Covariance Modeling in Applications of Data Assimilation to High-dimensional Earth and Geospace Systems

Tomoko Matsuo
Ann and H.J. Smead Department of Aerospace Engineering Sciences, University of Colorado
Boulder, UCB 429, Boulder, CO 80309-0429, USA
E-mail: tomoko.matsuo@colorado.edu

Abstract. Data assimilation and inverse methods play a key role in integrating remote-sensing and in-situ Earth and Geospace observations into models of the Earth and Geospace system and subsystems, enabling weather prediction and climate projection of high societal relevance. The problem presents considerable methodological and computational challenges because of the high-dimensionality and non-linearity of the underlying dynamics of the Earth and Geospace system and the large volumes of heterogeneous observations from various observing platforms, including Low-Earth-Orbit and geosynchronous-Earth-orbit satellites as well as ground-based and airborne platforms. To make the problem tractable, most data assimilation methods resort to scalable methods derived from well-established Gaussian process-based formulation of the sequential Bayesian estimation problem. One such approach is to use approximate Kalman filters that rely on low-rank approximations of the covariance matrices. In ensemble data assimilation, the covariance is expressed in the subspace spanned by the model ensemble that evolves dynamically. Though the low-rank ensemble-based approximation allows the covariance to track non-linear dynamics with a moderate computational cost, the impact of the observations that lie outside of the subspace represented by the ensemble cannot be effectively incorporated into the assimilation analysis. One heuristic remedy widely used in numerical weather prediction is to adopt a hybrid method that uses the full-rank stationary covariance and the ensemble-based covariance. As a scalable approach to modeling the full-rank nonstationary covariance, the use of a multiresolution-based covariance is discussed.

1. Introduction
From a statistical point of view, the process of data assimilation can be viewed as sequential Bayesian estimation of geophysical random processes in which the most likely states of the system are continuously updated as observations of time-evolving states become available sequentially over time. The states of our interest, denoted by a multivariate random vector $x \in \mathbb{R}^n$, are usually observed incompletely in space and time, and the relationship of observations $y \in \mathbb{R}^p$ to $x$ at a given time $t$ is characterized as

$$y_t = H(x_t) + \epsilon_t,$$  \hspace{1cm} (1)

where $H : \mathbb{R}^n \rightarrow \mathbb{R}^p$ represents a non-linear forward operator and $\epsilon_t \in \mathbb{R}^p$ is a vector of the observational errors. The process of inferring $x$ from $y$ requires inverting the operation
represented by $\mathcal{H}$ as well as incorporating the prior information provided by the model of the Earth and Geospace system or subsystems into the inference of $x$. The relationship of $x$ to its prediction by a forecast model is given as

$$x_t = x_t^f + \eta_t,$$

where $\eta_t \in \mathbb{R}^n$ is a vector of the forecast errors. The evolution of the state $x$ from time $t-1$ to $t$ can be expressed with a non-linear operator $M : \mathbb{R}^n \to \mathbb{R}^n$ as

$$x_t^f = M(x_{t-1}^a) + \gamma_t,$$

where $\gamma_t \in \mathbb{R}^n$ is a vector of the model errors.

In the application under consideration, in which the size of $x$ can be as large as $10^8$ and the size of $y$ as large as $10^6$, most data assimilation methods end up resorting to scalable methods based on well-established Gaussian processed-based formulation of the sequential Bayesian estimation problem to make the problem tractable [1]. The prior distribution of $x$ is therefore given as $x \sim \mathcal{N}(x^f, P^f)$ where $P^f \in \mathbb{R}^{n \times n}$ is the covariance matrix of $\eta$. The likelihood of $y$ conditioned on $x$ is $y|x \sim \mathcal{N}(H(x), R)$ where $R \in \mathbb{R}^{p \times p}$ is the covariance of $e$. Under the assumption that $\mathcal{H}$ can be linear or linearized as $H = \frac{\partial \mathcal{H}}{\partial x}|_{x=x^f}$, the posterior distribution is given as $x^a \sim \mathcal{N}(x^a, P^a)$, where

