Nonparametric forecasting of low-dimensional dynamical systems

Tyrus Berry,1 Dimitrios Giannakis,2 and John Harlim1,3

1Department of Mathematics, the Pennsylvania State University, University Park, PA 16802-6400, USA
2Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, USA
3Department of Meteorology, the Pennsylvania State University, University Park, PA 16802-5013, USA

(Dated: November 20, 2014)

This letter presents a non-parametric modeling approach for forecasting stochastic dynamical systems on low-dimensional manifolds. The key idea is to represent the discrete shift maps on a smooth basis which can be obtained by the diffusion maps algorithm. In the limit of large data, this approach converges to a Galerkin projection of the semigroup solution to the underlying dynamics on a basis adapted to the invariant measure. This approach allows one to quantify uncertainties (in fact, evolve the probability distribution) for non-trivial dynamical systems with equation-free modeling. We verify our approach on various examples, ranging from an inhomogeneous anisotropic stochastic differential equation on a torus, the chaotic Lorenz three-dimensional model, and the Niño-3.4 data set which is used as a proxy of the El-Niño Southern Oscillation.

A significant challenge in modeling is to account for physical processes which are often not well understood but for which large data sets are available. A standard approach is to perform regression fitting of the data into various parametric models. While this approach is popular and successful in many applied domains, the resulting predictive skill can be sensitive to the choice of models, the parameter fitting algorithm, and the complexity of the underlying physical processes. An alternative approach is to avoid choosing particular models and/or parameter estimation algorithms and instead apply a non-parametric modeling technique. In particular, nonparametric modeling based on local linearization of discrete shift maps on paths has been successful in predicting mean statistics of data sets generated by dynamical systems with low-dimensional attractors.

In this letter, we generalize this nonparametric approach to quantify the evolving probability distribution of the underlying dynamical system. The key idea is to project the shift map on a set of basis functions that are generated by the diffusion maps algorithm with a variable bandwidth diffusion kernel. This approach has connections with a recently developed family of kernels, which utilize small shifts of a deterministic time series to estimate the dynamical vector field. The method also generalizes a recently introduced non-parametric modeling framework for gradient systems to inhomogeneous stochastic systems having non-gradient drift and anisotropic diffusion. Consider a dynamical system,

$$dx = a(x) dt + b(x) dW_t$$ (1)

where $W_t$ is a standard Brownian process, $a$ a vector field, and $b$ a diffusion tensor, all defined on a manifold $\mathcal{M} \subset \mathbb{R}^n$. Given a time series $x_i = x(t_i)$, generated by at discrete times $\{t_i\}_{i=1}^{N+1}$, we are interested in constructing a forecast model so that given an initial density $p_0(x)$ we can estimate the density $p(x, t) = e^{t\mathcal{L}^*} p_0(x)$ at time $t > 0$, without the Fokker-Planck operator, $\mathcal{L}^*$, of and without knowing or estimating $a$ and $b$. We will assume that is ergodic so that $\{x_i\}$ can be considered samples from the invariant measure $p_{eq}(x)$ of $\mathcal{H}$. Note that all probability densities are relative to a volume form $\mu$ which $\mathcal{M}$ inherits from the ambient space, and adjoints are with respect to the inner product $\langle \cdot, \cdot \rangle$ of $L^2(\mathcal{M}, \mu)$.

A straightforward application of the diffusion maps algorithm with a variable bandwidth kernel on the data set $x_i$ produces a basis of eigenfunctions of the generator, $\mathcal{L}$, of a stochastically forced gradient flow with potential $U(x) = -\log(p_{eq}(x))$. Let $\lambda_j$ and $\varphi_j$ be eigenvalues and eigenfunctions of $\mathcal{L}$; these eigenfunctions are orthogonal on $L^2(\mathcal{M}, p_{eq})$. Numerically, we obtain $\varphi_j(x_i)$ as eigenvectors of a stochastic matrix, constructed by evaluating a variable bandwidth kernel of all pairs of data points and then applying an appropriate normalization. For gradient flow systems, it is easy to check that $\psi_j = p_{eq} \varphi_j$ are the eigenfunctions of the adjoint operator $\mathcal{L}^*$, which are orthogonal on $L^2(\mathcal{M}, p_{eq-1})$. Let us write the solution,

