

Data-driven forecasting without a model and with a partially known model

Tyrus Berry

The Pennsylvania State University

March 13, 2015

Collaborators: John Harlim (PSU) and Dimitrios Giannakis (Courant)
Support: Office of Naval Research

Nonparametric/Data Driven/Model Free

What is Nonparametric Modeling?

Nonparametric Forecasting

The probabilistic perspective

The geometric prior

Example: nonparametric forecast on a torus

Applications of nonparametric forecast

Semiparametric Forecasting

Model Error and the Curse of Dimensionality

Semiparametric modeling

Example: Lorenz-96 with model error

Nonparametric Model: Definition and Intuition

- ▶ Definition
 - ▶ Parameter set is infinite
 - ▶ Only a finite subset of parameters are used
 - ▶ Data set determines which parameters are used
 - ▶ As data is added the parameter set evolves

- ▶ Intuition: Gaussian Parameterization vs. Histogram.
 - ▶ Gaussian Parameters: μ, σ
 - ▶ Histogram Parameters: $\{f([a_i, a_{i+1}]) : i \in \mathbb{Z}\}$

Goal of nonparametric forecasting

(Loading Video...)

Forecasting observables with the Generator

- ▶ Consider a dynamical system $dx = a(x) dt + b(x) dW_t$
- ▶ For an observable $f(x(t))$, by the Itô formula we have:

$$\begin{aligned} df &= \left(\nabla f^\top a + \frac{1}{2} \sum_{i,j,k} b_{ik} b_{jk} \frac{\partial^2 f}{\partial x_i \partial x_j} \right) dt + \nabla f^\top b dW_t \\ &= \mathcal{L}f dt + \nabla f^\top b dW_t \end{aligned}$$

- ▶ The linear operator \mathcal{L} is the *generator* of the system
- ▶ $\mathbb{E}[df] = \mathbb{E}[\mathcal{L}f] dt$
- ▶ Feynman-Kac formula, $\mathbb{E}[f(x(t))] = e^{t\mathcal{L}}f(x(0))$

Forecasting full densities with the Fokker-Plank PDE

- ▶ Consider an uncertain initial state $x(0)$ with density $p(x, 0)$

$$\begin{aligned}\left\langle f, \frac{\partial}{\partial t} p(\cdot, t) \right\rangle &= \frac{d}{dt} \langle f, p(\cdot, t) \rangle = \frac{d}{dt} \mathbb{E}[f(x(t))] \\ &= \langle \mathcal{L}f, p(x, t) \rangle = \langle f, \mathcal{L}^* p(x, t) \rangle\end{aligned}$$

- ▶ *Fokker-Planck* operator \mathcal{L}^* is the adjoint of the generator \mathcal{L}
- ▶ Density solves Fokker-Planck PDE, $p_t = \mathcal{L}^* p$
- ▶ Semigroup solution, $p(x, t) = e^{t\mathcal{L}^*} p(x, 0)$
- ▶ Invariant measure satisfies $\mathcal{L}^* p_{\text{eq}} = 0$

If we had a model...

- ▶ If we knew the model $dx = a(x) dt + b(x) dW_t$
- ▶ We could construct the Fokker-Planck operator \mathcal{L}^*
- ▶ Then solve the F-P PDE $p_t = \mathcal{L}^* p$ with initial condition $p(x, 0)$
- ▶ This is hard, even when everything is known
- ▶ The data is a solution of the true system
- ▶ Given data we can bypass the model entirely

The Shift Map

- ▶ Given data samples $x_i = x(t_i)$ with $\tau = t_{i+1} - t_i$
- ▶ Define the *shift map* of a function by $Sf(x_i) = f(x_{i+1})$
- ▶ Using the Itô lemma we can show:

$$Sf(x_i) = e^{\tau \mathcal{L}} f(x_i) + \int_{t_i}^{t_{i+1}} \nabla f^\top b \, dW_s + \int_{t_i}^{t_{i+1}} Bf \, ds$$

- ▶ Where $B = \mathcal{L}f - \mathbb{E}[\mathcal{L}f]$ so that $\mathbb{E}[Bf] = 0$
- ▶ Unbiased estimate $\mathbb{E}[Sf(x_i)] = e^{\tau \mathcal{L}} f(x_i)$
- ▶ Need to minimize the stochastic integrand $\nabla f^\top b$