$$x^a = x^f + \frac{P^f H^T}{HP^f H^T + R}(y - Hx^f),$$

$$P^a = (I - KH)P^f. \quad (4)$$

These constitute well-known formula for the Kalman filter state and covariance update. Under the assumption of linear forward and forecast models, the Kalman filter produces optimal analysis. The Kalman filter is however impractical for high-dimensional systems of the order of $O(n) = 10^6 - 10^8$, because it requires handling matrices of dimension $n \times n$. A variety of approximate Kalman filters have been developed for use in high-dimensional systems, and all of these approximate Kalman filters rely on low-rank approximations of the covariance matrices and yields sub-optimal analysis.

2. Reduce-rank ensemble-based approach

The reduced-rank methods approximate the Kalman filter by restricting the representation of covariances to a smaller subspace spanned by $q$ eigenvectors. For instance, the eigenvalue decomposition of a $n \times n$ covariance matrix of rank $q$ can be represented as, $P = V\Gamma V^T \approx V_q \Gamma_q V_q^T$, where $V_q$ is a $n \times q$ matrix that store $q$ eigenvectors. This section shows that the ensemble Kalman filter is a section of the low-rank Kalman filter.

Suppose that $P$ has rank $q << n$, and that the covariance can be expressed as

$$P = DD^T, \quad (6)$$

where $D \in \mathbb{R}^{n \times q}$. The Kalman Gain $K$ can then be re-written as

$$K = P^f H^T (R + HP^f H^T)^{-1}$$
$$= DD^T H^T (R + HDD^T H^T)^{-1}$$
$$= D(HD)^T (R + H(DH)^T)^{-1}. \quad (7)$$
The state update is given as

\[ x^a = x^f + K(y - Hx^f) = x^f + P^f H^T (R + HP^f H^T)^{-1} (y - Hx^f) = x^f + D(HD)^T (R + HD(HD)^T)^{-1} (y - Hx^f). \] (8)

The analysis increment is a linear combination of the columns of \( D \), and is thus confined to the subspace of dimension \( q \).

The ensemble Kalman filters are reduced-rank Kalman filters, with their covariance approximated as the sample covariance \( C \).

\[ P \approx C = \frac{1}{m-1} GG^T, \] (9)

where \( m \) is the size of the ensemble and \( G \in \mathbb{R}^{n \times m} \) is a matrix with each column filled with a member of the ensemble of state vector \( x \) as

\[ G = [ (x^{(1)} - \mu) \ (x^{(2)} - \mu) \ \ldots \ (x^{(m)} - \mu) ] , \] (10)

where \( x^{(m)} \) denotes the \( m \)th-member of \( x \) and \( \mu \in \mathbb{R}^n \) is the ensemble (sample) mean state. Note that the rank of \( C \) is \( m - 1 \) at most. Usually \( m \ll n \), so \( P \) used in EnKF is rank deficient. For \( m \ll n \) case, \( C \) is not uniquely determined. In actual implementation of EnKFs, the full matrix \( C \) is never explicitly computed and Equation (8) is not used to update the forecast ensemble members \([2, 3]\). The update step can be applied to each ensemble member in a stochastic or deterministic fashion to make the sample statistics consistent with the analysis update of the state and covariance stated in Equations (4) and (5). Deterministic methods have the advantage of avoiding sampling issues associated with the use of “perturbed observations” in stochastic analysis update methods. Though the low-rank ensemble-based approximation allows the covariance to track non-linear dynamics with a moderate computational cost, the impact of the observations that lie outside of the subspace represented by the ensemble cannot be effectively incorporated into the assimilation analysis. One heuristic remedy widely used in numerical weather prediction is to adopt a hybrid method that uses the full-rank stationary covariance and the ensemble-based covariance.

3. Full-rank wavelet-based approach

Modeling and estimation of a full-rank nonstationary covariance for high-dimensional systems pose computational and methodological challenges. For instance, representation of covariance with multiresolution compactly-supported basis functions such as wavelets offers a practical approach scalable to high-dimensional problems \([4]\), and complements the reduce-rank ensemble-based approach described in the previous section.