$$p(x, \tau) = e^{\tau \mathcal{L}^*} p_0(x),$$

in the following form,

$$p(x, \tau) = \sum_{l=1}^{\infty} (e^{\tau \mathcal{L}^*} p_0, \psi_l)_{p_{eq}} \psi_l(x) = \sum_{l=1}^{\infty} (p_0, e^{\tau \mathcal{L}^*} \varphi_l)_{p_{eq}} \psi_l(x).$$

Similarly, the initial density is $p_0(x) = \sum_j c_j(0) \psi_j(x)$, where $c_j(0) = (p_0, \psi_j)_{p_{eq}}$. We therefore obtain,

$$p(x, \tau) = \sum_{l=1}^{\infty} \sum_{j=1}^{\infty} c_j(0) A_{lj}(\tau) p_{eq}(x) \varphi_l(x),$$ (2)

where $A_{lj}(\tau) := \langle \varphi_j, e^{\tau \mathcal{L}^*} \varphi_l \rangle_{p_{eq}}$. With this formulation, the evolution of the dynamics can be determined if we can compute $A_{lj}$ and the key contribution in this paper is in approximating $A_{lj}$ based on the following observation.

Our approach is stimulated by a rigorous connection between the shift map, $S$, and the semi-group solution, $e^{\tau \mathcal{L}^*}$, of the underlying dynamical system. To see this, let us define the shift map, $S\varphi_i(x_i) = \varphi_i(x_{i+1})$, for the eigenfunction $\varphi_i$ defined on the data set. Applying Itô’s
Lemma to the smooth function $\varphi_l(x)$, one can show that,

$$S\varphi_l(x_i) = e^{t_i}\varphi_l(x_i) + \int_{t_i}^{t_{i+1}} \nabla \varphi_l^\top b dW_s + \int_{t_i}^{t_{i+1}} \left(L\varphi_l(x(s)) - \mathbb{E}[L\varphi_l(x(s))]\right) ds,$$

where $t = t_{i+1} - t_i$ and the expectation is with respect to paths of $\mathbf{1}$ conditional to $x(t_i)$. Setting $\mathbf{1}_{ij} \equiv \langle \varphi_j, S\varphi_l \rangle_{p_{\text{eq}}}$, (3) implies,

$$\hat{A}_{ij} = A_{ij} + \int_{t_i}^{t_{i+1}} \langle \varphi_j, L\varphi_l - \mathbb{E}[L\varphi_l] \rangle_{p_{\text{eq}}} dW_s + \int_{t_i}^{t_{i+1}} B_{ij} ds,$$

where $B_{ij} \equiv \langle \varphi_j, L\varphi_l - \mathbb{E}[L\varphi_l] \rangle_{p_{\text{eq}}}$ is bounded so that the final integral is order-$\tau$. Note that $\hat{A}_{ij}$ can be approximated by the Monte-Carlo integral,

$$\hat{A}_{ij} = \frac{1}{N} \sum_{i=1}^{N} \varphi_j(x_i)\varphi_l(x_{i+1}).$$

Assuming the diffusion tensor is bounded above, $\|b\| \leq b_0$, in a tensor norm, we can bound the inner product,

$$|\langle \varphi_j, \nabla \varphi_l^\top b \rangle_{p_{\text{eq}}} | \leq b_0 \|\varphi_j\|_{p_{\text{eq}}} \|\nabla \varphi_l\|_{p_{\text{eq}}} = b_0 \lambda_l,$$

by the Cauchy inequality. The last equality in (4) follows from $\{\varphi_l\}$ being orthonormal on $L^2(M, p_{\text{eq}})$ which successively minimize $\|\nabla f\|_{p_{\text{eq}}}$ with minima $\|\nabla \varphi_l\|_{p_{\text{eq}}} = \lambda_l$.

