Sparse Regression and Adaptive Feature Generation for the Discovery of Dynamical Systems

Chinmay Kulkarni, Abhinav Gupta and Pierre F. J. Lermusiaux

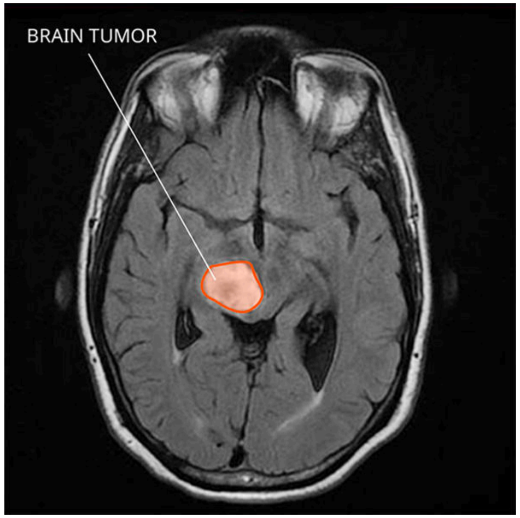
Department of Mechanical Engineering, MIT

Dynamic Data Driven Applications Systems (DDDAS), 2020

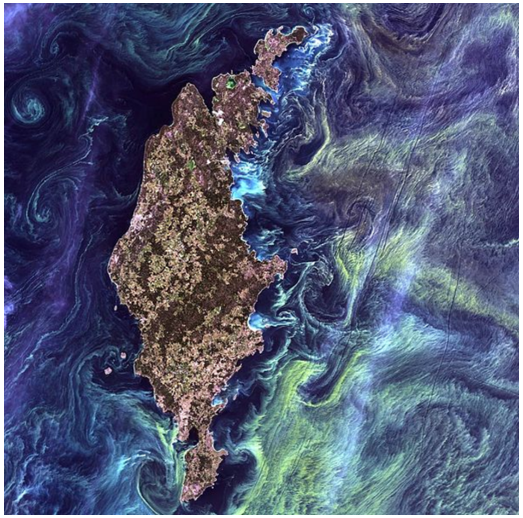
Introduction and Motivation

Data-driven modeling:

- Commonly used for a variety of research and societal needs
- Energy, food, sustainability, security, medical applications etc.



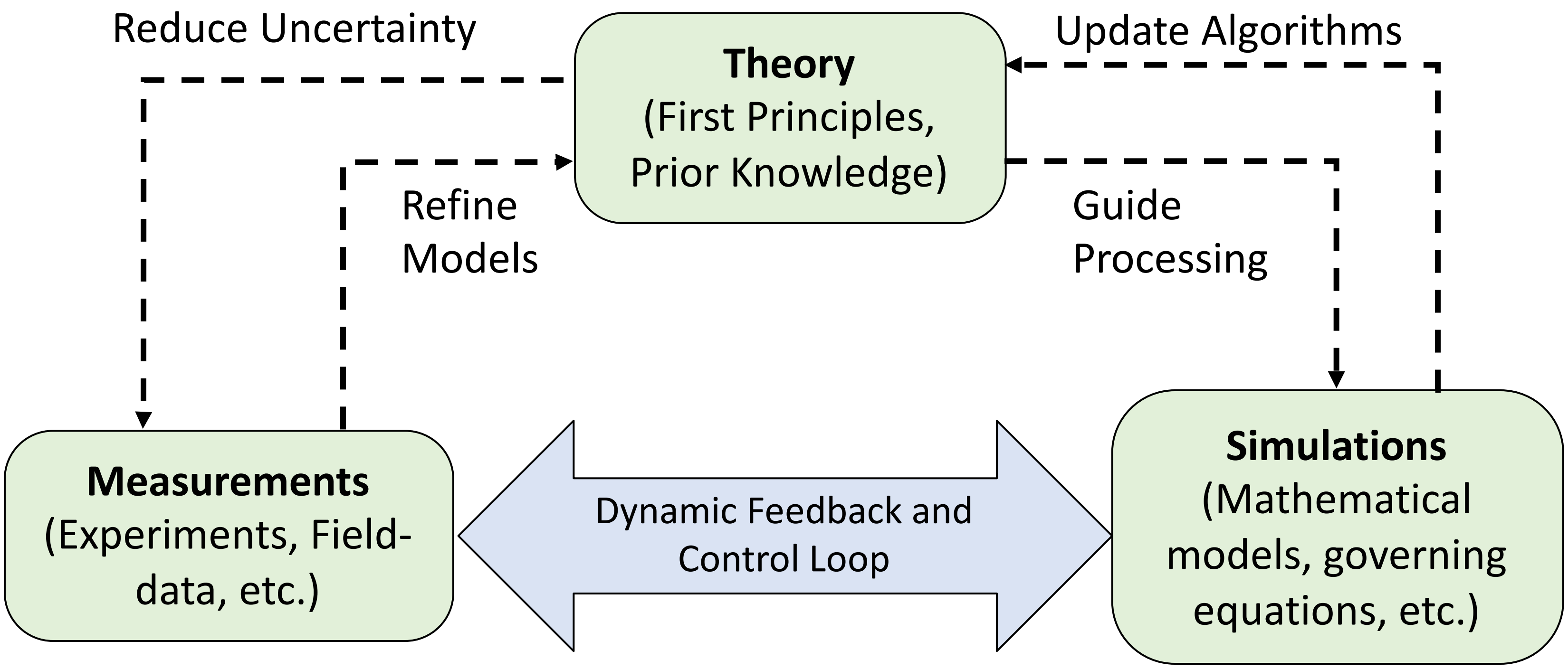
Diffuse Midline Gliomas (National Cancer Institute)



