A quick introduction to data assimilation

Alexandre Fournier

Institut de Physique du Globe de Paris, Paris, France

LabEx UnivEarthS Fall School, October 2015

Tentative plan

1. **Introduction**
2. **Stochastic Estimation. The BLUE**
3. **The Kalman filter**
4. **Variational assimilation**
5. **Numerical Weather Prediction**
6. **Other examples**

What is data assimilation?

The basic purpose of data assimilation is to combine different sources of information in order to produce the best possible estimate of the state of a system.

- observations of the system
- physical laws describing its behaviour (numerical model)

Why not simply use observations?

- too sparse or partial in geo- and astrophysics
- interpolation: numerical model
- combining (by means of the model) several noised data: filter out part of the noise & more accurate estimate ("accuracies are added")

Different mathematical approaches

The problem of data assimilation can be tackled using different mathematical approaches: signal processing, control theory, estimation theory, ...

- Stochastic methods (eg Kalman filter): estimation theory.
- Variational methods (3D-Var, 4D-Var...): control theory.

Fields of application

Meteorology (initialization of a forecast): sole field of application until early 1990ies.
Today:

- atmospheric chemistry
- oceanic biochemistry
- glaciology
- physical oceanography
- geomagnetism
- solar magnetism
- seismology
- ...

A variety of purposes

- the estimation of the trajectory of a system to study its variability (reanalyses)
- the identification of systematic errors in numerical models
- the estimation of unobserved field variables (e.g. the magnetic field inside Earth's core)
- the estimation of parameters (e.g. a structural Earth model in seismology)
- the optimization of observation networks

A scalar example

Assume we have two distinct measurements,

$$y_1 = 1$$

and

$$y_2 = 2,$$

of the same unknown quantity x.

What estimation of its true value can we make?

First approach

We seek x which minimizes $(x-1)^2+(x-2)^2$, and we find the estimate

$$\widehat{x} = 3/2 = 1.5$$

(this is the least-squares solution).
This solution has the following problems:

- it is sensitive to any change of units. If $y_1 = 1$ is a measurement of x and $y_2 = 4$ is a measurement of $2x$, then minimizing $(x-1)^2+(2x-4)^2$ leads to $\widehat{x} = 9/5 = 1.8$.
- it does not reflect the quality of the various measurements.

Reformulation in a statistical framework

We define

$$Y_i = x + \epsilon_i, \tag{1}$$

where the observation errors ϵ_i satisfy the following hypotheses

- $\mathrm{E}(\epsilon_i) = 0$ (unbiased measurements)
- $\mathrm{Var}(\epsilon_i) = \sigma_i^2$ (accuracy is known)
- $\mathrm{Covar}(\epsilon_1, \epsilon_2) = 0$, i.e. $\mathrm{E}(\epsilon_1 \epsilon_2) = 0$, errors are independent.

We next seek an estimator (i.e. a random variable) $\widehat{X}$ which is

- linear: $\widehat{X} = \alpha_1 Y_1 + \alpha_2 Y_2$
- unbiased: $\mathrm{E}\left(\widehat{X}\right) = x$
- of minimum variance: $\mathrm{Var}\left(\widehat{X}\right)$ minimal (optimal accuracy)

BLUE

This estimator is called the **BLUE**: Best Linear Unbiased Estimator. To compute the α_i we use the unbiased hypothesis

$$\mathrm{E}\left(\widehat{X}\right) = x = (\alpha_1 + \alpha_2)x + \alpha_1 \mathrm{E}\left(\epsilon_1\right) + \alpha_2 \mathrm{E}\left(\epsilon_2\right) = (\alpha_1 + \alpha_2)x, \quad (2)$$

so that $\alpha_1 + \alpha_2 = 1$, or $\alpha_2 = 1 - \alpha_1$. Next we compute the variance of $\widehat{X}$.

$$\begin{aligned}
\mathrm{Var}\left(\widehat{X}\right) &= \mathrm{E}\left[\left(\widehat{X} - x\right)^2\right] = \mathrm{E}\left[\left(\alpha_1\epsilon_1 + \alpha_2\epsilon_2\right)^2\right] \\
&= \alpha_1^2 \mathrm{E}\left(\epsilon_1^2\right) + 2\alpha_1\alpha_2 \mathrm{E}\left(\epsilon_1\epsilon_2\right) + \alpha_2^2 \mathrm{E}\left(\epsilon_2^2\right) \\
&= \alpha_1^2\sigma_1^2 + \alpha_2^2\sigma_2^2 \\
&= \alpha_1^2\sigma_1^2 + (1 - \alpha_1)^2\sigma_2^2.
\end{aligned}$$

Our estimator $\widehat{X}$ has to minimize this quantity.