With this observation, we approximate $A_{ij}$ in (3) with $\hat{A}_{ij}$ defined in (4), such that, the diffusion forecast is defined as follows:

$$p(x, \tau) \approx \sum_{i=1}^{\infty} p_{\text{eq}}(x) \varphi_l(x) \sum_{j=1}^{\infty} \hat{A}_{ij}(\tau) c_j(0).$$

Computationally, we will make a Galerkin truncation onto $M$-eigenfunctions such that $\hat{A}$ becomes an $M \times M$ matrix. In this sense, the matrix-vector multiplication of $A\mathbf{c}$ in (6), where $\mathbf{c} = \{c_j(0)\} \in \mathbb{R}^M$, evolves the coefficients $c_j(0)$ forward in time with lag $\tau$.

With this approximation, the expected error between the discrete estimate $\mathbf{1}_{ij}$ and the true coefficients $A_{ij}$ is zero and the variance is,

$$\mathbb{E}[(\hat{A}_{ij} - A_{ij})^2] \leq b_0^2 \lambda_l^2 \tau N^{-1} + \mathcal{O}(\tau^2 N^{-1}),$$

assuming that $x_i$ are independent. Since $x_i$ form a time series, they are not independent and the convergence of the Monte-Carlo integral will be slower if the dependence is strong. In that case, one may need to subsample the time series which simultaneously requires a larger data set. Assuming independence, by the Chebyshev bound,

$$P(|\hat{A}_{ij} - A_{ij}| \geq \epsilon) \leq b_0^2 \lambda_l^2 \epsilon^{-2} N^{-1} + \mathcal{O}(\tau^2 \epsilon^{-2} N^{-1})$$

and balancing these error terms requires $\lambda_l < b_0^2 \sqrt{\tau}$ and the errors are of order $\epsilon = \mathcal{O}(\tau^{-1/2})$ in probability.

**Non-gradient drift anisotropic diffusion:** We first verify the above approach for a system of SDE’s of the form $\mathbf{1}$ on a torus defined in the intrinsic coordinates $(\theta, \phi) \in [0, 2\pi]^2$ with drift and diffusion coefficients,

$$a(\theta, \phi) = \left(\frac{1}{4} + \frac{1}{8} \cos(\theta) \cos(2\phi) + \frac{1}{16} \cos(\theta + \pi/2)\right),$$

$$b(\theta, \phi) = \left(\frac{1}{4} + \frac{1}{8} \sin(\theta) \cos(\theta + \phi) - \frac{1}{16} \cos(\theta + \phi)\right).$$

This example is chosen to exhibit non-gradient drift, anisotropic diffusion, and multiple time scales. A training data set of 20000 points is generated by numerically solving the SDE in (1) with a discrete time step $\Delta t = 0.1$ and then mapped into the ambient space, $\mathbb{R}^3$, via the standard embedding of the torus given by $(x, y, z) = ((2 + \sin(\theta)) \cos(\phi), (2 + \sin(\theta)) \sin(\phi), \cos(\theta))$. We then define a Gaussian initial density $p_0(\theta, \phi)$ with a randomly selected mean and a diagonal covariance matrix with variance $1/10$. The initial density is projected into a basis of $M = 1000$ eigenfunctions giving coefficients $c_j(0) \equiv \langle p_0/p_{\text{eq}}, \varphi_j \rangle_{p_{\text{eq}}} \approx \sum_{i=1}^{1000} p_0(\theta_i, \phi_i) \varphi_j(\theta_i, \phi_i)/p_{\text{eq}}(\theta_i, \phi_i)$. The coefficients $c_j(0)$ are evolved forward in time in discrete steps of length $\Delta t = 0.1$ by $\hat{A}$, constructed by (1) so that the forecast at time $\tau = n\Delta t$ is effectively $A(\Delta t)^n$.

