Sparse Approximation for Nonlinear Dynamics and Stationary Processes

Giang Tran

Department of Applied Mathematics, University of Waterloo, Canada

joint work with

Lam Si Tung Ho, Dalhousie University
Hayden Schaeffer, Carnegie Mellon University
Rachel Ward, University of Texas at Austin

Numerical Analysis and Approximation Theory meet Data Science
Banff International Research Station, April 2018

Problem Set-up

- Given: (possibly noisy or corrupted) samples of a nonlinear continuous function $\mathbf{y} = \mathbf{f}(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^n$

$$\left(\mathbf{x}^1, \mathbf{y}^1\right), \ldots, (\mathbf{x}^m, \mathbf{y}^m)$$

- Goal: recover the underlying equation $\mathbf{f} = (f_1, f_2, \ldots, f_n)$

- Suppose $\mathbf{f} = (f_1, f_2, \ldots, f_n)$ are multivariate polynomials of maximal degree p:

$$f_k(\mathbf{x}) = \sum_{\alpha_1 + \cdots + \alpha_d \leq p} c_\alpha^k x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_d^{\alpha_d}.$$

- Goal is then to recover polynomial coefficients $\{c_\alpha^k\}_{k,\alpha}$

Problem Set-up (cont'd)

Problems of interest:

- Nonlinear dynamical systems with bifurcation
- High-dimensional nonlinear dynamical systems
- Chaotic systems in 3D with corrupted data
- Stationary processes (identically distributed + concentration inequality)

Main ideas: sparse optimization + compressed sensing

Problem Set-up (cont'd)

- Learning nonlinear dynamics:
  - Nonlinear dynamical systems with bifurcation
  - High-dimensional nonlinear dynamical systems
  - Chaotic systems in 3D with corrupted data
- Given m samples (possibly noisy or corrupted) of snapshots:

$$(\mathbf{x}(t_1), \dot{\mathbf{x}}(t_1)), \ldots, (\mathbf{x}(t_m), \dot{\mathbf{x}}(t_m))$$

- Goal: learn the multivariate polynomial $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^d$ s.t.

$$\left.\frac{\mathrm{d}x}{\mathrm{d}t}\right|_{t=t_i} = \mathbf{f}(x(t_i)), \quad i = 1, \ldots, m.$$

Example 1: Nonlinear Systems with Bifurcation

- Given multiple data sets that follow the same physical law, what can we say about its governing equation? For example,

$$\begin{cases} \dot{x}_1 &= 10(x_2 - x_1) \\ \dot{x}_2 &= -x_1 x_3 + (24 - 4\lambda)x_1 + \lambda\, x_2 \\ \dot{x}_3 &= x_1 x_3 - \dfrac{8}{3}x_3, \end{cases}$$

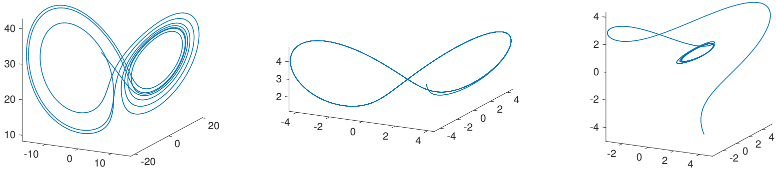


Figure: State space plots for $\lambda = -1$ (left), $\lambda = 7.075$ (middle), and $\lambda = 7.73$ (right), where the dynamics are chaos, a pitchfork bifurcation, and limit cycles, respectively.

Example 2: High-Dimensional Nonlinear Systems

- **Lorenz 96:** a canonical family of ODEs for approximating dynamics of atmosphere:

$$\frac{\mathrm{d}x_k}{\mathrm{d}t} = x_{k+1}x_{k-1} - x_{k-2}x_{k-1} - x_k + F, \quad k = 1, \cdots, d,$$

  where $x_{-1} = x_{d-1},\ x_0 = x_d$, and $x_{d+1} = x_1$ and F is a forcing constant.