BLUE

Computing α_1 such that

$$\frac{\mathrm{d}}{\mathrm{d}\alpha_1}\mathrm{Var}\left(\widehat{X}\right) = 0 \tag{3}$$

yields

$$\alpha_1 = \frac{\sigma_2^2}{\sigma_2^2 + \sigma_1^2}. \tag{4}$$

It follows that

$$\widehat{X} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} y_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} y_2. \tag{5}$$

Note that we get the same result if we try to minimize the functional

$$\mathcal{J}(x) = \frac{1}{2}\left[\frac{(x - y_1)^2}{\sigma_1^2} + \frac{(x - y_2)^2}{\sigma_2^2}\right]. \tag{6}$$

Comments

- This statistical approach solves the problem of sensitivity to units and it incorporates measurement accuracies.
- The accuracy of the estimator is given by the second derivative of $\mathcal{J}$

$$\left.\frac{\mathrm{d}^2\mathcal{J}}{\mathrm{d}x^2}\right|_{x=\widehat{x}} = \frac{1}{\mathrm{Var}\left(\widehat{X}\right)} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}, \tag{7}$$

  so that "accuracies are added".
- If we consider that $y_1 = x^b$ is a first guess of x (with standard deviation $\sigma_b = \sigma_1$) and $y_2 = y$ is an additional observation (with std dev $\sigma = \sigma_2$), then we can rearrange Eq. (5) as

$$\widehat{X} = x^b + \frac{\sigma_b^2}{\sigma^2 + \sigma_b^2}\left(y - x^b\right). \tag{8}$$

  The quantity $y - x^b$ is called the **innovation**. It contains the additional information provided by y with respect to x^b.

Data assimilation methods

Two classes of methods

- statistical methods: direct computation of the BLUE thanks to algebraic computations (the Kalman filter);
- variational mehods: minimization of the functional $\mathcal{J}$ (4D-Var).

Shared properties

- they provide the same result (in the linear case);
- their optimality can only be demonstrated in the linear case;

Shared difficulties

- accounting for non-linearities
- dealing with large problems
- error statistics are required but sometimes only poorly known

Notations

There exists some sort of standard notations, summarized by Ide et al. (1997).

- $\mathbf{x}$ state vector
- $\mathbf{x}^t$ true state
- $\mathbf{x}^b$ background state
- $\mathbf{x}^a$ analyzed state

Superscripts denote vector types, subscripts refer to space or time. In the following: unless otherwise noted, all vectors will be column vectors. If $\mathbf{a}$ and $\mathbf{b}$ are two column vectors of equal size n, with the superscript T denoting transposition, then

$$\mathbf{a}^T\mathbf{b} \quad \text{is their scalar product } = \sum a_i b_i, \tag{9}$$

$$\mathbf{a}\mathbf{b}^T \quad \text{is a matrix of coefficients } a_i b_j, (i,j) \in \{1,\ldots,n\}^2. \tag{10}$$

Discretization and true state

Most of the time, our goal will be to estimate as accurately as possible a geophysical field that varies continuously in space and time. This real, continuous (and possibly multivariate) field is denoted by $\mathfrak{x}$.
Numerical models are often used for the estimation. Numerical models operate in a discrete world and only handle discrete representations of physical fields. Therefore we will try to estimate a projection of the real state $\mathfrak{x}$ onto a discrete space. Let $\mathbf{\Pi}$ denote the associated projector, and $\mathbf{x}^t$ be the projection of $\mathfrak{x}$

$$\mathbf{x}^t = \mathbf{\Pi}(\mathfrak{x}). \tag{11}$$

$\mathbf{x}^t$ is called the true state (see above); this is the state we wish to estimate in practice.