Representing the Shift Map

- ▶ Choose a basis $\{\varphi_j\}$ orthonormal with respect to $\langle \cdot, \cdot \rangle_{p_{\text{eq}}}$
- ▶ The coefficients $c_l(t) = \langle p(x, t), \varphi_l \rangle$ have evolution:

$$\begin{aligned} c_l(t + \tau) &= \langle p(x, t + \tau), \varphi_l \rangle = \left\langle e^{\tau \mathcal{L}^*} p(x, t), \varphi_l \right\rangle = \langle p(x, t), e^{\tau \mathcal{L}} \varphi_l \rangle \\ &= \left\langle \sum_j c_j(t) \varphi_j p_{\text{eq}}, e^{\tau \mathcal{L}} \varphi_l \right\rangle = \sum_j c_j(t) \langle \varphi_j, e^{\tau \mathcal{L}} \varphi_l \rangle_{p_{\text{eq}}} \\ &= \sum_j A_{lj} c_j(t) \end{aligned}$$

- ▶ So $\vec{c}(t + \tau) = A \vec{c}(t)$
- ▶ Where $A_{lj} = \langle \varphi_j, e^{\tau \mathcal{L}} \varphi_l \rangle_{p_{\text{eq}}} \approx \frac{1}{N} \sum_{i=1}^N \varphi_j(x_i) \varphi_l(x_{i+1})$

Forecasting with the Shift Map

$$\begin{array}{ccc}
 p(x, t) & \xrightarrow{\text{Nonparametric Forecast}} & p(x, t + \tau) \\
 \downarrow \langle p, \varphi_j \rangle & & \uparrow \sum_j c_j \varphi_j p_{\text{eq}} \\
 \vec{c}(t) & \xrightarrow{A_{lj} \equiv \mathbb{E}[\langle \varphi_j, S \varphi_l \rangle]_{p_{\text{eq}}}} & \vec{c}(t + \tau) = A \vec{c}(t).
 \end{array}$$

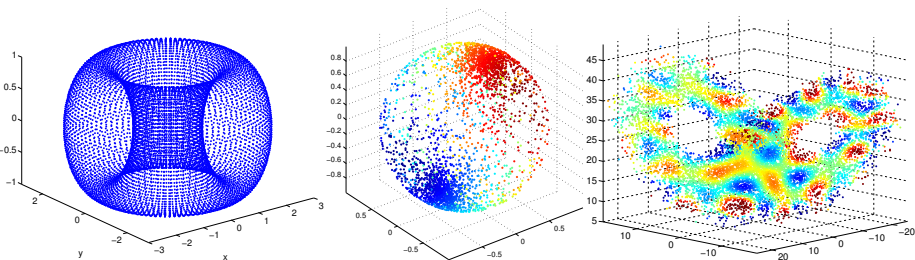
- ▶ We approximate $c_l(t) \approx \frac{1}{N} \sum_{i=1}^N \varphi_l(x_i) p(x_i, t) / p_{\text{eq}}(x_i)$
- ▶ We approximate A_{lj} with $\hat{A}_{lj} = \frac{1}{N} \sum_{i=1}^N \varphi_j(x_i) \varphi_l(x_{i+1})$
- ▶ $\mathbb{E}[\hat{A}_{lj}] = A_{lj}$ with error $\mathcal{O}(\|\nabla \varphi_l\|_{p_{\text{eq}}} \sqrt{\tau/N})$

Choosing a basis

- ▶ Need to minimize the error term $\mathcal{O}(\|\nabla\varphi_I\|_{\rho_{\text{eq}}}\sqrt{\tau/N})$
- ▶ The minimizers of $\|\nabla\varphi_I\|_{\rho_{\text{eq}}}$ are a generalized Fourier basis
- ▶ Let $\Delta_{\rho_{\text{eq}}} = \Delta + \frac{\nabla\rho_{\text{eq}}}{\rho_{\text{eq}}} \cdot \nabla$ be the Laplacian weighted by ρ_{eq}
- ▶ The eigenfunctions $\Delta_{\rho_{\text{eq}}}\varphi_j = \lambda_j\varphi_j$ minimize $\|\nabla\varphi_j\|_{\rho_{\text{eq}}} = \lambda_j$
- ▶ If the data fills the ambient space then $\Delta_{\rho_{\text{eq}}}$ is known, but it is unlikely that we will have enough data
- ▶ What if the data does not fill the ambient space?
- ▶ What is $\Delta_{\rho_{\text{eq}}}$ and how do we find φ_j ?

