Journal of Computational Physics 257 (2014) 782–812

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

An ensemble Kalman filter for statistical estimation of physics constrained nonlinear regression models

John Harlim [a,*], Adam Mahdi [b], Andrew J. Majda [c]

[a] Department of Mathematics and Department of Meteorology, the Pennsylvania State University, University Park, PA 16802, Unites States
[b] Department of Mathematics, North Carolina State University, Raleigh, NC 27695, United States
[c] Department of Mathematics and Center for Atmosphere and Ocean Science, Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, United States

ARTICLE INFO

Article history:
Received 27 June 2013
Received in revised form 1 October 2013
Accepted 4 October 2013
Available online 23 October 2013

Keywords:
Ensemble Kalman filter
Nonlinear regression models
Parameter estimation of stochastic differential equations
Multi-level models

ABSTRACT

A central issue in contemporary science is the development of nonlinear data driven statistical–dynamical models for time series of noisy partial observations from nature or a complex model. It has been established recently that ad-hoc quadratic multi-level regression models can have finite-time blow-up of statistical solutions and/or pathological behavior of their invariant measure. Recently, a new class of physics constrained nonlinear regression models were developed to ameliorate this pathological behavior. Here a new finite ensemble Kalman filtering algorithm is developed for estimating the state, the linear and nonlinear model coefficients, the model and the observation noise covariances from available partial noisy observations of the state.

Several stringent tests and applications of the method are developed here. In the most complex application, the perfect model has 57 degrees of freedom involving a zonal (east–west) jet, two topographic Rossby waves, and 54 nonlinearly interacting Rossby waves; the perfect model has significant non-Gaussian statistics in the zonal jet with blocked and unblocked regimes and a non-Gaussian skewed distribution due to interaction with the other 56 modes. We only observe the zonal jet contaminated by noise and apply the ensemble filter algorithm for estimation. Numerically, we find that a three dimensional nonlinear stochastic model with one level of memory mimics the statistical effect of the other 56 modes on the zonal jet in an accurate fashion, including the skew non-Gaussian distribution and autocorrelation decay. On the other hand, a similar stochastic model with zero memory levels fails to capture the crucial non-Gaussian behavior of the zonal jet from the perfect 57-mode model.

1. Introduction

A central issue in contemporary science is the development of data driven statistical–dynamical models for the time series of a partial subset of observed variables, $u_1(t) \in \mathbb{R}^{N1}$, which arise from observations of nature or from an extremely complex physical model [1–11]. This is an important issue in systems ranging from bio-molecular dynamics to climate science to engineering turbulence. Examples of such data driven dynamical models are multi-level linear autoregressive models with external factors [2,6] as well as ad-hoc quadratic nonlinear regression models [12–15]. Such purely data driven ad-hoc regression models are developed through various criteria to fit the data but by design, do not respect the underlying physical dynamics of the partially observed system or the causal processes in the dynamics; nevertheless, the goal of purely

* Corresponding author.
E-mail addresses: jharlim@psu.edu (J. Harlim), amahdi@ncsu.edu (A. Mahdi), jonjon@cims.nyu.edu (A.J. Majda).

0021-9991/$ – see front matter
http://dx.doi.org/10.1016/j.jcp.2013.10.025

data driven statistical modeling is to provide simplified low order models with high predictive skill for central features of the underlying physical system and not just fit (or over-fit, see [2]) the given data. Indeed, the work in [16] provides rigorous mathematical theory and examples with straightforward numerical experiments where the ad-hoc quadratic multi-level regression models proposed in [12–14] necessarily have non-physical finite-time blow-up of statistical solutions and also pathological behavior of the related invariant measure even though these models match a long time series of the observed data produced from the physical model with high accuracy.