Discretization and true state

In a data assimilation problem, one deals with **dynamical** models that compute the time evolution of the simulated state. Let $\mathfrak{x}_i$ and $\mathfrak{x}_{i+1}$ be the real (continuous) states at two consecutive observation times, i being a time index. These two states are related by a causal link (the physical model)

$$\mathfrak{x}_{i+1} = g\left(\mathfrak{x}_i\right). \tag{12}$$

Projecting this equality into the discrete world, we get

$$\mathbf{x}_{i+1}^t = \mathbf{\Pi}\left[g\left(\mathfrak{x}_i\right)\right]. \tag{13}$$

The dynamical model g is not strictly known, even though we hopefully know most of the physics involved in it. This physics is represented in the discrete world by our numerical model $\mathcal{M}$, which operates on discrete states such as $\mathbf{x}^t$. Introducing this model into Eq. (13), we get

$$\mathbf{x}_{i+1}^t = \mathcal{M}_{i,i+1}\left(\mathbf{x}_i^t\right) + \boldsymbol{\eta}_{i,i+1}, \tag{14}$$

in which

$$\boldsymbol{\eta}_{i+1} = \mathbf{\Pi}\left[g\left(\mathfrak{x}_i\right)\right] - \mathcal{M}_{i,i+1}\left(\mathbf{x}_i^t\right). \tag{15}$$

Discretization and true state

The **model error** $\boldsymbol{\eta}_{i+1}$ term accounts for the errors in the numerical models (e.g. misrepresentation of some physical processes) and for the errors due to the discretization. The covariance matrix $\mathbf{Q}_{i+1}$ of the model error is given by

$$\mathbf{Q}_{i+1} = \mathrm{Covar}\left(\boldsymbol{\eta}_{i+1}\right) = \mathrm{E}\left[\left(\boldsymbol{\eta}_{i+1} - \langle\boldsymbol{\eta}_{i+1}\rangle\right)\left(\boldsymbol{\eta}_{i+1} - \langle\boldsymbol{\eta}_{i+1}\rangle\right)^{T}\right], \quad (16)$$

where $\langle\boldsymbol{\eta}_{i+1}\rangle = \mathrm{E}\left(\boldsymbol{\eta}_{i+1}\right)$ is the average error.

Observations I

The real, continous field $\mathscr{x}$ results in a signal $\mathscr{y}$ in the space of observations. This involves a mapping $\mathscr{h}$

$$\mathscr{y} = \mathscr{h}(\mathscr{x}). \tag{17}$$

Despite its simplicity, this equation can not be used in practice. First, we do not have access to the real $\mathscr{y}$: the observed field $\mathbf{y}^o$ is contaminated with measurement errors, denoted by $\boldsymbol{\epsilon}^\mu$. Accordingly,

$$\mathbf{y}^o = \mathscr{h}(\mathscr{x}) + \boldsymbol{\epsilon}^\mu. \tag{18}$$

Second, $\mathscr{h}$, which represents the physics of the measurement process (which might be exactly known), is a continuous mapping. In practice, this physics is represented by a numerical operator $\mathcal{H}$, which is applied to the discrete state we wish to estimate, $\mathbf{x}^t$. Incorporating $\mathcal{H}$ and $\boldsymbol{\Pi}$ in Eq. (18) yields

$$\mathbf{y}^o = \mathcal{H}(\mathbf{x}^t) + \underbrace{\mathscr{h}(\mathscr{x}) - \mathcal{H}[\boldsymbol{\Pi}(\mathscr{x})]}_{\boldsymbol{\epsilon}^r} + \boldsymbol{\epsilon}^\mu, \tag{19}$$

Observations II

where $\boldsymbol{\epsilon}^r$ is often termed the error of representativeness (Lorenc, 1986), which includes the errors related to the representation of the physics in $\mathcal{H}$ and those errors due to the projection $\mathbf{\Pi}$ of the real state $\mathfrak{x}$ onto the discrete state space (due for instance to numerical interpolation). The sum of the measurement error and the error of representativeness is the **observation error**

$$\boldsymbol{\epsilon}^o = \boldsymbol{\epsilon}^\mu + \boldsymbol{\epsilon}^r. \tag{20}$$

This allows us to write the final form of the equation relating the discrete true state $\mathbf{x}^t$ and the observations

$$\mathbf{y}^o = \mathcal{H}\left(\mathbf{x}^t\right) + \boldsymbol{\epsilon}^o. \tag{21}$$

The covariance matrix of the observation error $\boldsymbol{\epsilon}^o$ is defined by

$$\mathbf{R} = \mathrm{Covar}\left(\boldsymbol{\epsilon}^o\right) = \mathrm{E}\left[\left(\boldsymbol{\epsilon}^o - \left\langle\boldsymbol{\epsilon}^o\right\rangle\right)\left(\boldsymbol{\epsilon}^o - \left\langle\boldsymbol{\epsilon}^o\right\rangle\right)^T\right]. \tag{22}$$

A priori (background) information

It can be that we have some a priori knowledge of the state $\mathbf{x}^t$, under the form of a vector $\mathbf{x}^b$ having the same dimension as $\mathbf{x}^t$. This is the **background state**. Following a similar logic, the background error is defined as

$$\boldsymbol{\epsilon}^b = \mathbf{x}^b - \mathbf{x}^t. \tag{23}$$

Often the estimate of the background state comes from a model simulation. In this case, the background is a **forecast** and is rather denoted by $\mathbf{x}^f$, with forecast error $\boldsymbol{\epsilon}^f$.