- Finite difference discretization of many PDEs with applications range from population dynamics to combustion physics.
  - For example, Fisher's equation can be written as

$$\frac{\mathrm{d}x_k}{\mathrm{d}t} = x_{k+1} - 2x_k + x_{k-1} + \gamma(x_k - x_k^2), \quad k = 1, \cdots, d.$$

"Kernel Trick" to Linearize Problem

- Form data and velocity matrices from given snapshots:

$$X = \begin{bmatrix} | & \cdots & | \\ X_1 & \cdots & X_d \\ | & \cdots & | \end{bmatrix} = \begin{bmatrix} x_1(t_1) & \cdots & x_d(t_1) \\ x_1(t_2) & \cdots & x_d(t_2) \\ \vdots & \cdots & \vdots \\ x_1(t_m) & \cdots & x_d(t_m) \end{bmatrix}_{m \times d}, \quad \dot{X} = \begin{bmatrix} | & \cdots & | \\ \dot{X}_1 & \cdots & \dot{X}_d \\ | & \cdots & | \end{bmatrix}_{m \times d}$$

- Construct *dictionary matrix* from data:

$$\Phi = \begin{bmatrix} | & | & & | & | & | & & | & \\ 1 & X_1 & \cdots & X_d & X_1^2 & X_1 X_2 & \cdots & X_d^2 & \cdots \\ | & | & & | & | & | & & | & \end{bmatrix}_{m \times N}$$

where $N = \binom{p+d}{d}$ is number of multivariate monomials of degree $\leq p$.

"Kernel Trick" to Linearize Problem

- Recovering poly. coefficients $\mathcal{C} = [c_\alpha^1, c_\alpha^2, \dots c_\alpha^d]_{|\alpha| \le p} \in \mathbb{R}^{N \times d}$ as solution to the linear inverse problem [1]

$$\dot{X} = \Phi\, \mathcal{C}.$$

- In the presence of measurement errors (in data, time-derivative approximations,...), the problem becomes

$$\dot{X} = \Phi\, \mathcal{C} + \mathcal{E}.$$

- We will investigate properties of the matrix Φ in various type of input data.

[1] Brunton, Proctor, and Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems", *PNAS* 2016.

Sparse Optimization and Bifurcation

Nonlinear Dynamics with Bifurcation

- Consider the Lorenz system with a single bifurcation parameter λ:

$$\begin{cases} \dot{x}_1 &= 10(x_2 - x_1) \\ \dot{x}_2 &= -x_1 x_3 + (24 - 4\lambda)x_1 + \lambda\, x_2 \\ \dot{x}_3 &= x_1 x_3 - \dfrac{8}{3} x_3, \end{cases}$$

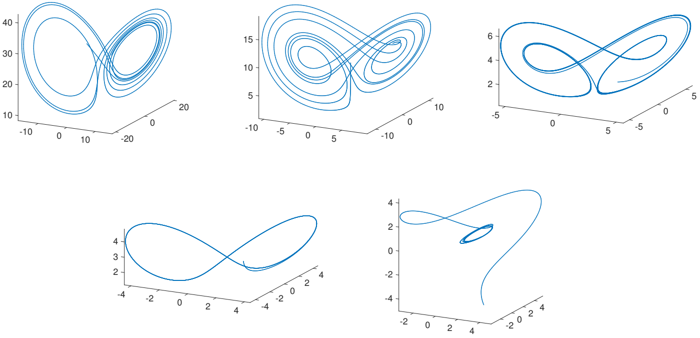


Figure: State space plots for different λ with $dt = 0.005$.