Satellite image of phytoplankton in the Baltic Sea around Gotland (USGS)

Question: Is the role of data limited to the verification of first-principles or finding empirical relations, or can be used to discover the underlying governing model?

Dynamic Data-Driven Application Systems Paradigm



Bayesian Learning and Deep Learning of Dynamical Models

Learning with Prior

Use data and uncertain prior knowledge, to evolve the pdf of model equations, states, parameters, etc.

Learning without Prior

Use only data and no prior knowledge to estimate the model equations, states, parameters, etc.

Dynamic Bayesian Learning:

Estimate the pdf of model equations (while learning states, parameters, etc.)

- GMM-DO: [Lu, SM-MIT '13; Lu and Lermusiaux, 2016; Lin and Lermusiaux, 2020; Gupta and Lermusiaux, in prep]
- ESSE: [Lermusiaux et al, 2004, 2007]

Deep Learning:

Predict future states without finding explicit representation of model equations

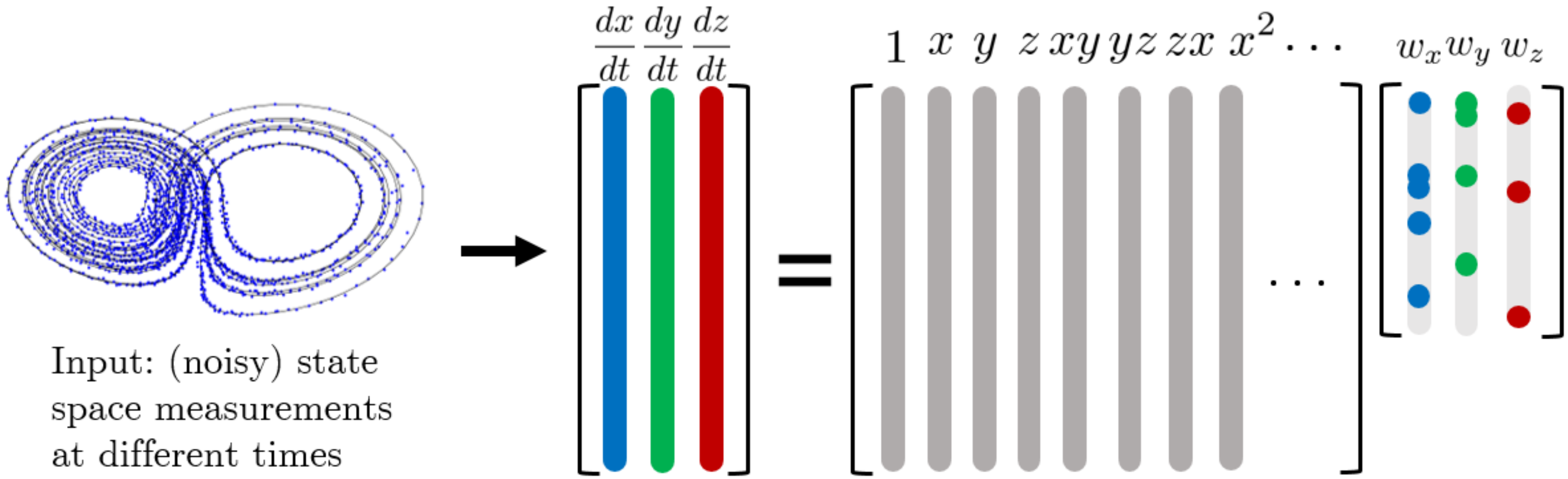
- ODEs: [Ogunmolu et al., '16; Trischer et al, '16; Yeo, '16]
- PDEs: [Kulkarni et al., 2020, in prep.]

