MATHEMATICAL THEORY OF DATA PROCESSING IN MODELS (DATA ASSIMILATION PROBLEMS)

A.W. Heemink
Delft University of Technology, 2600 GA Delft, The Netherlands

Keywords: State space models, data assimilation, variational method, adjoint method, stochastic models, linear Kalman filtering, Extended Kalman filtering, Ensemble Kalman filter, Reduced-Rank Square Root filter.

Contents

Summary

To understand and predict the behavior of an environmental system one can use measurements or develop physically based models. In many applications however neither of these approaches is able to provide an accurate description of the dynamic behavior of the system. A model is always a simplification of the real world while measurements seldom produce a complete picture of the system behavior. Using data assimilation techniques measurements and model results are both used to obtain an optimal estimate of the state of the system. In this chapter we present an overview of methods available to assimilate data into a numerical model. Attention is concentrated on variational methods and on Kalman filtering. The main problem of using these advanced data assimilation schemes is the huge computational burden that is required for solving real life problems. For variational methods the adjoint model implementation is essential to obtain an efficient data assimilation algorithm. For Kalman filtering problems a number of approximate algorithms have been introduced recently: Ensemble Kalman filters and Reduced Rank filters. These algorithms make the application of Kalman filtering to large-scale data assimilation problems feasible. After a brief introduction to the most important data assimilation approaches we will discuss

the advantages and disadvantages of the various methods.

1. Introduction

Measurements can be used to develop statistical models for predicting the behavior of environmental processes. However these types of models are derived from the data and do not include physical knowledge of the process. Furthermore, measurements alone do generally not provide a complete picture of the process. Especially in case of processes that vary in space and time it is very hard to reconstruct the spatial and temporal patterns only from data. Physically based models, deterministic or stochastic, produce results that are spatially and temporally consistent. However these models are usually not able to accurately reproduce the measurements that are available. The information provided by the models and by the measurement information is often complementary. Therefore it is important to study a methodology for integrating measurements and physically based mathematical models. This methodology is called data assimilation. By using models that are based on physical laws and that are continuously adapted by the measurements available the two sources of information of the process, model information and measurement information can be integrated.

Data assimilation can be defined as a procedure to incorporate data into a model simulation so as to improve the predictions. However, assimilating data into a numerical model is far from trivial. The simplest data assimilation procedure is to overwrite the model values at the measurement locations with the observed data. Inserting the data in this way into a numerical model is in general not a satisfactory method. It leaves the model dynamically unbalanced and introduced spurious waves into the model. These short waves may even cause instabilities of the underlying numerical model.

The most common data assimilation technique used in numerical weather prediction is optimal interpolation. Here some estimates of the error statistics of the numerical model are used to correct the results of the model using the measurements. However, since these error statistics have to be determined by adopting some *ad-hoc* statistical assumptions, the correction produced by optimal interpolation is again not consistent with the underlying numerical model. As a consequence the use of optimal interpolation still often yields unrealistic correction or instabilities.

More accurate data assimilation methods are variational data assimilation and Kalman filtering. The basic idea of these data assimilation methods is to use the data to only correct the weak points in the model. Weak parts of the model may be due to uncertainty in initial and boundary conditions or imperfectly known model parameters. The data is not allowed to modify the accurate parts of the model. Therefore a two-step procedure is introduced:

Step 1: Specify the uncertainties in the model

Step 2: Use the data to estimate the uncertainties as accurate as possible

As a result these types of data assimilation problems are in fact, inverse problems. The specified inputs (model uncertainties) have to be reconstructed from the output

(measurements). A variational approach or Kalman filtering solves these inverse problems accurately. For linear problems it can be shown that both approaches produce exactly the same results for the same problem formulation. Optimal interpolation does not solve an inverse problem. It produces corrections for the model output without reconstructing model uncertainties. As a result the variational method and Kalman filtering are superior to optimal interpolation. In fact, optimal interpolation can be considered as a simplified Kalman filter.

In the last decennium the variational approach and Kalman filtering have gained acceptance as powerful frameworks for data assimilation. However, both methods require a very large computational burden, at least an order of magnitude larger then the computational effort required for the underlying numerical model. This is the main disadvantage of these methods compared to optimal interpolation that requires only a small increase in computer time.

Starting point for the data assimilation methodology is a state space representation of the model and the measurements. Let us assume that modeling techniques have provided us with a deterministic state space representation of the form:

$$X_{k+1} = f(X_k, k) + B(k)u_k, \quad X_0 = x_0 \tag{1}$$

Here the X_k is the system state, u_k is the input of the system, f is a nonlinear function, and $B(k)$ is an input matrix. For a numerical model that describes the behavior of an environmental process in space and time, the state consists of all the variables in all the grid points of the model at a certain time, while the function f in this case represents one time step of the numerical scheme of the model.