Sparse Group Penalization

- Denote the coefficient matrix

$$C_j = \begin{bmatrix} | & | & | & | \\ c_j^{(1)} & c_j^{(2)} & \dots & c_j^{(m)} \\ | & | & | & | \end{bmatrix}_{\bar{n}\times m}$$

- Observe: The vectors $c_j^{(i)}$ have the same support set for all i!
- Solve the following group-sparse optimization problem:

$$\boxed{\min_{C_j} \sum_{i=1}^{m} \|\Phi^{(i)} c_j^{(i)} - V_j^{(i)}\|_2^2 + \gamma \|C_j\|_{2,0}}$$

where the $\ell^{2,0}$ penalty is defined as:

$$\|A\|_{2,0} := \# \left\{ k : \left(\sum_{\ell} |a_{k,\ell}|^2 \right)^{1/2} \neq 0 \right\}.$$

Numerical Method

- Proximal descent method + Hard-iterative thresholding
- Step 1: Gradient descent

$$\left(\widetilde{c^{(i)}}\right)^{k+1} = \left(c^{(i)}\right)^{k} - (\Phi^{(i)})^{T}\left(\Phi^{(i)}\left(c^{(i)}\right)^{k} - V^{(i)}\right)$$

- Step 2: Hard-iterative thresholding

$$(c^{(i)})^{k+1} = \begin{cases} 0, \quad \text{if} \quad \|row\| < \sqrt{\gamma} \\ \underset{c^{(i)}}{\operatorname{argmin}} \|\Phi^{(i)} c^{(i)} - V^{(i)}\|_2^2 \quad \text{otherwise.} \end{cases}$$

Group Hard-Iterative Thresholding Algorithm

Given: initialization matrix C^0, *tol* and parameters γ.

while $\|C^{k+1} - C^k\|_\infty > tol$ **do**

for $i = 1$ **to** m:

$$\left(\widetilde{c^{(i)}}\right)^{k+1} = \left(c^{(i)}\right)^k - (\Phi^{(i)})^T \left(\Phi^{(i)} \left(c^{(i)}\right)^k - V^{(i)}\right)$$

end for

$$S^{k+1} = \text{supp}\left(H_{\sqrt{\gamma}}\left[\widetilde{c^{(1)}}, \widetilde{c^{(2)}}, \cdots, \widetilde{c^{(m)}}\right]\right)$$

for $i = 1$ **to** m:

$$(c^{(i)})^{k+1} = \underset{c^{(i)}}{\operatorname{argmin}} \|\Phi^{(i)} c^{(i)} - V^{(i)}\|_2^2 \quad \text{s.t.} \ \ \text{supp}(c^{(i)}) \subset S^{k+1}$$

end for

end while

Convergence Guarantees

$$\min_{C} \; F(C) := \sum_{i=1}^{m} \|\Phi^{(i)} c^{(i)} - V^{(i)}\|_2^2 + \gamma \|C\|_{2,0}$$

Theorem

Let C^k be the sequence generated by the proposed numerical scheme, then $F(C^{k+1}) \leq F(C^k)$ and there are subsequences that converge to local minimizers. In addition, if

$$D := diag[\Phi^{(1)}, \ldots, \Phi^{(m)}]$$

is coercive then the sequence C^k converges to a local minimizer.

Convergence Guarantees: General Bound

Proposition (General bound)

Suppose, for each i, $\overline{n} \leq \ell_i$, there exists a subset $S \subset [\ell_i]$ of size $|S| = \overline{n}$ such that $\{X^{(i)}(k, -) \mid k \in S\}$ do not belong to a common algebraic hypersurface of degree $\leq p$.

Convergence Guarantees: General Bound

Proposition (General bound)

Suppose, for each i, $\overline{n} \leq \ell_i$, *there exists a subset* $S \subset [\ell_i]$ *of size* $|S| = \overline{n}$ *such that* $\{X^{(i)}(k, -) \mid k \in S\}$ *do not belong to a common algebraic hypersurface of degree* $\leq p$.