A priori (background) information

The covariance $\mathbf{P}^b$ of the background error is given by

$$\mathbf{P}^b = \mathrm{Covar}\left(\boldsymbol{\epsilon}^b\right) = \mathrm{E}\left[\left(\boldsymbol{\epsilon}^b - \langle\boldsymbol{\epsilon}^b\rangle\right)\left(\boldsymbol{\epsilon}^b - \langle\boldsymbol{\epsilon}^b\rangle\right)^T\right]. \tag{24}$$

Analysis

The result of the assimilation process is often called the analysis, and is denoted by $\mathbf{x}^a$. The analysis error is defined by

$$\boldsymbol{\epsilon}^a = \mathbf{x}^a - \mathbf{x}^t, \tag{25}$$

while the covariance matrix of the analysis error $\boldsymbol{\epsilon}^a$ is defined by

$$\mathbf{P}^a = \mathrm{Covar}\left(\boldsymbol{\epsilon}^a\right) = \mathrm{E}\left[\left(\boldsymbol{\epsilon}^a - \langle\boldsymbol{\epsilon}^a\rangle\right)\left(\boldsymbol{\epsilon}^a - \langle\boldsymbol{\epsilon}^a\rangle\right)^T\right]. \tag{26}$$

An important comment

the problem is entirely set-up once the physical model and the observations have been chosen, and the covariances (and possibly the background) defined. All the physics has been introduced at this stage. The remaining part (the production of the analysis) is technical.

Useful references

- "Discrete Inverse and State Estimation Problems", by Wunsch (2006), provides a very personal and powerful account of adjoint methods and their application in geophysical fluid dynamics (oceanography).
- In her book entitled "Atmospheric Modelling, Data Assimilation and Predictability ", E. Kalnay (2003) has two comprehensive and very well-written chapters on the basics and applications of data assimilation techniques to atmospheric dynamics.
- In addition, Evensen (2009) provides a very complete treatment of data assimilation techniques, with a strong and useful emphasis on the basics and applications of the ensemble Kalman filter he invented.
- Last, but not least, Blayo et al. (2014) is an excellent compilation covering theoretical and practical aspects of data assimilation in geosciences.

Useful references

For a start, I would highly recommend the review paper by Talagrand (1997), "Assimilation of observations, an introduction" which provides an extremely concise and well-written overview of the topic.
In addition, if you are looking for references related to the geophysical inverse problem in general, Parker (1994) and Tarantola (2005) provide two very personal, insightful, and sometimes contradictory views on how we should go about making inference on the Earth based on a finite number of noisy observations and on physical laws governing its behaviour.

Bibliography I

Blayo, E., M. Bocquet, E. Cosme, and L. F. Cugliandolo, 2014: Advanced data assimilation for geosciences. *International Summer School-Advanced Data Assimilation for Geosciences*, Oxford University Press, 608.

Evensen, G., 2009: *Data assimilation: The ensemble Kalman filter*. 2d ed., Springer, Berlin, doi:10.1007/978-3-642-03711-5.

Ide, K., P. Courtier, M. Ghil, and A. C. Lorenc, 1997: Unified notation for data assimilation: Operational, sequential and variational. *Journal of the meteorological society of Japan*, **75**, 181–189.

Kalnay, E., 2003: *Atmospheric modeling, data assimilation, and predictability*. Cambridge University Press, Cambridge.

Lorenc, A. C., 1986: Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, **112 (474)**, 1177–1194, doi:10.1002/qj.49711247414.

Parker, R. L., 1994: *Geophysical inverse theory*. Princeton University Press, Princeton, NJ.

Bibliography II

Talagrand, O., 1997: Assimilation of observations, an introduction. *Journal of the Meteorological Society of Japan*, **75 (1B)**, 191–209.

Tarantola, A., 2005: *Inverse problem theory and methods for model parameter estimation*. Society for Industrial Mathematics, Philadelphia, PA.

Wunsch, C., 2006: *Discrete inverse and state estimation problems*. Cambridge University Press, Cambridge.