Defining the Geometric Prior

- ▶ Data does not actually fill the high-dimensional data space \mathbb{R}^n



- ▶ Assume data lies on a manifold \mathcal{M} embedded in \mathbb{R}^n
- ▶ Δ is the Laplace-Beltrami operator and ∇ is the gradient operator

Accessing the Geometric Prior

- ▶ Assume data set $\{x_i\}_{i=1}^N \subset \mathbb{R}^n$ is sampled from $x_i \in \mathcal{M} \subset \mathbb{R}^n$
- ▶ The sampling density $x_i \sim \pi(x)$ is singular as a density on \mathbb{R}^n
- ▶ Consider $x_i \sim q(x)$, where q is the sampling density taken with respect to the volume form on \mathcal{M} so $\pi d\mu = q dV$
- ▶ Averaging a function $f : \mathcal{M} \rightarrow \mathbb{R}^n$ is a Monte-Carlo integral biased by q :

$$\frac{1}{N} \sum_{i=1}^N f(x_i) = \int_{\mathcal{M}} f(x) q(x) dV(x) + \mathcal{O}(N^{-1/2})$$

- ▶ We will assume ergodicity so that $q = p_{\text{eq}}$

Accessing the Geometric Prior

- ▶ If $0 \leq K_\epsilon(x, y) \leq ce^{-\sigma\|x-y\|^2/\epsilon}$ then

$$\tilde{q}(x) \equiv \frac{\epsilon^{-d/2}}{N} \sum_{i=1}^N K_\epsilon(x, x_i) \approx \int_{\mathcal{M}} \frac{K_\epsilon(x, y)}{\epsilon^{d/2}} q(y) dV(y) = m_0 q(x) + \mathcal{O}(\epsilon)$$

- ▶ Continuing the expansion, we find,

$$\frac{1}{N} \sum_{i=1}^N K_\epsilon(x, x_i) \frac{f(x_i)}{\tilde{q}(x_i)} = f(x) + \epsilon m(\omega(x)f(x) + \Delta f(x)) + \mathcal{O}(\epsilon^2)$$

- ▶ With an appropriate normalization we can extract $L \approx \Delta$
- ▶ Can also approximate other important operators such as $\Delta_{p_{\text{eq}}}$

Representing the Geometric Prior

- ▶ Represent functions $f : \mathcal{M} \rightarrow \mathbb{R}$ by vectors $(\vec{f})_i = f(x_i)$
- ▶ Dot products are Monte-Carlo integrals:

$$\frac{1}{N} \vec{f} \cdot \vec{g} \approx \int_{\mathcal{M}} f(x)g(x)q(x) dV(x) = \langle f, g \rangle_{L^2(\mathcal{M},q)}$$

- ▶ Represent operators $f \mapsto \mathcal{L}f$ by matrices $(L\vec{f})_i = (\mathcal{L}f)(x_i)$
- ▶ Applying an appropriate normalization to the kernel matrix $K_{ij} = K(x_i, x_j)$ we can approximate $L_{p_{\text{eq}}} \approx \Delta_{p_{\text{eq}}}$
- ▶ The eigenvectors of $L_{p_{\text{eq}}}$ approximate the eigenfunctions φ_j which minimize $\|\nabla\varphi_j\|_{p_{\text{eq}}}$

Nonparametric forecast on a torus

- ▶ Stochastic dynamics on a torus $(\theta, \phi) \in [0, 2\pi)^2$

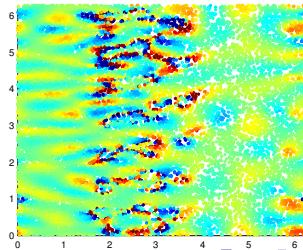
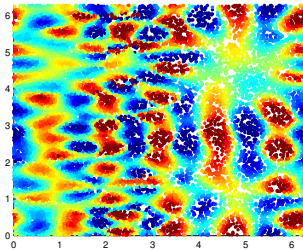
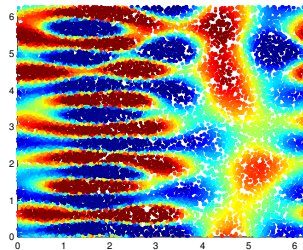
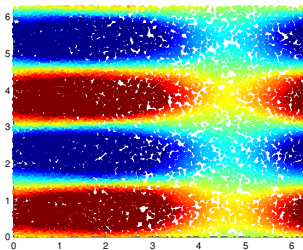
$$d(\theta, \phi)^\top = a(\theta, \phi) dt + b(\theta, \phi) dW_t$$

