Physica D 320 (2016) 57–76

Contents lists available at ScienceDirect

Physica D

journal homepage: www.elsevier.com/locate/physd

Forecasting turbulent modes with nonparametric diffusion models: Learning from noisy data

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

[a] Department of Mathematics, The Pennsylvania State University, University Park, PA 16802, United States

[b] Department of Meteorology, The Pennsylvania State University, University Park, PA 16802, United States

HIGHLIGHTS

- We examine the skill of the diffusion forecast model in predicting turbulent modes.
- A novel Bayesian filtering method is introduced to initialize the forecast given noisy data.
- The diffusion forecast is competitive with the perfect model given the same set of noisy data.
- A test on geophysical turbulence indicates that the long-term forecasts are unbiased.

ARTICLE INFO

Article history:
Received 27 January 2015
Received in revised form
9 December 2015
Accepted 31 January 2016
Available online 8 February 2016
Communicated by H.A. Dijkstra

Keywords:
Nonparametric forecasting
Kernel methods
Diffusion maps
Diffusion models
Time-lagged embedding
Diffusion forecast

ABSTRACT

In this paper, we apply a recently developed nonparametric modeling approach, the "diffusion forecast", to predict the time-evolution of Fourier modes of turbulent dynamical systems. While the diffusion forecasting method assumes the availability of a noise-free training data set observing the full state space of the dynamics, in real applications we often have only partial observations which are corrupted by noise. To alleviate these practical issues, following the theory of embedology, the diffusion model is built using the delay-embedding coordinates of the data. We show that this delay embedding biases the geometry of the data in a way which extracts the most stable component of the dynamics and reduces the influence of independent additive observation noise. The resulting diffusion forecast model approximates the semigroup solutions of the generator of the underlying dynamics in the limit of large data and when the observation noise vanishes. As in any standard forecasting problem, the forecasting skill depends crucially on the accuracy of the initial conditions. We introduce a novel Bayesian method for filtering the discrete-time noisy observations which works with the diffusion forecast to determine the forecast initial densities.

Numerically, we compare this nonparametric approach with standard stochastic parametric models on a wide-range of well-studied turbulent modes, including the Lorenz-96 model in weakly chaotic to fully turbulent regimes and the barotropic modes of a quasi-geostrophic model with baroclinic instabilities. We show that when the only available data is the low-dimensional set of noisy modes that are being modeled, the diffusion forecast is indeed competitive to the perfect model.

1. Introduction

A long-standing issue in modeling turbulent dynamics is the so-called turbulent closure problem (see e.g. [1]) where the goal is to find a set of effective equations to represent low-order statistics of the coarse-grained variables of interest. The main difficulty of this problem is largely due to the infinite dimensionality and nontrivial coupling of the governing equations of the statistics. In order to predict a few lower-order statistics of some resolved variables, common closure approaches were developed using physical insights to choose a parametric ansatz to represent the feedback from the unresolved scales (see e.g., [2] for various closure approximations for predicting passive scalar turbulence and [3–6] for various stochastic modeling approaches for predicting geophysical turbulence).

Despite these successes, the parametric modeling approaches have practical issues due to model error when the necessary physical insights are not known. If the parametric model (or ansatz) is not chosen appropriately, one can end up with a model with poor predictive skills (or even with solutions which diverge

* Corresponding author.
E-mail addresses: tbh11@psu.edu (T. Berry), jharlim@psu.edu (J. Harlim).

http://dx.doi.org/10.1016/j.physd.2016.01.012

catastrophically) even when the parameters can be obtained by a standard regression fitting procedure [7]. Moreover, even when an appropriate parametric form is chosen, specifying the parameters from noisy observations of the physical variables can be nontrivial since the parameters are typically not directly observed. Indeed, it was shown that an appropriate parameterization scheme is crucial for accurate filtering and equilibrium statistical prediction even when the parametric forms are appropriately chosen [8].

Recently, a nonparametric modeling approach, called the *diffusion forecast*, for predicting the evolution of the probability density of low-dimensional dynamical system was introduced in [9]. The approach of [9] can be intuitively viewed as extending the standard nonparametric statistical models (such as kernel density estimates) which are used to estimate time-independent densities [10]. The key idea behind the diffusion forecast is to use a basis of smooth functions to represent probability densities, so that the forecast model becomes a linear map in this basis. Numerically, this linear map is estimated by exploiting a rigorous connection between the discrete time shift map and semi-group solution associated to the backward Kolmogorov equation. In [9], it was shown that the resulting model estimates the semigroup solutions of the generator of the underlying dynamics in the limit of large data. Moreover, the smooth basis is defined on the training data set, using the diffusion maps algorithm [11,12], which means that the data requirements only depend on the intrinsic dimensionality of the dynamics.