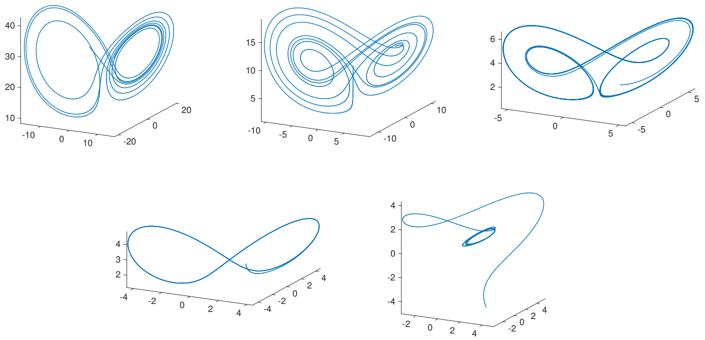
This is a necessary and sufficient condition for the dictionary matrix D *to be full rank: for each* $D^{(i)}$, *there exists a* $\delta_i > 0$ *such that*

$$\inf_u \frac{\|\Phi^{(i)} u\|_2}{\|u\|_2} \geq \delta_i. \tag{1}$$

Numerical Results: Lorenz 3D

- Consider the Lorenz system with a single bifurcation parameter λ:

$$\begin{cases} \dot{x}_1 &= 10(x_2 - x_1) \\ \dot{x}_2 &= -x_1 x_3 + (24 - 4\lambda)x_1 + \lambda\, x_2 \\ \dot{x}_3 &= x_1 x_3 - \dfrac{8}{3} x_3, \end{cases}$$



Numerical Results: Lorenz 3D

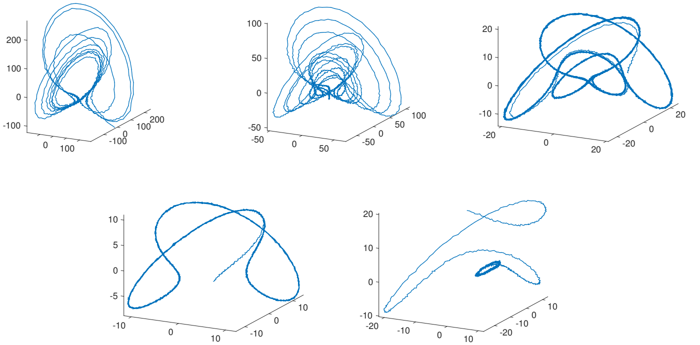


Figure: Noisy velocity space plots corresponding to the data given in Figure 3 with noise level $\sigma_{noise} = 0.5\%$.

Numerical Results: Lorenz 3D

Coeff.	Set 1	Set 2	Set 3	Set 4	Set 5
1	0	0	0	0	0
x_1	28.0232 (28)	5.2104 (5.2)	-3.6068 (-3.6)	-4.2960 (-4.3)	-6.9246 (-6.92)
x_2	-1.0093 (-1.0)	4.6970 (4.7)	6.9020 (6.9)	7.0719 (7.075)	7.7310 (7.73)
x_3	0	0	0	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_1 x_3$	-1.0002 (-1)	-1.0003 (-1)	-0.9989 (-1)	-1.0002 (-1)	-0.9992 (-1)
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_3^4	0	0	0	0	0

Recovered coefficients from all five sets for $\frac{dx_2}{dt}$. The true values are highlighted in (red).

Sparse Approximation in High-Dimensional Nonlinear Dynamical Systems

Sparsity-of-Effect Principle

- For high-dimensional systems ($d \gg 1$), system is usually dominated by main effects and first- and second-order interactions.
  - For d large, consider polynomial dictionary only up to degree $p = 2$, then D is $m \times N$ where $N = \binom{2+d}{d} = \frac{(d+1)(d+2)}{2}$.
- In systems of practical interest, low-order interactions also sparse – exploit!
- Reformulate as a basis pursuit problem for an underdetermined system ($m \ll N$):

$$\boxed{\min \|\mathcal{C}\|_1, \quad \text{s.t} \quad \|\dot{X} - D\,\mathcal{C}\| \leq \sigma}$$

where σ represents error in time-derivative approximations.

Sparse Approximation in High-Dimensional Systems

$$\boxed{\min \|\mathcal{C}\|_1, \quad \text{s.t} \quad \|\dot{X} - D\,\mathcal{C}\| \leq \sigma}$$

- Limitation in theory for high-dimensional nonlinear dynamical system compared to rich (but technical) theory for ergodicity/chaos in 3D nonlinear systems.