Sparse Regression:

Learn functional form of model equations only from data

- [Brunton et al., '16; Schaeffer et al., '17; Rudy et al., '17]
- Adaptive & Dynamic: [Kulkarni et al., 2020, subm.]

Discovering the Governing Model Dynamics



Sparse regression using a library of nonlinear features

Thresholding to obtain component-wise optimal feature space

$$\frac{dx}{dt} = \sigma(y - x) \qquad \frac{dy}{dt} = x(\rho - z) - y \qquad \frac{dz}{dt} = xy - \beta z$$

Given: State measurements, state rate of change measurements at discrete times

Construct: Feature library – typically using polynomials etc.

Solve: Sparse regression problem to identify the active features in the feature library (typically sparse, as functional form only contains a few terms on the RHS)

Obtain: Equations in symbolic form by identifying the active feature in the library

Brunton et al. ‘16, Schaeffer et al. ‘17, Rudy et al. ‘17

Discovering the Governing Model Dynamics

State-of-the-Art:

- L_1 regularized regression (LASSO) for promoting sparsity
- Fixed feature space – typically polynomials
- Common hyperparameter: weight penalty

Issues:

- LASSO does not yield unique, sparse, and robust coefficient vector
- The actual functional representation may not be contained in the library
- Scaling of different states not accounted for

Solutions:

- Dual LASSO for feature selection – robust model selection even from highly correlated features
- Adaptive feature growth, scale-based thresholding – grow the feature library

Brunton et al. '16, Schaeffer et al. '17, Rudy et al. '17

Mathematical Setup, Notation

States: $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$ at discrete times $t_1, t_2, \ldots, t_k$

Rates of Change: $\dot{X} = [\dot{x}_1, \dot{x}_2, \ldots, \dot{x}_n]$ at the same time instants $t_1, t_2, \ldots, t_k$

Feature Library: $X(\boldsymbol{x}, t)$ contains polynomials of states up to maximum degree p

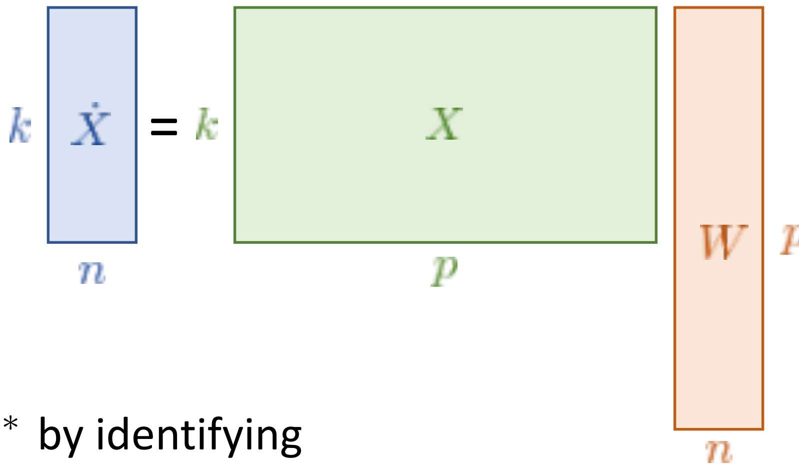
Total number of terms $m = \dfrac{(n+p)!}{n!p!} >> n$

Solve optimization: $W^* = \arg\min_{W} \mathcal{L}(W) = \arg\min_{W} \left[\left(\dot{X} - XW \right)^2 + \mathcal{P}(W) \right].$

Sparsity: Only $\mathcal{O}(n)$ present (active) polynomials in the feature library – impose sparsity

Ideal sparsity is L_0 $\mathcal{P}(W) = \lambda |W|_0$

Convex counterpart: L_1 $\mathcal{P}(W) = \lambda |W|_1$



Obtain: Equations of the form $\dot{X} = f(\boldsymbol{x}, t) = XW^*$ by identifying the active components in the feature library

Analysis and Pitfalls of LASSO

$$W^* = \arg\min_{W} \left[\left(\dot{X} - XW \right)^2 + \lambda |W|_1 \right] .$$