In Figure 1, we show the evolution of the first two-moments in the ambient space, for the fast and slow variables, $x$ and $z$, respectively, created by the diffusion forecast in (6). To verify the accuracy of the diffusion forecast, we also show the corresponding moments produced by an ensemble forecast of 50000 initial conditions, randomly sampled from the initial distribution $p_0(\theta, \phi)$, evolved using the true dynamical system. Notice the long-time pathwise agreement of both moments constructed via the diffusion forecast and those constructed by an ensemble forecast. See also a video of the evolution of $p$ in the Supplementary Material.
**Lorenz-63 model:** Next, we apply the diffusion forecasting algorithm to data generated by the Lorenz-63 model \( \mathbb{R}^3 \), a deterministic chaotic dynamical system. We will compare our approach to the classical nonparametric prediction method which uses a local linear approximation of the shift map \( \mathbb{R}^3 \) and a standard ensemble forecasting method with the true model, applied with 50000 initial conditions, sampled from the same initial distribution \( p_0 \). In the local linear approach it is known that iterating the 1-shift forecast multiple times could be unstable and typically produces worse forecasts than linearizing a multiple-shift map. In contrast, the diffusion forecast creates a single Markovian approximation to the coarse scale dynamics. Alternatively, when few eigenfunctions are available, it may be advantageous to use multi-shift operators with the diffusion forecast, however this would introduce a bias in the long term forecast.

In this simulation, we generate 10000 data points with discrete time step \( \Delta t = 0.1 \) by solving the Lorenz-63 model. We use the first 5000 data points as training data for both nonparametric models, and the remaining 5000 data points to verify the forecasting skill. For each of the 5000 verification points, we initiate the forecast as a Gaussian \( p_0 \) (evaluated on the training data set) with a variance of 10% of the mean distance to the 15 nearest neighbors. The diffusion forecast is performed with 4500 eigenfunctions \( \varphi_j \), constructed with the diffusion maps algorithm with a variable bandwidth \( \mathbb{R}^3 \) (we show examples in Figure 2). For the local linear forecast, we use an ordinary least squares method that fits an affine model to the \( n \)-step shift map to find the mean forecast. The variance estimate of the local linear model is given by conjugating the covariance matrix of \( p_0 \) (estimated on the training data) with the linear part of the affine \( n \)-step forecast model. To measure the performance, we use the root mean squared error (RMSE) of the mean forecast compared to the true state, averaged over the verification period of 5000 steps. We also show the standard deviation of the forecast density, so that a forecasting method has good uncertainty quantification (UQ) if the standard deviation agrees with the RMSE.

Of course, neither nonparametric model can match the performance of the full model ensemble forecast, however the diffusion forecast is a considerable improvement over the local linear forecast. The diffusion forecast produces better intermediate forecasting skill than the local linear model (see also the 23 step lead time comparison in Figure 2). Also, in the very long term, both the diffusion and ensemble forecasts provide reasonable estimates of the invariant measure, while the local linear forecast has a slight bias. However, for the 1-4 step lead time, the local linear forecast is more accurate than the diffusion forecast due to the reconstruction error in the diffusion model. While the diffusion basis provides good control over the error estimates, other bases, such as the diffusion wavelet basis \( \mathbb{R}^3 \), may provide better localization to improve short-term prediction. In terms of forecast uncertainty, the ensemble forecast provides the most consistent UQ since it has access to the true model. The diffusion forecast produces reasonable estimates without knowing the true model: slight underestimation of forecast error which converges to the true error in the long term. The local linear forecast error estimate is not useful since it significantly underestimates the variance of intermediate to long term forecasts. We include a video comparing the diffusion forecast density to an ensemble in the Supplementary Material.

![Image of Lorenz-63 model](image-url)

**FIG. 2.** Comparison of forecasting methods for Lorenz-63. Bottom: Eigenfunctions \( \varphi_{40}, \varphi_{500}, \varphi_{1500}, \text{and } \varphi_{4000} \) of the coarse approximation of the fractal attractor by a manifold.
El-Niño data set: We now apply our method to a real world data set, namely the Niño-3.4 index, which records the monthly anomalies of sea surface temperature (SST) in the central equatorial Pacific region (the raw dataset is available from NOAA ftp://ftp.cpc.ncep.noaa.gov/wd52dg/data/indices/). In applying this method, we implicitly assume that there is an underlying dynamics on a low-dimensional manifold that generates these SST anomalies. Since the time series is one-dimensional, we apply the time-delay embedding technique to the data set, following [10–12], which will recover this low dimensional manifold if it exists. We use a 5-lag embedding, empirically chosen to maximize the forecast correlation skill between the true time series and the mean estimate. We construct the $A_{ij}$ matrix with 80 eigenfunctions obtained by the diffusion maps algorithm with a variable bandwidth kernel, applied on the lag-embedded data set.