- In our work, we show that in many scenarios, we can obtain sparse recovery for $\mathcal{C}$.
  - Main idea: random initializations + multiple trajectories.

Sparse Approximation in High-Dimensional Systems

- Given snapshots from K different trajectories:

$$\{\mathbf{x}(t_1,1),\ldots,\mathbf{x}(t_m,1)\},\quad \{\dot{\mathbf{x}}(t_1,1),\ldots,\dot{\mathbf{x}}(t_m,1)\},\ldots$$
$$\{\mathbf{x}(t_1,K),\ldots,\mathbf{x}(t_m,K)\},\quad \{\dot{\mathbf{x}}(t_1,K),\ldots,\dot{\mathbf{x}}(t_m,K)\}$$

- Form dictionary matrix D of size $mK \times N$ and solve for $\mathcal{C}$.

Sparse Approximation in High-Dimensional Systems

- Given snapshots from K different trajectories:

$$\{\mathbf{x}(t_1,1),\ldots,\mathbf{x}(t_m,1)\},\quad \{\dot{\mathbf{x}}(t_1,1),\ldots,\dot{\mathbf{x}}(t_m,1)\},\ldots$$
$$\{\mathbf{x}(t_1,K),\ldots,\mathbf{x}(t_m,K)\},\quad \{\dot{\mathbf{x}}(t_1,K),\ldots,\dot{\mathbf{x}}(t_m,K)\}$$

- Form dictionary matrix D of size $mK \times N$ and solve for $\mathcal{C}$.

Theorem (Schaeffer, **T'**, and Ward, 2017)

Assume each component of $f(x) = (f_1(x),\ldots,f_d(x))$ is quadratic and has at most s non-zero polynomial coefficients; the K initializations $\{\mathbf{x}(t_1,1),\ldots,\mathbf{x}(t_1,K)\}$ are drawn i.i.d. uniformly from $[-1,1]^d$; and the number of bursts $K \geq 9c_*\, s \log N \log(\varepsilon^{-1})$.

Sparse Approximation in High-Dimensional Systems

- Given snapshots from K different trajectories:

$$\{\mathbf{x}(t_1,1),\ldots,\mathbf{x}(t_m,1)\},\quad \{\dot{\mathbf{x}}(t_1,1),\ldots,\dot{\mathbf{x}}(t_m,1)\},\ldots$$
$$\{\mathbf{x}(t_1,K),\ldots,\mathbf{x}(t_m,K)\},\quad \{\dot{\mathbf{x}}(t_1,K),\ldots,\dot{\mathbf{x}}(t_m,K)\}$$

- Form dictionary matrix D of size $mK \times N$ and solve for $\mathcal{C}$.

Theorem (Schaeffer, T', and Ward, 2017)

Assume each component of $f(x) = (f_1(x), \ldots, f_d(x))$ is quadratic and has at most s non-zero polynomial coefficients; the K initializations $\{\mathbf{x}(t_1,1),\ldots,\mathbf{x}(t_1,K)\}$ are drawn i.i.d. uniformly from $[-1,1]^d$; and the number of bursts $K \geq 9c_* \, s \log N \log(\varepsilon^{-1})$.

Then with probability $1-\varepsilon$, $\mathcal{C}$ is the unique solution to the ℓ_1-minimization problem:

$$\min \|\mathcal{C}\|_1 \quad \textit{subject to} \quad \dot{X} = D\,\mathcal{C},$$

and recovery is stable with respect to inexact sparsity and robust with respect to additive noise (as from approximating derivatives).

Sparse Approximation in High-Dimensional Systems

- The initializations could be taken to be i.i.d. from other distributions such as Gaussian or Chebyshev distribution
  - Choose appropriate orthonormal monomials w.r.t different distributions: uniform dist. vs Legendre polynomials, Gaussian dist. vs Hermite polynomials, ...
- The reconstruction guarantees can be extended to
  - higher-order polynomial systems (the constant in the theoretical result will increase)
  - other bounded orthonormal basis such as sines and cosines,...
- The basis pursuit problem can be solved using *spgl1*[2], *SpaRSA*[3], or *cvx*.