- Using the Rademacher averages and the symmetrization lemma, we get:

$$\mathbb{E} \max_{g \in \mathcal{G}} \left[\mathbb{E} g(x_i) - \frac{1}{n} \sum_{i=1}^{n} g(x_i) \right] \approx \mathcal{O}\left(\sqrt{\frac{\log(m)}{n}} \right)$$

As $m >> n$, this bound is impractical

- Another important task is to choose the penalty weight λ
  - Higher the correlation amongst the features in $X(\boldsymbol{x}, t)$, lower the value of λ
  - Analytical suggestion: $\lambda = \mathcal{O}\left(\sqrt{k \log(m)} \right)$, significant tuning required

- Main issues: LASSO fit (i.e. XW^*) is unique, but the weight vector W^* is not
- LASSO tends to choose a feature at random amongst the correlated features

Kakade et al. '09; Bickel et al. '09; Ghaoui et al. '10

Dual LASSO for Feature Selection

- LASSO is convex $\Longrightarrow$ solve the dual optimization of LASSO instead of the primal

- Dual LASSO: $\theta^* = \arg\max_{\theta} \mathcal{D}(\theta) = ||\dot{X}||_2^2 - ||\theta - \dot{X}||_2^2$ such that $||X^T\theta||_\infty \leq \lambda$

- Stationarity condition: $\theta^* = \dot{X} - XW^*$

  Unique for dual LASSO

  Unique for LASSO

- KKT conditions: $(\theta^*)^T X_i \begin{cases} = \text{sign}(W_i^*) \text{ if } W_i^* \neq 0 \\ \in (-1,1) \text{ if } W_i^* = 0 \end{cases}$ Choose dual active features using this

- Strong duality $\Longrightarrow$ optimal primal and dual active features are the same! (with h.p.)

- Dual LASSO tells us the correct active features robustly, but does not yield a good fit for their coefficient values

- Once the active features are determined using dual LASSO, perform ridge (L_2) regression to determine the coefficient values

Dual LASSO for Feature Selection

- Dual LASSO: $\theta^* = \arg\max_\theta \mathcal{D}(\theta) = ||\dot{X}||_2^2 - ||\theta - \dot{X}||_2^2 \;\text{ such that }\; ||X^T\theta||_\infty \leq \lambda$
- Stationarity condition: $\theta^* = \dot{X} - XW^*$
- KKT conditions: $(\theta^*)^T X_i \begin{cases} = \text{sign}(\hat{W}_i) \text{ if } \hat{W}_i \neq 0 \\ \in (-1, 1) \text{ if } \hat{W}_i = 0 \end{cases}$

Algorithm 1 Sparse regression using dual LASSO feature selection

---

Require: state parameters: $\boldsymbol{x} = x_i^t, \dot{X} = \dot{x}_i^t$; LASSO penalty λ, ridge penalty λ_2

Construct the feature library appropriately

Compute the primal LASSO solution: $W^* = \min_W \left[\left(\dot{X} - XW\right)^2 + \lambda|W|_1\right]$

Compute the unique dual solution $\theta^* = \dot{X} - XW^*$

Compute dual active set (same as primal active with h.p.) $S^d = \{1 \leq j \leq m : (\theta^*)^T X_j \notin (-1, 1)\}$

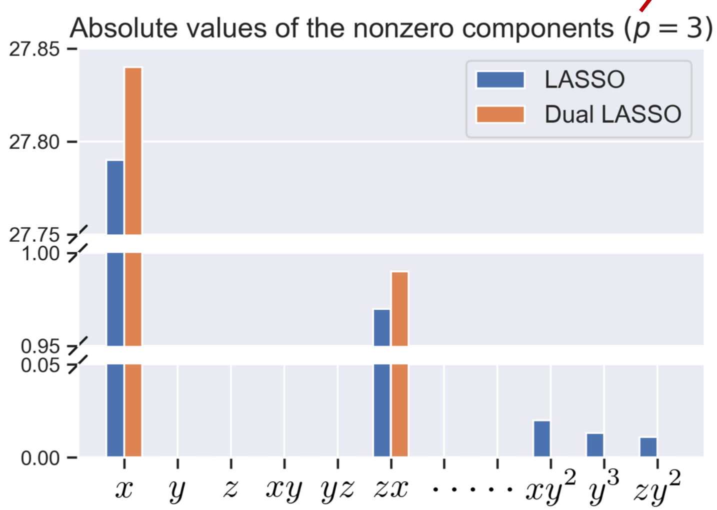
Construct reduced feature matrix $\tilde{X}$ by only considering features whose indices are in S^d