In [17] a new class of physics constrained multi-level regression models were proposed to systematically ameliorate the above difficulties. These models begin with an observed time series for a variable $u_1 \in \mathbb{R}^{N_1}$ and augment u_1 with a hidden variable u_2 so that primary nonlinear energy conserving interactions occur in the u variables with $u = (u_1, u_2) \in \mathbb{R}^N$, $N = N_1 + N_2$, while the stochastic memory effects of noise are linear, conditional on the state u. If $\Pi_2 u = (0, u_2)^\top$ denotes the projection on u_2, then the *Physics Constrained Multi-Level Regression Models* developed in [17] have the structural form,

$$\begin{aligned} \frac{du}{dt} &= Lu + B(u,u) + F + \Pi_2 r_1, \\ \frac{d\vec{r}}{dt} &= Qu + \tilde{A}\vec{r} + \sigma \dot{W}, \end{aligned} \tag{1}$$

where the nonlinear interactions $B(u,u)$ conserve energy so that

$$\langle u, B(u,u) \rangle = 0, \tag{2}$$

for a suitable inner product $\langle \cdot,\cdot \rangle$; and $\dot{W}$ is a white noise vector with noise matrix σ. The interactions of $(r_1, \ldots, r_p)$ with $r_j \in \mathbb{R}^{N_2}$, where p memory levels are characterized by the triangular matrix $\tilde{A}$; their component are given as follows,

$$\begin{aligned} \frac{dr_i}{dt} &= Q_i u + \sum_{j=1}^{i} \tilde{a}_{i,j} r_j + r_{i+1}, \quad 1 \leqslant i \leqslant p-1, \\ \frac{dr_p}{dt} &= Q_p u + \sum_{j=1}^{p} \tilde{a}_{p,j} r_j + \sigma \dot{W}. \end{aligned}$$

Here p denotes the number of memory levels and $p = 0$ denotes the special *zero-memory level model*

$$\frac{du}{dt} = Lu + B(u,u) + F + \Pi_2 \sigma \dot{W}. \tag{3}$$

Rigorous theorems in [17] show that such models do not blow-up (under certain hypotheses) and have nice invariant measures. Guidelines involving stability of L as well as observability and controllability of (1) for the primary variables $u = (u_1, u_2)$ when only u_1 is observed are also given in [17] together with a counterexample with statistical blow-up with neutral stability.

The practical issue treated here involves the inference of an appropriate stochastic model in (1) with (2) from partial knowledge involving observations of u_1 alone, contaminated by noise. This estimation problem involves determining the coefficients L, Q, $\tilde{A}$, B and estimating the noise amplitude σ in (1) and the observation noise covariance, besides estimating u_2.

The filtering or estimation algorithm implemented in [17] was based on using Extended Kalman Filter (EKF) combined with Belanger's method [18] for noise and parameter estimation. This algorithm was used successfully in [17] to estimate the parameters in a reduced stochastic model for the first Fourier mode of the Truncated Burgers–Hopf (TBH) model [19] as well as related nonlinear oscillators with memory. In Section 2 of this paper we provide an instructive motivating example, showing the skill of this algorithm with complete observation of u and a spectacular failure of this algorithm due to linear instability with partial observation when only u_1 is observed. A new finite Ensemble Kalman Filter (EnKF) based scheme for estimating all the parameters L, B, Q, $\tilde{A}$ and the noise σ in the models from (1) is developed in Section 3.

An important area for application of the regression models is low frequency climate variability and a family of stringent paradigm models for this behavior [9,20], which are studied through the new estimation algorithm in Section 4. The most complex problem considered in Section 4 involves a perfect model with 57 degrees of freedom with a large scale zonal (east–west) jet, two topographic Rossby waves, and 54 nonlinearly interacting Rossby waves; the perfect models have significant non-Gaussian statistics with blocked and unblocked regimes of the zonal jet due to stochastic backscatter from the Rossby waves even though the total perfect system exactly conserves energy. We only observe the zonal jet contaminated by noise and seek a physics constrained stochastic model with memory with form in (1) which mimics the statistics of the 57-mode model. We apply our EnKF-based algorithm for parameter estimation introduced in Section 3 and find that a three dimensional nonlinear stochastic model with one level of memory, $p = 1$, is sufficient to mimic the statistical effects of the other 56 modes on the zonal jet u in an accurate fashion, including both the skewed non-Gaussian distribution for u and its autocorrelation decay. On the other hand, a similar model with zero level memory fails to capture any of these crucial non-Gaussian behavior for the zonal jet from the perfect 57-mode model.