- ▶ Drift and diffusion coefficients,

$$a(\theta, \phi) = \begin{pmatrix} \frac{1}{2} + \frac{1}{8} \cos(\theta) \cos(2\phi) + \frac{1}{2} \cos(\theta + \pi/2) \\ 10 + \frac{1}{2} \cos(\theta + \phi/2) + \cos(\theta + \pi/2) \end{pmatrix},$$

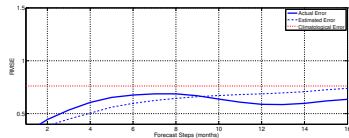
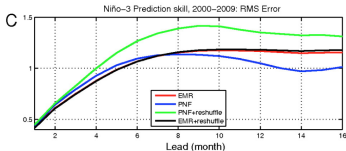
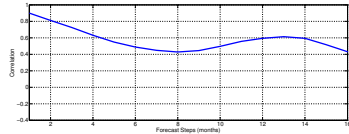
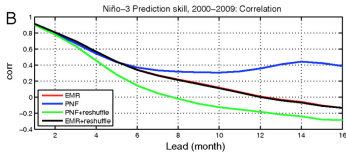
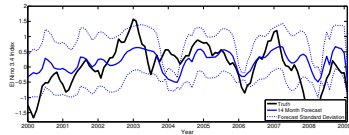
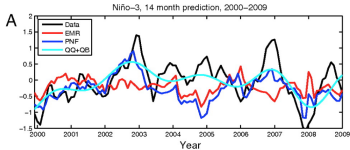
$$b(\theta, \phi) = \begin{pmatrix} \frac{1}{4} + \frac{1}{4} \sin(\theta) & \frac{1}{4} \cos(\theta + \phi) \\ \frac{1}{4} \cos(\theta + \phi) & \frac{1}{40} + \frac{1}{40} \sin(\phi) \cos(\theta) \end{pmatrix}.$$

Eigenfunctions on the torus



(Loading Video...)

Forecasting the El Niño Index

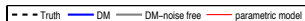
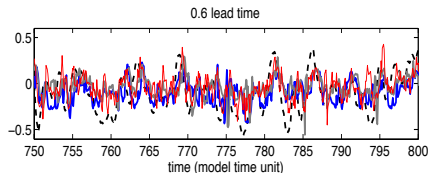
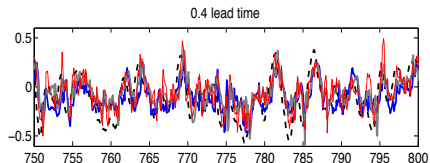
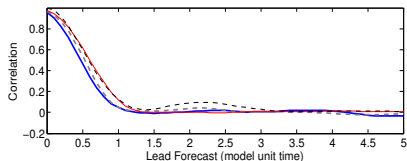
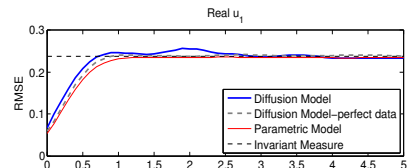


Diffusion Forecast

Chekrouna, Kondrashov, and Ghil, PNAS 2011,108,no.29

Forecasting a Fourier mode of TBH

Diffusion Forecast compared to parametric model of A.Majda,J.Harlim,
Physics constrained nonlinear regression models for time series, *Nonlinearity* 26 (2013).