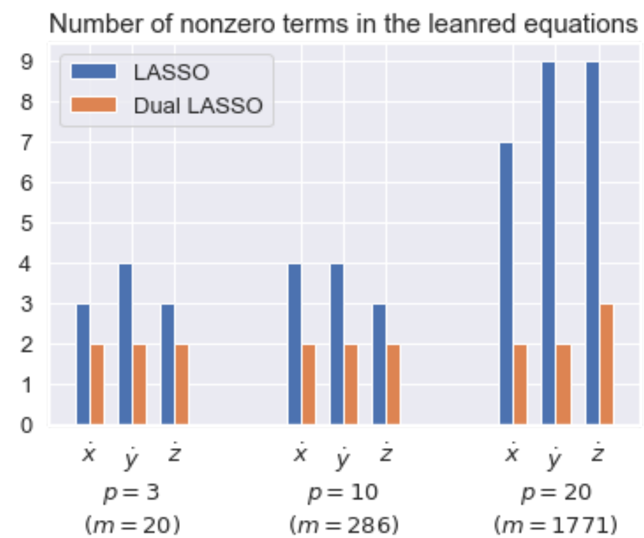
Solve ridge regression $W^* = \arg\min_W \left[\left(\dot{X} - XW\right)^2 + \lambda_2|W|_2^2\right]$

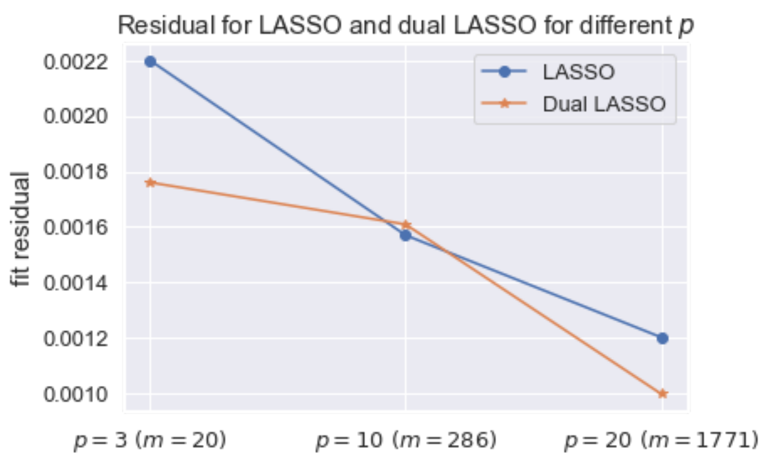
Results: Lorentz 63 Attractor

Actual system: $\dot{x} = 10(yz - x)\ ;\quad \dot{y} = x(28 - z)\ ;\quad \dot{z} = xy - 2.667z$

Learned system:



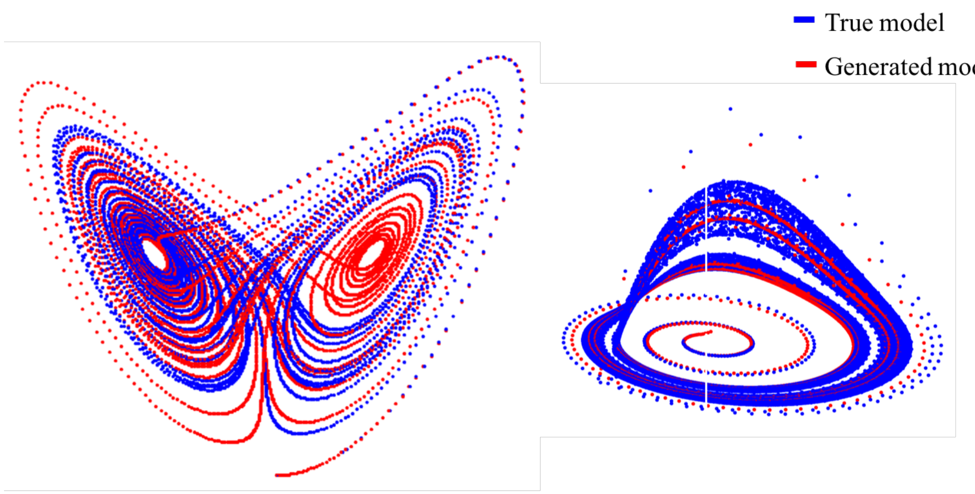




Results: Other Dynamical Systems

	True system	Clean Data	Noisy Data ($\sigma = 0.1$)
ODEs			
Lorenz attractor	$\dot{x} = 10(y - x)$ $\dot{y} = x(28 - z) - y$ $\dot{z} = xy - \frac{8}{3}z$	$\dot{x} = 10y - 10x$ $\dot{y} = x(27.9941 - 0.9998z) - 0.9985y$ $\dot{z} = xy - 2.6667z$	$\dot{x} = 9.98537y - 9.6639x$ $\dot{y} = x(27.6240 - 0.8926z) - 0.9890y$ $\dot{z} = 0.9861xy - 2.7170z$
Rossler attractor	$\dot{x} = -y - z$ $\dot{y} = x + 0.2y$ $\dot{z} = 0.2 + z(x - 5.7)$	$\dot{x} = -1.000y - 1.000z$ $\dot{y} = 1.000x + 0.2y$ $\dot{z} = 0.2 + z(x - 5.700)$	$\dot{x} = -0.9960y - 0.9969z$ $\dot{y} = 1.0006x + 0.2036y$ $\dot{z} = 0.1761 + z(1.0271x - 5.66876)$
Hyperchaotic attractor	$\dot{x} = -y - z$ $\dot{y} = x + 0.25y + w$ $\dot{z} = 3 + xz$ $\dot{w} = -0.5z + 0.05w$	$\dot{x} = -1.000y - 1.000z$ $\dot{y} = 1.000x + 0.250y + 1.000w$ $\dot{z} = 3.000 + 1.000xz$ $\dot{w} = -0.500z + 0.050w$	$\dot{x} = -0.9999y - 1.0005z$ $\dot{y} = 0.9998x + 0.2479y + 0.9933w$ $\dot{z} = 2.5034 + 0.9959xz$ $\dot{w} = -0.4987z + 0.0503w$
PDEs			
Sommerfeld equation	$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0$	$\frac{\partial u}{\partial t} + 0.9976\frac{\partial u}{\partial x} = 0$	$\frac{\partial u}{\partial t} + 1.0942\frac{\partial u}{\partial x} = 0$
Burgers' equation	$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0$	$\frac{\partial u}{\partial t} + 0.9953u\frac{\partial u}{\partial x} = 0$	$\frac{\partial u}{\partial t} + 0.9949u\frac{\partial u}{\partial x} = 0$

Figure: Showcasing generative capabilities of learned model



Lorenz attractor

Rossler attractor

Adaptive Feature Library

- Symbolic equation discovery is inaccurate if the true functional form is not contained in the span of the considered function library

- Adding and removing features in a greedy manner is not optimal

Solution:

- Use an functional library consisting of orthogonal functions. Grow / shrink the library adaptively. As functions are orthogonal, once a component is dropped, it should never reappear

- Perform sequentially thresholded ridge regression[1], as the formulation may not be sparse in this new (orthogonal basis)

- Perform symbolic simplification[2] to obtain the final governing equations in functional form

1 Rudy et al. '17; Bailey et al. '14

Discovering the Governing Model Dynamics

Algorithm 2 Learning the Governing Equations through Adaptive Growth of the Feature Library

Require: state parameters: $\boldsymbol{x} = x_i^t, \dot{X} = \dot{x}_i^t$; orthogonal family $F_j(\bullet)$;
feature addition / removal thresholds: $r_a\ (\leq 1), r_r\ (\geq 1), \lambda_0$; removal step frequency k_r
Initialize: $X = \emptyset, W = \mathbf{0}, t = 0, \mathcal{L} = \infty$
while True **do**
  $X_t = \text{append}(X, F_k(\boldsymbol{x}))$
  Solve the STRidge problem: $W_t = \text{STRidge}(\dot{X}, X_t, \lambda_0)$
  Compute the loss $\mathcal{L}_t = \left(\dot{X} - X_t W_t\right)^2$
  if $\mathcal{L}_t \leq r_a \mathcal{L}$ **then**
    $X = X_t\ ;\ W = W_t$
  if $\mod (k, k_r) == 0$ **then**
    for $i = 1, \ldots, X.\text{shape}[1]$ **do** (number of columns of X)
      $X_t = \text{append}\left(X[:, 1 : i-1], X[:, i+1 : \text{end}]\right)$ (ignore the i^{th} column of X)
      Solve the STRidge problem: $W_t = \text{STRidge}(\dot{X}, X_t, \lambda_0)$
      Compute the loss $\mathcal{L}_t = \left(\dot{X} - X_t W_t\right)^2$
      if $\mathcal{L}_t \leq r_r \mathcal{L}$ **then**
        $X = X_t\ ;\ W = W_t$
  $k = k + 1.$
  break if no change in feature space over multiple iterations.
Perform symbolic simplification of $\dot{X} = XW$ to obtain the final form of the equations