In this paper, we test this nonparametric modeling approach as a method of forecasting noisy observations of Fourier modes from a selection of well-studied high-dimensional dynamical systems in various turbulent regimes. A novel aspect of this paper is that we consider building a forecasting model given a noisy training data set consisting of partial observations of the dynamics, as is common in practical applications, in contrast to the work in [9] which used noiseless full observations to train the diffusion forecasting model. A key ingredient for solving initial value problems in any forecasting problem is accurate initial conditions. While initial conditions were assumed to be given in [9], in this paper, we introduce a novel Bayesian filtering method to iteratively assimilate each observation and find the initial probability densities given all of the past noisy observations up to the corresponding initial time.

We should note that the diffusion forecasting method [9] could be naively applied to signals corrupted by observation noise, however the resulting nonparametric model would implicitly include the observation noise in the model, which would limit the forecast skill compared to treating the noise as a separate process. Treating the noise as a separate process requires first learning the 'correct' model from the noisy training data set, and then generating 'clean' initial conditions for forecasting from the noisy observations. In [13–15] it was shown that applying diffusion maps to the delay-embedded data reduces the influence of the noise on the diffusion maps basis. Building upon the work in [9], we apply the theory of [13] to show that building the nonparametric model using the delay-embedded data biases the geometry of the data in a way which extracts the most predictable component of the dynamics. We extend the theory of [13] by giving a rigorous justification for the reduction of the influence of independent additive observation noise on the resulting diffusion forecast model.

One interesting question which we address here is whether it is possible to build a skillful nonparametric forecasting model for a turbulent mode given only a small amount of noisy training data, when the true dynamics are solutions of a high-dimensional dynamical system with chaotic behavior. This question arises because the nonparametric model has a practical limitation in terms of modeling dynamics with high-dimensional attractors, namely: it will require an immense amount of data to unwind the attractors since the required data increases exponentially as a function of the dimension of the attractor. Moreover, even given a sufficiently large data set, the required computational power would be a limiting factor since the diffusion maps algorithm requires storing and computing eigenvectors of a sparse $N \times N$ matrix, where N is the number of data points. Constrained by a small data set, the curse-of-dimensionality implies that we cannot unwind the full high-dimensional attractor. We attempt to circumvent the curse-of-dimensionality by decomposing the data into Fourier modes in the hope that delay reconstruction of each mode projects onto a different component of the dynamics. We do not claim that the Fourier decomposition can completely resolve this issue but we will numerically demonstrate that the Fourier decomposition will map an isotropic turbulent field in the spatial domain (which implies that each spatial component is as predictable as any other spatial component) to a coordinate system in which some modes are more predictable than others. Of course, the standard theory of embedology [16] suggests that the delay-embedding of a single Fourier mode would reconstruct the entire high-dimensional attractor, which would again be inaccessible to our nonparametric model due to the curse-of-dimensionality. This would suggest that nothing could be gained by building separate models based on delay-embedding of each mode. However, the full attractors reconstructed from each mode are only equivalent in a topological sense, and the geometries of these reconstructed attractors are dramatically different. The biased geometry influences the nonparametric model of [9] through the use of the diffusion map algorithm which is known to preserve the geometry which the data inherits from the embedding space [11,13]. The diffusion maps algorithm preserves the biased geometry of the delay embedding as was shown in [13]; and we will see that this biased geometry projects the full dynamical system onto the most stable components of the dynamics in the direction of the chosen observations. When we apply the nonparametric model of [9] using the basis arising from this biased geometry, we find improved forecasting skill and robustness to observation noise.

The remainder of this paper is organized as follows. In Section 2, we introduce the problems under consideration and establish the necessary background, including a brief overview of the nonparametric modeling approach introduced in [9] as well as a discussion on how the theory of [13] is applied to mitigate the effect of noise on the model. We conclude Section 2 by introducing the iterative Bayesian filter which we use to generate initial conditions for forecasting with the nonparametric model. In Section 3, we numerically compare predicting Fourier modes of the Lorenz-96 model in various chaotic regimes using the nonparametric model with the persistence model, perfect model, and various parametric models, including the autoregressive models of order-1 (MSM [17]). In Section 4, we numerically compare the nonparametric model with a stochastic model with additive and multiplicative noises (SPEKF model [18,19]) in predicting the barotropic modes of a geostrophic turbulence. We close this paper with a short summary in Section 5.

2. Nonparametric diffusion modeling

Let $u(x, t) \in \mathbb{R}^s$ be solutions of an ergodic system of nonlinear PDEs,

$$\frac{\partial u}{\partial t} = \mathcal{A}(u), \tag{1}$$

where $\mathcal{A}$ denotes nonlinear differential operators, for smooth initial conditions $u(x, 0)$ and periodic boundary conditions on a non-dimensionalized periodic domain $x \in [0, 2\pi]^n$. To simplify