Journal of Computational Physics 308 (2016) 305–321

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

Semiparametric modeling: Correcting low-dimensional model error in parametric models

Tyrus Berry [a,*], John Harlim [a,b]

[a] Department of Mathematics, the Pennsylvania State University, 109 McAllister Building, University Park, PA 16802-6400, USA
[b] Department of Meteorology, the Pennsylvania State University, 503 Walker Building, University Park, PA 16802-5013, USA

ARTICLE INFO

Article history:
Received 23 June 2015
Received in revised form 20 November 2015
Accepted 19 December 2015
Available online 29 December 2015

Keywords:
Diffusion maps
Model error
Diffusion forecast
Semiparametric modeling
Kalman filter
Nonparametric modeling

ABSTRACT

In this paper, a semiparametric modeling approach is introduced as a paradigm for addressing model error arising from unresolved physical phenomena. Our approach compensates for model error by learning an auxiliary dynamical model for the unknown parameters. Practically, the proposed approach consists of the following steps. Given a physics-based model and a noisy data set of historical observations, a Bayesian filtering algorithm is used to extract a time-series of the parameter values. Subsequently, the *diffusion forecast* algorithm is applied to the retrieved time-series in order to construct the auxiliary model for the time evolving parameters. The semiparametric forecasting algorithm consists of integrating the existing physics-based model with an ensemble of parameters sampled from the probability density function of the diffusion forecast. To specify initial conditions for the diffusion forecast, a Bayesian semiparametric filtering method that extends the Kalman-based filtering framework is introduced. In difficult test examples, which introduce chaotically and stochastically evolving hidden parameters into the Lorenz-96 model, we show that our approach can effectively compensate for model error, with forecasting skill comparable to that of the perfect model.

1. Introduction

A long-standing problem in modeling natural processes is model error. This issue is attributed to our incomplete understanding of the underlying physics and our lack of computational resources to resolve physical processes at various time and length scales. Serious efforts to combat this issue have been proposed with parametric modeling approaches (for examples see [1–8] just to name a few). In fact, it was rigorously shown in a simple context that it is possible to fully compensate for model error arising from unresolved scales with appropriate stochastic parametric models. In particular, it was shown that one can simultaneously obtain optimal filtering and climatological statistical estimates assuming that one has access to the *right* stochastic parametric model [9] which can be quite demanding in general.

While stochastic parameterization techniques have been successful in many particular problems, this approach also has another limitation beyond the selection of the parametric form. Namely, determining the parameters in these models from a limited amount of noisy data can be nontrivial, especially when the parameters are not directly observed. This was shown in [9], which found that even when the appropriate parametric form is known, the resulting forecasting skill is sensitive to the stochastic parameter estimation scheme. While a class of adaptive parameter estimation schemes [8,10,11] produce

* Corresponding author.
E-mail addresses: thb11@psu.edu (T. Berry), jharlim@psu.edu (J. Harlim).

http://dx.doi.org/10.1016/j.jcp.2015.12.043

much more accurate results relative to the standard linear regression fitting method [12], they are numerically expensive when the dimensionality of the problems increases.

The main goal of the present paper is to introduce a modeling paradigm that avoids the two fundamental issues in parametric modeling; namely the choice of parametric form and the fitting of parameters. Our approach is philosophically motivated by the semiparametric modeling method for *time-independent* densities, which is well known in the statistics literature as a flexible work-around for overcoming the curse-of-dimensionality [13]. Classical nonparametric models such as histograms or kernel density estimates can reconstruct an unknown *time-independent* density from a training data set, however the amount of data needed grows exponentially in the dimensionality of the data [14]. Alternatively, estimating parameters of a parametric statistical model can often recover a high-dimensional density from a relatively small data set. However, imposing the rigid structure of the parametric form opens up the possibility of model error. Semiparametric models in statistics attempt to take an intermediate road by choosing a parametric form but allowing the parameters to vary and building nonparametric models for the parameters. Choosing the structure of the parametric form allows one to encode the prior knowledge about the data set, hopefully reducing the intrinsic dimensionality of the parameter space to the point where a nonparametric model can be used to capture the remaining variability. In this paper, we will extend this statistical idea of semiparametric modeling to estimate *time-dependent* densities in order to improving forecasting and filtering of dynamical systems.

In particular, the semiparametric approach introduced in the present paper will be based on a recently developed nonparametric modeling approach for learning a dynamical system from a training data set. This nonparametric approach, called the *diffusion forecast*, was introduced for stochastic gradient flow systems in [15] and has been extended to model diffusion processes on smooth manifolds [16]. Essentially, the diffusion forecast is a data-driven method for solving the Fokker–Planck equations of any autonomous diffusion processes on a smooth manifold without requiring knowledge of the corresponding drift coefficients, diffusion tensor, or even the underlying manifold itself. It was shown in [17] that diffusion forecast provides high forecasting skill, and is competitive with some physics-based parametric models for predicting energetic modes of turbulent dynamical systems using relatively small training data sets (on the order of 10^3 data points).

The novel aspect of the present paper, beyond [16,17], is to overcome the curse-of-dimensionality which restricts the diffusion forecast to modeling low dimensional dynamics. We overcome this limitation by assuming that we have an approximate or incomplete parametric model and then using the diffusion forecast to fill in the missing components, in other words, to correct the model error. By assuming that the parametric model captures most of the variability, so that the model error is low-dimensional, the nonparametric model becomes feasible. Simultaneously, from the perspective of the modeling community, the goal of semiparametric modeling is to overcome model error by using historical data to 'correct' the existing physical models. These two perspectives indicate that semiparametric modeling has the potential to seamlessly blend the strengths of the parametric and nonparametric modeling approaches, thereby overcoming their complementary weaknesses.

Combining the partial parametric model with the nonparametric, diffusion forecasting, model requires new semiparametric forecasting and filtering algorithms. Since the framework is introduced for correcting model error in existing parametric models, an important consideration is that we wish to maintain as much of the standard parametric ensemble forecasting and filtering framework as possible. To achieve these goals, in Section 2 we will define the form of the model error that we will be able to address here, and we briefly review some simple stochastic parametric models which have been successful for certain types of model error. We will assume that the model error can be described by dynamically varying certain parameters in the parametric model, and that the evolution of these parameters is independent of the state variables of the parametric model. In Section 3 we introduce a semiparametric forecasting algorithm which combines diffusion forecast of [16] with a standard ensemble forecast method for the parametric model. For simplicity and clarity, in Section 3 we assume that a training data set of the varying parameters is available and that we are given noisy initial conditions for forecasting. In Sections 4 and 5, we will discuss additional strategies for dealing with the more realistic scenario when we only have noisy observations of the state variables of the parametric model, which is the common situation in practice. In particular, in Section 4 we use an adaptive filtering method developed in [10,9] to extract a time series of the varying parameters from the noisy observations. In Section 5 we introduce a semiparametric filter which combines an Ensemble Kalman Filter (EnKF) for the parametric model with the nonparametric model (learned from the time series recovered in Section 4) in order to find initial conditions from the noisy observations. As a proof-of-concept example, we will demonstrate the semiparametric filter and forecast on the Lorenz-96 model with model error arising from a parameter which evolves according to low-dimensional chaotic or stochastic models. We close this paper with a short summary in Section 6.

2. Problem statement and background

We consider the problem of forecasting in the presence of model error. We assume that we are given a noisy time series of observations from a known dynamical model,

$$\dot{x} = f(x, \theta), \tag{1}$$

with an observation function,

$$y = h(x, \theta), \tag{2}$$

that depends on parameters θ which evolve according to an unknown stochastic model,