Available online at www.sciencedirect.com
ScienceDirect
Computers and Mathematics with Applications 55 (2008) 2247–2265
An International Journal
computers & mathematics with applications
www.elsevier.com/locate/camwa

A numerical comparative study on data assimilation using Kalman filters

G. Dimitriu

Department of Mathematics and Informatics, University of Medicine and Pharmacy, Faculty of Pharmacy, 700115 Iaşi, Romania

Abstract

Kalman filtering has become a powerful framework for solving data assimilation problems. Of interest here are the low-rank filters which are computationally efficient for solving large-scale data assimilation problems. Together with theoretical aspects on the basis of which some common low-rank filters are designed, the paper also presents numerically comparative results of data assimilation using an air pollution model. The performance of such filters, as depending on the distance between the measurement locations and emission points, is investigated.

Keywords: Data assimilation; Kalman filters; Low-rank filters; Air pollution modeling; Advection–diffusion equation

1. Introduction

Originally designed for guidance problems, the Kalman filter [1] has a long history of merging models and measurements in electrical engineering and control. The growing availability of cheap computing power during the last decade made the filter approach feasible for large air pollution models too. Kalman filtering represents a powerful framework for solving data assimilation problems [2]. For the implementation of a Kalman filter, the evolution of the state and observation of measurements can be described with the stochastic system:

$$\mathbf{x}^t[k+1] = \mathbf{A}[k]\mathbf{x}^t[k] + \eta[k], \qquad \mathbf{y}^o[k] = \mathbf{H}[k]'\mathbf{x}^t[k] + \nu[k] \tag{1}$$

with $\mathbf{x}^t[k] \in \mathbb{R}^n$ being the true state vector at time $t[k]$, $\mathbf{A}[k]$ a deterministic model, $\eta[k] \in \mathbb{R}^n$ a Gaussian distributed model error (zero mean, covariance $\mathbf{Q}$), and $\mathbf{y}^o[k] \in \mathbb{R}^r$ a vector of observations with $\nu[k]$ the representation error (Gaussian with zero mean and covariance $\mathbf{R}$). The superscripts t, o, and later on f and a refer to the true, observed, forecasted and analyzed entities, respectively. We also mention that the time indices for $\mathbf{A}$ and $\mathbf{H}'$ will be omitted in the following equations, assuming that the time is implied by the state where the operators act on.

The aim of the filter operations is to obtain the mean $\hat{\mathbf{x}}^a$ and covariance $\mathbf{P}^a$ for the probability density of the true state. The filter equations for this system contain the *forecast* stage given by:

$$\hat{\mathbf{x}}^f[k+1] = \mathbf{A}[k]\hat{\mathbf{x}}^a[k], \qquad \mathbf{P}^f[k+1] = \mathbf{A}\mathbf{P}^a[k]\mathbf{A} + \mathbf{Q}[k] \tag{2}$$

E-mail address: dimitriu@umfiasi.ro.

0898-1221/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2007.11.009

and *analysis* stage expressed by

$$\mathbf{x}^a = \hat{\mathbf{x}}^f + \mathbf{K}(\mathbf{y}^o - \mathbf{H}'\hat{\mathbf{x}}^f) \tag{3}$$

$$\mathbf{P}^a = \begin{cases} (\mathbf{I} - \mathbf{K}^{MV}\mathbf{H}')\mathbf{P}^f, & \mathbf{K}^{MV} = \mathbf{P}^f\mathbf{H}'(\mathbf{H}'\mathbf{P}^f\mathbf{H} + \mathbf{R})^{-1} \\ (\mathbf{I} - \mathbf{K}\mathbf{H}')\mathbf{P}^f(\mathbf{I} - \mathbf{K}\mathbf{H}')' + \mathbf{K}\mathbf{R}\mathbf{K}', & \text{arbitrary gain } \mathbf{K}. \end{cases} \tag{4}$$

In the case of a large model, the propagation of the covariance matrix in (2) represents the most expensive part in the full rank filter. If $\mathbf{A}$ is defined by an $n \times n$ matrix, then the dynamical model is called $2n$ times to perform the operation $\mathbf{A}(\mathbf{A}\mathbf{P})'$. Limitation of both the number of model evaluations as well as the storage requirements will be achieved in this study by reducing the rank of the covariance matrix.

