frac{1}{2}[\mathbf{f}_A - \mathbf{f}_o]^T \mathbf{W}[\mathbf{f}_A - \mathbf{f}_o] \qquad (32)$$

where $\mathbf{f}_A$, $\mathbf{f}_o$ are column vectors of length K:

and **W** is a K x K (square) matrix with diagonal elements only:

$$w_k \sim \langle \varepsilon^2_o (\mathbf{r}_k) \rangle^{-1}$$

Note that eq. (32) is written in this form to ensure that the matrices are conformal for multiplication. Thus, in (32), the $[\quad]^T$ term creates a (1, K) row vector that is conformal with **W** whereas column vector [ ] is not.

To illustrate, $[\mathbf{f}_A - \mathbf{f}_o]^T \quad \mathbf{W}$ is conformal and yields a [1, K] row

(1, K) (K x K)

vector that is conformal with the [K, 1] column vector $[\mathbf{f}_A - \mathbf{f}_o]$. The result of this second multiplication (1, K)(K, 1) yields a (1, 1) matrix - i.e. – a <u>scalar</u>.

Eq. (32) is often written is the form

$$I = \frac{1}{2}[\mathbf{f}_A - \mathbf{f}_o]^T \mathbf{O}^{-1}[\mathbf{f}_A - \mathbf{f}_o] \qquad (33)$$

where **O** is a K x K <u>observational error variance matrix</u> with elements $< \varepsilon_o^2(r_k) >$

Here, **O** is a diagonal matrix, but if the errors are <u>spatially correlated</u>, that is,

$< \varepsilon_o(\mathbf{r}_k)\, \varepsilon_o(\mathbf{r}_l) > \neq 0$, then **O** is a "full" <u>observational error covariance matrix</u>.

Note that **O** needs to be nonsingular so that its inverse exists.

<u>Spatial applications of least squares</u> - <u>N = 2 Case</u>

Now, as in the zero-dimensional application, we have 2 pieces of information: an <u>observation</u> $f_o(\mathbf{r})$ and a **<u>background estimate</u>** of f, $f_B(\mathbf{r})$. Note that in the older literature, $f_B(\mathbf{r})$ is often called the "first guess field". Please read p. 25-27 in Daley's book, especially Fig. 1.12, to see the role of the background field.

The background field is vital in atmospheric analysis problems since the observations are not uniformly distributed. Sources for $f_B(\mathbf{r})$ could be climatology, a previous analysis or a previous forecast valid at the time of the new analysis - the latter is the most common in NWP applications.

The background field also has errors, denoted by $\varepsilon_B(\mathbf{r}_k)$. The estimation of these errors is a major research topic in data assimilation research. In NWP, they represent the errors in the short-range forecast that provides the background field. Three possible ways to obtain these errors are:

1. Compare the forecast values with the observations at each point. The flaw here is that the obs. also have errors, which are not precisely known.

2. (NMC method) Subtract two forecasts of different length that are valid at the same time. The difference between them represents forecast error growth.

3. Calculate them from an ensemble of many short-range forecasts made in real-time (basis of ensemble Kalman filter data assimilation)

Now, assume both $\varepsilon_o(\mathbf{r}_k)$ and $\varepsilon_B(\mathbf{r}_k)$ to be unbiased, random, and normally distributed. Also assume that ε_o and ε_B are spatially correlated (e.g., the error field has synoptic structure) but *not* with each other. We express these assumptions as:

$$< \varepsilon_o(\mathbf{r}_k)\, \varepsilon_o(\mathbf{r}_l) > \neq 0 \; ; \qquad < \varepsilon_B(\mathbf{r}_k)\, \varepsilon_B(\mathbf{r}_l) > \neq 0 \qquad (34)$$

$$< \varepsilon_o(\mathbf{r}_k)\, \varepsilon_B(\mathbf{r}_l) > = 0 \quad \text{for all } k, l \qquad (35)$$

The quantity to be minimized is given by eq. (33) for N = 2:

$$\mathrm{I} = \frac{1}{2}[\mathbf{f}_A - \mathbf{f}_o]^T \mathbf{O}^{-1}[\mathbf{f}_A - \mathbf{f}_o] + \frac{1}{2}[\mathbf{f}_A - \mathbf{f}_\mathrm{B}]^T \mathbf{B}^{-1}[\mathbf{f}_A - \mathbf{f}_B] \qquad (36)$$

where **B** is the background error covariance matrix.

[Note that $\mathbf{f}_\mathrm{A}$, $\mathbf{f}_\mathrm{B}$ and $\mathbf{f}_\mathrm{o}$ are all defined at observation points here, so we don't have an objective analysis scheme yet!]

To provide one more example of the simplifying properties of matrix notation, here is what eq. (36) looks like in summation notation:

$$\begin{aligned}\mathrm{I} = \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{K}\Big\{ & [f_A(\mathbf{r}_k) - f_o(\mathbf{r}_k)][f_A(\mathbf{r}_l) - f_o(\mathbf{r}_l)]O_{kl} \\ & + [f_A(\mathbf{r}_k) - f_B(\mathbf{r}_k)][f_A(\mathbf{r}_l) - f_B(\mathbf{r}_l)]B_{kl} \Big\}\end{aligned} \qquad (37)$$

where O_{kl} , B_{kl} are the elements of $\mathbf{O}^{-1}$, $\mathbf{B}^{-1}$. K is the total number of observations.