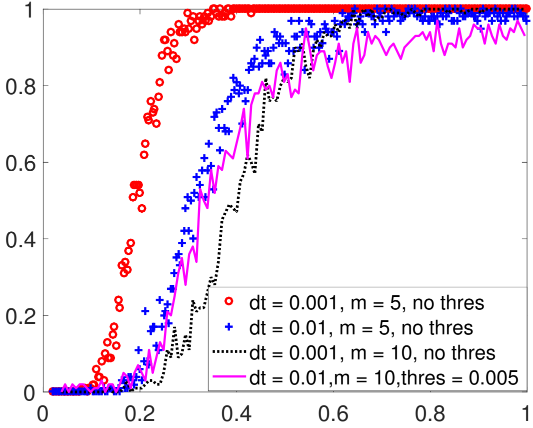
---

[2] Van Den Berg and Friedlander, "Probing the Pareto frontier for basis pursuit solutions", SIAM Journal on Scientific Computing, 2008.

[3] Wright, Nowark, and Figueiredo, "Sparse reconstruction by separable approximation", IEEE Trans. on Signal Processing, 2009.

Example 1: Lorenz 96 – Phase Transition

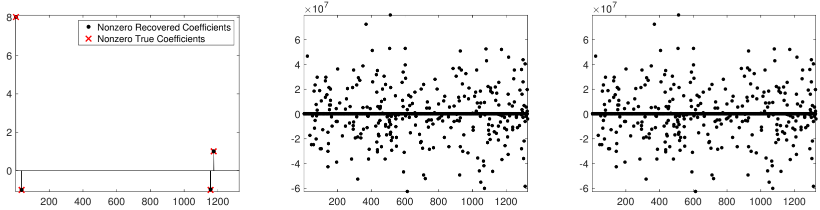
$$\text{Lorenz 96:} \quad \frac{dx_k}{dt} = -x_{k-2}\,x_{k-1} + x_{k-1}\,x_{k+1} - x_k + F, \quad k = 1, \ldots, d$$



Probability of exact recovery vs the undersampling rate K/N with $N = 1326$, $F = 8$.

For dt $= 0.001$, $K = 80$ is needed to achieve 90% prob. of success for both m.

Example 1: Lorenz 96 – Comparison



The coefficients learned from basis pursuit method (left), the least-square algorithm (middle), and the sequential thresholding algorithm (right) for the 35th component of the Lorenz 96 with $d = 50, \mathrm{dt} = 0.001$. The threshold parameter for the last two methods is set to be 0.05.

- The least-square and the sequential thresholding solutions have sparsity $s \gg 10$ and coefficients on the order of 10^7.
- Our solution is 5-sparse in the Legendre basis, when transformed back, it is nearly exact (up to a few significant digits)!

Other Sampling Strategies

- Depend on the degree of prior knowledge about the data or the governing equations, the number of initializations can be reduced.
  - Due to the localization of ODE system discretized from a PDE
    $$K \sim c\, s\, \log(\ell) \log(\varepsilon^{-1}).$$
  - Due to a strong decay of correlations of chaotic systems $\Rightarrow$ Sample both at small time-scale (for time-derivative approximation) and at large-scale (for ergodicity).