Bierman [3] proposed to write the equations for the Kalman filter using the factorization $\mathbf{P} = \mathbf{S}\mathbf{S}'$. Numerical inaccuracies made in the computation and storage of the matrix $\mathbf{S}$ will never affect the property of the positive definiteness of $\mathbf{P}$. Inaccuracies will even be reduced, since the condition number of $\mathbf{S}$ is only the square root of the condition number of $\mathbf{P}$.

The idea of factorization is useful to reduce the storage requirements of $\mathbf{P}$. Consider a covariance matrix $\mathbf{P}$ written as the product of a rectangular matrix square root $\mathbf{S}$ and its transpose:

$$\mathbf{P}_{n\times n} = \mathbf{S}_{n\times m}\mathbf{S}'_{m\times n}.$$

In order to obtain the Kalman filter in square root form, apart from the previous factorization $\mathbf{P} = \mathbf{S}\mathbf{S}'$ for the covariance of the true state, we also introduce the factorizations $\mathbf{Q} = \mathbf{T}\mathbf{T}'$ and $\mathbf{R} = \mathbf{U}\mathbf{U}'$ for the covariance of the forecast and representation error, respectively. Further, a matrix $\Psi' = \mathbf{H}'\mathbf{S}$ is introduced for the mapping of the forecast covariance root to the observation space.

After (2), the forecasts of mean and covariance become:

$$\hat{\mathbf{x}}^f[k+1] = \mathbf{A}\hat{\mathbf{x}}^a[k] \tag{5}$$

$$(\mathbf{S}^f\mathbf{S}^{f\prime})[k+1] = \mathbf{A}(\mathbf{S}^a\mathbf{S}^{a\prime})[k]\mathbf{A} + \mathbf{T}\mathbf{T}'[k]$$

$$\text{or} \quad \mathbf{S}^f[k+1] = \left[\mathbf{A}\mathbf{S}^a[k], \mathbf{T}[k]\right]. \tag{6}$$

The second formula in (6) is able to reduce both the computational complexity and the numerical inaccuracies, since the condition number of $\mathbf{S}^f$ or $\mathbf{S}^a$ is only the square root of the condition number of $\mathbf{P}^f$ and $\mathbf{P}^a$, respectively. The introduction of a forecast error leads to the extension of the square root with the columns of $\mathbf{T}$. Each new column introduces a new direction for the uncertainty of the state vector. To prevent the number of modes from growing to infinity, filter algorithms based on factorizations include approximations or mechanisms to avoid the growth, for example avoiding the use of dynamic noise completely, projection of $\mathbf{T}$ on the base spanned by $\mathbf{AS}$, or reduction of the number of columns whenever necessary. If $\mathbf{T}$ is to be added to the covariance square root, the degree of freedom in the system noise (rank of $\mathbf{T}$) should be of order 10–100 to keep the storage and propagation of the covariance square root within feasible bounds.

This paper presents mathematical aspects of some Kalman filters in factorized form, together with comparative numerical results obtained by applying such filters to data assimilation problems. The paper is organized as follows. In Section 2 we briefly describe some factorized filters, namely the: Reduced Rank Square Root (RRSQRT) filter, Partially Orthogonal Ensemble Kalman (POENK) filter and its variant (COFFEE), also including the Ensemble Kalman filter. In Section 3, the performance of the various algorithms is illustrated by numerical tests carried out with an advection–diffusion model application. The last section contains some concluding remarks.

2. Description of some factorized filters

2.1. RRSQRT filter

In the *Reduced Rank SQuare RooT* (RRSQRT) formulation of the Kalman filter, the covariance matrix is expressed in a limited number of (orthogonal) modes, which are re-orthogonalized and truncated to a fixed number during each