Add the feature to the library if the loss decreases substantially

Remove the feature from the library if the loss does not increase much

Results: Quadratic Lorenz Attractor

Actual system: $\dot{x} = 10(yz - x)\ ;\qquad \dot{y} = x(28 - z)\ ;\qquad \dot{z} = (xy)^2 - 2.667z$

Learned system:

$$\left.\begin{aligned}
\dot{x} &= 9.93\mathbb{L}_1(y)\mathbb{L}_1(z) - 9.89\mathbb{L}_1(x)\\
\dot{y} &= 27.66\mathbb{L}_1(x) - 1.04\mathbb{L}_1(x)\mathbb{L}_1(z)\\
\dot{z} &= 0.43\mathbb{L}_2(x)\mathbb{L}_2(y) + 0.22\mathbb{L}_2(x) + 0.21\mathbb{L}_2(y)\\
&\quad -2.62\mathbb{L}_1(z) + 2.09\mathbb{L}_0(x) - 0.22\mathbb{L}_0(y) - 1.95\mathbb{L}_0(z)
\end{aligned}\right\}$$

After orthogonal feature regression

$$\underbrace{\begin{aligned}
\dot{x} &= 9.93yz - 9.89x\\
\dot{y} &= 27.66x - 1.04xz\\
\dot{z} &= 0.97(xy)^2 + 0.007(x^2 - y^2)\\
&\quad - 2.62z + 0.027
\end{aligned}}_{\text{after symbolic simplification}} \implies \underbrace{\begin{aligned}
\dot{x} &= 9.93yz - 9.89x\\
\dot{y} &= 27.66x - 1.04xz\\
\dot{z} &= 0.9675(xy)^2 - 2.62z
\end{aligned}}_{\text{after scale based thresholding}}$$

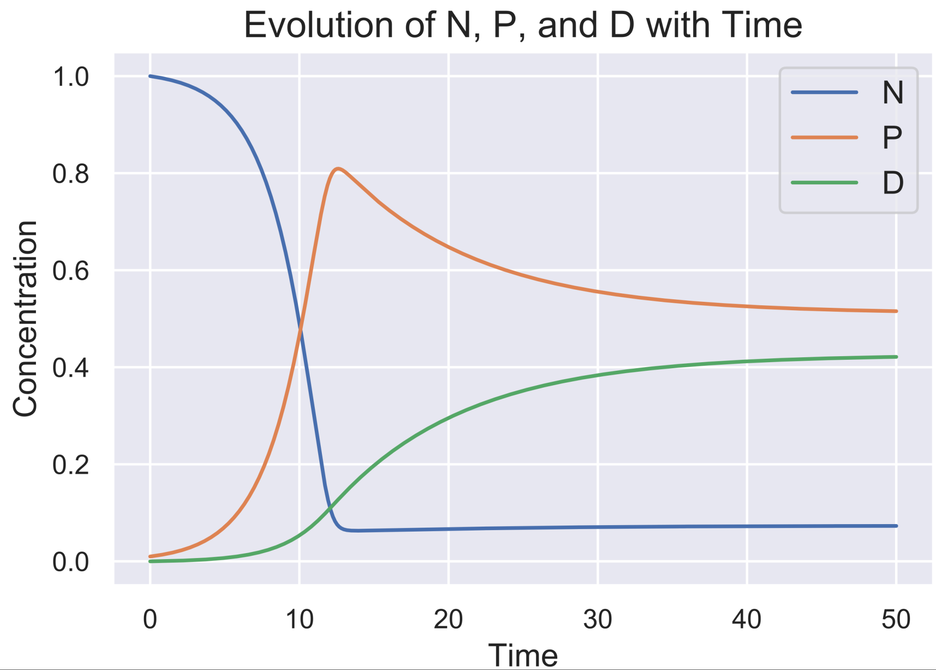
Results: Marine Ecosystem Model

3-Component Nutrient-Phytoplankton-Detritus (NPD) Model:

$$\frac{dN}{dt} = -\underbrace{\boxed{\frac{NP}{0.3+N}}}_{\text{N uptake by P}} + 0.5P + 0.06D$$

$$\frac{dP}{dt} = \frac{NP}{0.3+N} - \underbrace{\boxed{0.5P}}_{\text{P loss due to Respiration}} - \underbrace{\boxed{0.05P}}_{\text{P loss due to Mortality}}$$