The measurements taken from the actual system are assumed to be available according to the relation:

$$Z_k = m(X_k, k) \tag{2}$$

where Z_k is a vector containing the measurements and m is a nonlinear function that specifies the relation between the model results and the measurements.

In this chapter we describe in Section 2 the basic idea of the variational approach and discuss a number of extensions. In Section 3 we introduce the Kalman filter as data assimilation framework. Here we present a number of filter algorithms for solving large-scale data assimilation problems.

2. Variational Data Assimilation

2.1. Data Assimilation Formulated as a Minimization Problem

If it can be assumed that the only uncertainties of the model (1)-(2) are introduced by a number of poorly known parameters, the data assimilation problem can be formulated as a deterministic parameter estimation problem. Rewrite the model according to:

$$X_{k+1} = f_p(X_k, p, k) + B(k, p)u_k, \quad X_0 = x_0 \tag{3}$$

$$Z_k = m(X_k, k) \tag{4}$$

where p is the vector containing the uncertain parameters. Uncertain parameters may be model parameters, initial conditions or inputs.

In order to estimate the parameters we first define a criterion $J(p)$ as a measure for the distance between the measurements and the model results:

$$J(p) = \sum_{k=1}^{K} \left(Z_k - m(X_k, k)\right)^{\mathrm{T}} R(k)^{-1} \left(Z_k - m(X_k, k)\right) \tag{5}$$

Here the generalized least squares criterion or l_2-norm has been chosen to define J. The covariance matrix $R(k)$ is a weighting matrix that takes into account the errors associated with the measurements. This formulation is used very often in practice. The optimal parameter p is found by minimizing the criterion $J(p)$.

Prior information about the parameter values p_0 can be included by adding a regularization term to the criterion:

$$J(p) = \sum_{k=1}^{K} \left(\left(Z_k - m(X_k, k)\right)^{\mathrm{T}} R(k)^{-1} \left(Z_k - m(X_k, k)\right) \right) + \left(p - p_0\right)^{\mathrm{T}} P_0^{-1} \left(p - p_0\right) \tag{6}$$

Here P_0 is the covariance matrix of the prior information p_0, modeling the uncertainty associated with this prior information. Usually p_0 is the first guess of the uncertain parameters and the starting value for the optimization procedure. The regularization term in the criterion prevents that parameter estimates become unrealistic if the measurement information is limited. In this case the estimates will simply remain close to the first guess (as they should be).

-
-
-

Bibliography

Ghil, M., et al (eds.),, 1997 Data assimilation in Meteorology and Oceanography: Theory and Practice, Special issue of the Journal of the Meteorological Society of Japan, vol. 75, pp. 111-496. [Contains introductory papers on the methodology, papers on the various data assimilation algorithms and papers describing real life applications in meteorology, oceanography and hydrology.]

Daley, R., (1991) Atmospheric Data assimilation, Cambridge Atmospheric and Space Science Series, Cambridge University Press, Cambridge. [Focuses on variational methods and on atmospheric applications.]

Mayback, P.S., (1979) Stochastic models, Estimation and Control, Vol. 141-1, Mathematics in science and engineering, Academic Press, New York. [Mainly about Kalman filtering.]

Biographical Sketch

Arnold W. Heemink, of Delft University of Technology has developed for more than 20 years data assimilation techniques for large-scale numerical models and has published more than 100 papers on this subject. The techniques are often based on Kalman filtering or on the adjoint method. From 1981-1993, when Heemink was affiliated to Rijkswaterstaat (ministry of public works), the main applications were flow and transport problems in coastal seas. The data assimilation scheme for the operational storm surge forecasting system in the Netherlands is based on Kalman filtering and has been developed by Heemink in corporation with Rijkswaterstaat and KNMI. More recently he also worked on real life air pollution data assimilation problem in corporation with.

Education

1974-1978 University of Twente, Applied Mathematics, Systems theory, B. Sc. (cum laude)
1978-1980 University of Twente, Applied Mathematics, Mathematical Physics, M. Sc. (cum laude)
1980-1986 University of Twente, Applied Mathematics, Ph. D.

Employment history

1979-1981 University of Twente, Research assistant
1981-1986 Rijkswaterstaat, Researcher Applied Mathematics
1986-1990 Rijkswaterstaat, Senior researcher Applied Mathematics
1990-1993 Rijkswaterstaat, Head section Mathematics
1990-1993 Delft University of Technology, Applied Mathematics, part-time professor

Applied Analysis

1993-present Delft University of Technology, Applied Mathematics, full professor Applied Analysis