Now perform the least squares minimization by differentiating (36) or (37) w.r.t. the unknown analysis values $f_\mathrm{A}(\mathbf{r})$. Use of eq. (37) will yield:

$$\frac{\partial \mathrm{I}}{\partial f_A(\mathbf{r}_k)} = \sum_{l=1}^{K}\Big\{ [f_A(\mathbf{r}_k) - f_o(\mathbf{r}_k)]O_{kl} + [f_A(\mathbf{r}_k) - f_B(\mathbf{r}_k)]B_{kl} \Big\} = 0 \quad \text{for } k=1, \ldots, \mathrm{K}.$$

whereas use of eq. (36) gives us:

$$\frac{\partial \mathrm{I}}{\partial f_A(\mathbf{r}_k)} = \mathbf{O}^{-1}[\mathbf{f}_A - \mathbf{f}_o] + \mathbf{B}^{-1}[\mathbf{f}_A - \mathbf{f}_B] = 0$$

or $$\mathbf{O}^{-1}\mathbf{f}_A - \mathbf{O}^{-1}\mathbf{f}_o + \mathbf{B}^{-1}\mathbf{f}_A - \mathbf{B}^{-1}\mathbf{f}_B = 0$$

or $$(\mathbf{O}^{-1} + \mathbf{B}^{-1})\ \mathbf{f}_A = \mathbf{B}^{-1}\mathbf{f}_B + \mathbf{O}^{-1}\mathbf{f}_o$$

Solving for $\mathbf{f}_A$, we get

$$\mathbf{f}_A = [\mathbf{O}^{-1} + \mathbf{B}^{-1}]^{-1}[\mathbf{B}^{-1}\mathbf{f}_B + \mathbf{O}^{-1}\mathbf{f}_o] \qquad (38)$$

Class exercise: Show that eq. (38) can be rewritten as

$$\mathbf{f}_A = \mathbf{f}_B + \mathbf{B}[\mathbf{O} + \mathbf{B}]^{-1}[\mathbf{f}_o - \mathbf{f}_B]\,. \qquad (39)$$

[Note how (39) yields a (K,1) column vector of most likelihood estimates of the analysis values for the K obs. locations.]

This exercise is very challenging.   Here are two hints - make use of the following two lemmas:

Lemma #1:

$$\mathbf{X} [\mathbf{X} + \mathbf{Y}]^{-1} + \mathbf{Y} [\mathbf{X} + \mathbf{Y}]^{-1} = [\mathbf{X} + \mathbf{Y}] [\mathbf{X} + \mathbf{Y}]^{-1} = \mathbf{I}$$

Similarly,

$$\mathbf{X}^{-1}[\mathbf{X}^{-1} + \mathbf{Y}^{-1}]^{-1} + \mathbf{Y}^{-1}[\mathbf{X}^{-1} + \mathbf{Y}^{-1}]^{-1} = [\mathbf{X}^{-1} + \mathbf{Y}^{-1}] [[\mathbf{X}^{-1} + \mathbf{Y}^{-1}]^{-1} = \mathbf{I}$$

Lemma #2: $(\mathbf{X}\, \mathbf{Y})^{-1} = \mathbf{Y}^{-1} \mathbf{X}^{-1}$

In eq. (39), $\mathbf{f_o} - \mathbf{f_B}$ are the observation increments (these might be viewed as forecast

errors if the obs. were perfect, but, since they are not, we view this increment as simply a correction to the background field),

and $\mathbf{f_A} - \mathbf{f_B}$ are known as the analysis increments.

Again note that (39) provides analysis values at the observation locations; - it does not provide a gridded analysis.

As usual, we now determine the expected analysis error variance for this analysis system.

Define $\varepsilon_A = f_A(\mathbf{r}) - f_T(\mathbf{r})$, which is a column vector of analysis errors, where f_T

are the (unknown) true values.   Following the text, it is a **class exercise** to obtain

$$< \boldsymbol{\varepsilon}_{\mathbf{A}} \boldsymbol{\varepsilon}^{\mathbf{T}}_{\mathbf{A}} > \; = [\, \mathbf{B}^{-1} + \mathbf{O}^{-1} \,]^{-1} \qquad (40)$$

where the LHS is an analysis error covariance matrix with elements $< \varepsilon_A(\mathbf{r}_k)\, \varepsilon_A(\mathbf{r}_l) >$.

We define the analysis error **A** as the LHS of (40),

Please show that **A** can also be written as

$$\mathbf{A} = \mathbf{B}\,[\,\mathbf{B} + \mathbf{O}\,]^{-1}\,\mathbf{O} = \mathbf{B} - \mathbf{B}\,[\,\mathbf{B} + \mathbf{O}\,]^{-1}\,\mathbf{B} \qquad (41)$$

Note that equations (36), (39) and (41) are the vector equivalents to the scalar equations (27), (28) and (29).

Note that it is mathematically possible to minimize other than quadratic forms; i.e. – could minimize

$$I = \frac{1}{q} \sum_{n=1}^{N} \frac{[s_a - s_n]^q}{\sigma_n^q} \qquad (42)$$

The choice of q depends on what you think the probability density function for the errors is.

Here [e.g., eq. (36)], we have chosen $q = 2$, or minimization in the l_2 norm sense, because we believe that the normal or Gaussian distribution function is appropriate for most meteorological data errors (provided "gross errors" have been removed by a quality control procedure beforehand). Also, $q = 2$ leads to linear analysis equations.

A choice of $q = 1$ corresponds to a "long-tailed" error distribution – meaning large errors have a higher probability than with a Gaussian one.

See Daley book for more discussion of the differences between l_1 and l_2 minimization.