$$\frac{dD}{dt} = 0.05P - \underbrace{\boxed{0.06D}}_{\text{Remineralization}}$$



Fennel and Neumann, 2014

Results: Marine Ecosystem Model

Learned system:

$$
\begin{aligned}
\frac{dN}{dt} &= 27.92\mathbb{L}_1(P) + 0.053\mathbb{L}_1(D) - 199.18\mathbb{L}_1(N)\mathbb{L}_1(P) \\
&\quad + 77.13\mathbb{L}_2(N)\mathbb{L}_1(P) - 194.94\mathbb{L}_3(N)\mathbb{L}_1(P) + 27.90\mathbb{L}_4(N)\overbrace{\mathbb{L}_1(P)}^{\text{Legendre Polynomials}} \\
&\quad + 1.12\mathbb{L}_4(P)\mathbb{L}_2(D) - 51.50\mathbb{L}_5(N)\mathbb{L}_1(P) \\
\frac{dP}{dt} &= -28.65\mathbb{L}_1(P) + 199.18\mathbb{L}_1(N)\mathbb{L}_1(P) - 77.13\mathbb{L}_2(N)\mathbb{L}_1(P) + 196.71\mathbb{L}_3(N)\mathbb{L}_1(P) \\
&\quad - 0.94\mathbb{L}_3(N)\mathbb{L}_3(D) - 27.22\mathbb{L}_4(N)\mathbb{L}_1(P) + 52.12\mathbb{L}_5(N)\mathbb{L}_1(P) \\
\frac{dD}{dt} &= 0.0502\mathbb{L}_1(P) - 0.061\mathbb{L}_1(D) - 0.0003\mathbb{L}_3(N)\mathbb{L}_2(D)
\end{aligned}
$$

$$
\begin{aligned}
\frac{dN}{dt} &= 0.51P - 3.40NP + 11.55N^2P - 36.30N^3P \\
&\quad + 124.69N^4P - 382.72N^5P \\
\frac{dP}{dt} &= -0.56P + 3.30NP - 10.78N^2P \\
&\quad + 37.76N^3P - 127.16N^4P + 378.60N^5P \\
\frac{dD}{dt} &= 0.0505P - 0.062D - 0.0002N^2D
\end{aligned}
$$

After symbolic simplification and scale based thresholding

$$\frac{dN}{dt} = 0.51P - P\frac{N}{0.3}\overbrace{\left(1.02 - 1.04\frac{N}{0.3} + 0.98\left(\frac{N}{0.3}\right)^2 - 1.01\left(\frac{N}{0.3}\right)^3 + 0.93\left(\frac{N}{0.3}\right)^4\right)}^{\text{Taylor series}}$$

$$\Longrightarrow \boxed{\frac{dN}{dt} \approx 0.5P - \frac{PN}{0.3+N}}$$

$$\frac{dP}{dt} = -0.56P + P\frac{N}{0.3}\left(0.99 - 0.97\frac{N}{0.3} + 1.02\left(\frac{N}{0.3}\right)^2 - 1.03\left(\frac{N}{0.3}\right)^3 + 0.92\left(\frac{N}{0.3}\right)^4\right)$$

$$\Longrightarrow \boxed{\frac{dP}{dt} \approx -0.50P - 0.06P + \frac{PN}{0.3+N}}$$

$$\frac{dD}{dt} = 0.0505P - 0.062D + 0.00067ND^2$$

$$\Longrightarrow \boxed{\frac{dD}{dt} \approx 0.0505P - 0.062D}$$

Further:

- To accelerate learning, we can use a combination of orthogonal functions and common biogeochemical functional forms such as Michaelis-Menten, etc.

Conclusions and Future Work

- Developed 'dual LASSO feature selection', that relies on the uniqueness of the dual solution for the active set selection, to address the limitations of the current state-of-the-art LASSO based algorithm for model discovery
- Developed a new methodology learns the governing equations from scratch by iteratively building the feature library using appropriate orthogonal functional basis.
- We showcased results of the learning schemes on the classic Lorenz 63 system, a quadratic Lorenz system, and a marine ecosystem model with a non-polynomial nonlinearity.

Future Work:

- Using a mix of larger family of orthogonal functions, kernel composition etc.
- Applications in the presence of model and observation noise, and to higher dimensional systems.
- Using the learned system to guide future observations to close the loop for the Dynamic Data Driven Applications Systems (DDDAS) paradigm.