In this experiment, we train our nonparametric model with monthly data between Jan 1950-Dec 1999, following [13] and we verify the forecasting skill on the period of Jan 2000-Sept 2013. The initial distribution $p_0$ is generated with the same method as in the Lorenz-63 example. Based on the RMSE and correlation measure (see Figure 3), the forecasting skill decays to the climatological error in about 6 months but then the skill improves, peaking at 13-14 month lead time. In fact, our 14-month lead forecast skill, in terms of RMSE 0.60 and correlation 0.64, is significantly better than that of the method proposed in [13] (Fig. 3 in their paper suggests RMSE 1.4 and correlation 0.4) who claimed to beat the current operational forecasting skill. The 14-month lead forecast mean estimate gives a reasonably correct pattern except in the period of near January 2004, and the diffusion forecast provides a reasonable error bar (showing one stdev.) which is a useful UQ. We include a movie in the Supplementary Material showing the nontrivial evolution of the forecast distribution starting 14 months before January 2004. Notice that the forecasts never attain the large peaks of the truth, we suspect that this is due to the stochastic nature of the dataset and the limited amount of training data with only 600 data points.

Difficulty in improving the forecasts in this problem may be due to combinations of many factors, including the validity of our assumption of the existence of low-dimensional structures, a transient in the time series, a large stochastic component, and memory effects. Moreover, both the observational and dynamical noise in the data is currently treated as part of the process, however it would be advantageous to isolate the attractor and build a basis for the dynamics on the attractor itself. Finally, the empirical success of this method suggests that it is possible to approximate a fractal attractor with a smooth manifold, however, there is limited theoretical interpretation for such an approximation.

This research is supported under the ONR MURI grant N00014-12-1-0912. D.G. also acknowledges financial support from ONR DRI grant N00014-14-1-0150. J.H. also acknowledges financial support from ONR grant N00014-13-1-0797 and the NSF grant DMS-1317919.

[1] J. Farmer and J. Sidorowich, Phys. Rev. Lett. 59, 845 (1987).
[2] T. Sauer, in Time Series Prediction, Forecasting the Future and Understanding the Past (1994) pp. 175–193.
[3] D. Kugiumtzis, O. C. Lingurde, and N. Christophersen, Physica D: Nonlinear Phenomena 112, 344 (1998).
[4] R. Coifman and S. Lafon, Appl. Comput. Harmon. Anal. 21, 5 (2006).
[5] T. Berry and J. Harlim, Appl. Comput. Harmon. Anal. [in review] (2014), http://arxiv.org/abs/1406.5064.
[6] D. Giannakis, SIAM J. Appl. Dyn. Sys. [in review] (2014), http://arxiv.org/abs/1403.0361.
[7] T. Berry and J. Harlim, SIAM/ASA J. Uncer. Quant. [in review] (2014), http://arxiv.org/abs/1407.6072.
[8] E. N. Lorenz, J. Atmos. Sci. 20, 130 (1963).
[9] R. Coifman and M. Maggioni, Appl. Comp. Harmonic Anal. 21, 53 (2006).
[10] T. Sauer, J. Yorke, and M. Casdagli, J. Stat. Phys. 65, 579 (1991).
[11] D. Giannakis and A. J. Majda, Proc. Nat. Acad. Sci. 109, 2222 (2012).
[12] T. Berry, J. R. Cressman, Z. G. Ferencöö, and T. Sauer, SIAM J. Appl. Dyn. Syst. 12, 618 (2013).
[13] M. D. Chekroun, D. Kondrashov, and M. Ghil, Proc. Nat. Acad. Sci. 108, 11766 (2011).