Main advantages of the wavelet-based approach are its flexibility in representing departures from stationarity, and sparsity in the covariance matrix among wavelet basis coefficients. Suppose that \( x \) is composed of one type of the geophysical state on \( n \) regular model grid locations and \( W \in \mathbb{R}^{n \times n} \) is a matrix where \( n \) wavelet basis functions are evaluated at \( n \) locations. If \( W \) has full rank it is always possible that the covariance matrix \( P \) can be decomposed as

\[ P = WGG^T W^T , \] (11)

where \( G \in \mathbb{R}^{n \times n} \) is a real nonsingular symmetric matrix, and \( GG^T \) is the covariance of wavelet coefficients. Because of the compact support of wavelet basis functions, most of the scaling and coarse-scale wavelet functions are uncorrelated, justifying a sparse approximation of the
wavelet coefficient covariance, as well as $G$. As shown in [4] for wavelet coefficients of stationary processes, the covariance among coefficients decays geometrically with the scale of the coefficients where the rate of decay depends on the smoothness of the underlying random processes, and therefore the insignificant off-diagonal elements of $G$ can be set to equal zero such that $G$ has $O(n)$ rather than $O(n^2)$ nonzero elements, represented by $\theta$, yet has full rank.

The estimation of covariance parameters $\theta$ may not be trivial especially for high-dimensional problems. One way is to further parameterize $\theta$ with a smaller number of hyper-parameters and estimate them using the maximum-likelihood method. Another strategy is a non-parametric approach that takes an advantage of the scalability of algorithms facilitated by the availability of fast transforms for wavelet bases on regular grids. As demonstrated in [4], the covariance parameters $\theta$ can be estimated using a Monte-Carlo approach that draws on the expectation maximization algorithm. Let’s suppose that $x$ is partly observed, or composed of the observed state $x_1$ and unobserved state $x_2$, normally distributed as

$$
x = \begin{pmatrix} x_1 \\
x_2 \end{pmatrix} \sim \mathcal{N}(0, P_{\theta}) \quad \text{where} \quad P_{\theta} = \tilde{W}G^2(\theta)W^T.
$$

(12)

Then $\theta$ can be estimated iteratively by maximizing the following expression [5]:

$$
Q(\theta, \theta^*) = E[\mathcal{L}(x, \theta)|x_1, \theta^*] \\
\approx \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}\left( \begin{pmatrix} x_1 \\
x_2^{(m)} \end{pmatrix}, \theta \right),
$$

(13)

where $\theta^*$ denotes an initial guess for the covariance parameters and $\mathcal{L}$ represents a (complete data) log-likelihood function. $x_2^{(m)}$ is here sampled from $x_2^{(m)} \sim [x_2|x_1, G^2(\theta^*)W^T]$. Specific application examples can be found for surface ozone data in [4] and for upper atmosphere plasma data in [6].

4. Summary
Applications of data assimilation and inverse methods to Earth and Geospace system present considerable methodological and computational challenges because of the high-dimensionality and non-linearity of the underlying dynamics of the Earth and Geospace system, and the large volumes of heterogeneous observations. To make the problem tractable most data assimilation methods resort to scalable methods derived from well-established Gaussian process-based formulation of the sequential Bayesian estimation problem. A key component of a Gaussian process model is the covariance. The paper presents two scalable approaches to model the nonstationary covariance matrices: low-rank ensemble-based covariance and full-rank wavelet-based covariance.

Acknowledgments
The present study is supported by NASA Award NNX14AI17G.

References
The present study is supported by NASA Award NNX14AI17G.
[5] Wei, G.C., and M. A. Tanner 1990, A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms, *Journal of the American Statistical Association*, 85, 699-704.

[6] Sun, Y.-Y., T. Matsuo, E. A. Araujo-Pradere, and J. Y. Liu 2013, *Ground-based GPS observation of SED associated irregularities over CONUS*, *J. Geophys. Res. Space Physics*, 118, 2478-2489.