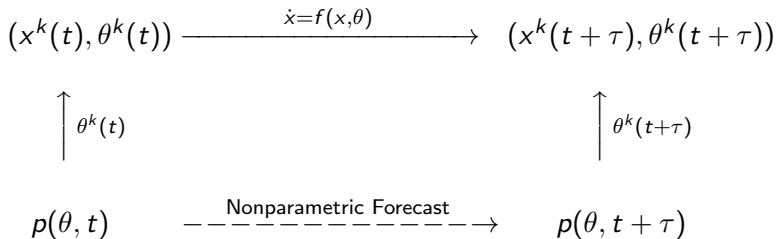


Model Error and the Curse of Dimensionality

- ▶ Nonparametric model *interpolates* from the training data
- ▶ Data required grows exponentially in the dimension of the manifold
- ▶ For high-dimensional systems we usually know something
- ▶ High-dimensional models are subject to model error
- ▶ Idea: Use the nonparametric model for the model error

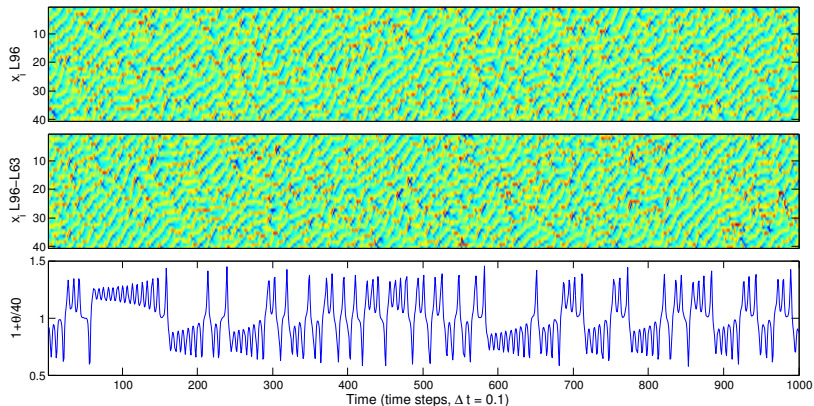
Semiparametric Model

- ▶ Assume we have a partially known model $\dot{x} = f(x, \theta)$
- ▶ Dynamics $d\theta = a(\theta) dt + b(\theta) dW_t$ are unknown
- ▶ Build a nonparametric model for $p(\theta, t)$
- ▶ Sample $\theta^k(t) \sim p(\theta, t)$ to use with ensemble forecast (x^k, θ^k)



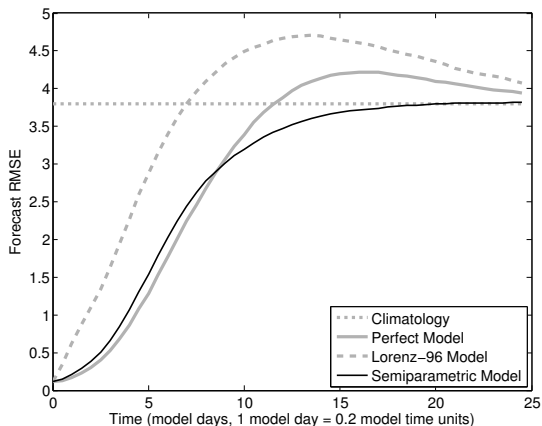
Example: 40-dimensional Lorenz-96 system

$$\dot{x}_i = f(x_i, \theta) = \theta x_{i-1} x_{i+1} - x_{i-1} x_{i-2} - x_i + 8$$



Example: 40-dimensional Lorenz-96 system

$$\dot{x}_i = f(x_i, \theta) = \theta x_{i-1} x_{i+1} - x_{i-1} x_{i-2} - x_i + 8$$



Additional challenges of semiparametric modeling

- ▶ Need a training data set for θ
- ▶ Need initial condition $p(\theta, t)$ for nonparametric forecast
- ▶ We developed semiparametric filtering to address these
- ▶ Still require that evolution of θ does not depend on x

For more information: <http://personal.psu.edu/thb11/>

Building the basis

- ▶ Coifman and Lafon, *Diffusion maps*.
- ▶ Berry and Harlim, *Variable Bandwidth Diffusion Kernels*.
- ▶ Berry and Sauer, *Local Kernels and the Geometric Structure of Data*.

Nonparametric forecast

- ▶ Berry, Giannakis, and Harlim, *Nonparametric forecasting of low-dimensional dynamical systems*.
- ▶ Berry and Harlim, *Forecasting Turbulent Modes with Nonparametric Diffusion Models*.

Semiparametric forecast

- ▶ Berry and Harlim, *Semiparametric forecasting and filtering: correcting low-dimensional model error in parametric models*.