Sparse Recovery in Low-Dimensional Nonlinear Dynamical Systems

Corrupted Chaotic Systems

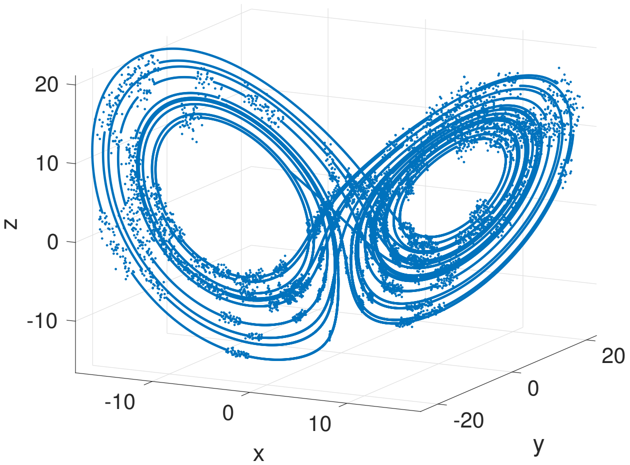


Figure: Lorenz System with 15% Corrupted Data

Problem Formulation

- Under same assumptions as before, now observe data corresponding to the time-Δ map, corrupted by outliers:

$$\mathbf{x}_j^o = \mathbf{x}(t_j) + O_{1,j}, \quad t_j = j\Delta, \quad j = 0, 1, \ldots, m,$$
$$\dot{\mathbf{x}}_j^o = \dot{\mathbf{x}}(t_j) + O_{2,j}, \quad t_j = j\Delta, \quad j = 0, 1, \ldots, m$$

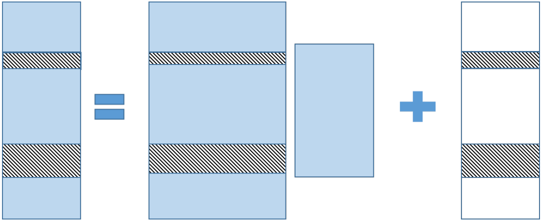
  where $\#\{j : |O_{1,j}| > 0\} + \#\{j : |O_{2,j}| > 0\} \leq s$ for $s < m$ but support is unknown a priori.

- In this setting, given the corrupted data matrices X^o and $\dot{X}^o$, the difference matrix

$$\dot{X}^o - \Phi(X^o)\mathcal{C}$$

  will have at most $2s$ nonzero rows, the locations of which are unknown a priori.

Problem Formulation (cont'd)



- Proposed reconstruction in corrupted data setting: solve the jointly sparse optimization problem

$$(\mathcal{C}, \mathcal{E}) = \operatorname*{argmin}_{\mathcal{C},\mathcal{E}} \|\mathcal{E}\|_{2,1} = \operatorname*{argmin}_{\mathcal{C},\mathcal{E}} \sum_{i=1}^{m} \|\mathcal{E}(i,:)\|_2,$$

$$\text{subject to} \quad \dot{X}^o = \Phi(X^o)\mathcal{C} + \mathcal{E} \quad \text{and} \;\; \mathcal{C} \text{ is sparse.}$$

Numerical Scheme

$$(\mathcal{C}, \mathcal{E}) = \underset{\mathcal{C},\mathcal{E}}{\operatorname{argmin}} \|\mathcal{E}\|_{2,1} = \underset{\mathcal{C},\mathcal{E}}{\operatorname{argmin}} \sum_{i=1}^{m} \|\mathcal{E}(i,:)\|_2,$$

$$\text{subject to} \quad \Phi(X^o)\mathcal{C} + \mathcal{E} = \dot{X}^o \quad \text{and } \mathcal{C} \text{ is sparse.}$$

- The corresponding augmented Lagrangian is of the form

$$(\mathcal{C}, \mathcal{E}, b) = \underset{\mathcal{C},\mathcal{E},b}{\operatorname{argmin}} \sum_{i=1}^{m} \|\mathcal{E}(i,:)\|_2 + \frac{\mu}{2}\|\Phi(X^o)\mathcal{C} + \mathcal{E} - \dot{X}^o + b\|_F^2,$$
$$\text{subject to } \mathcal{C} \text{ is sparse.} \tag{2}$$

- It can be solved via alternating directional method of multipliers (ADMM)/Split Bregman,...

Numerical Scheme (cont'd)

Algorithm

Given: $\mathcal{E}^0, b^0, tol$ and parameters λ, μ.

while $||\mathcal{E}^k - \mathcal{E}^{k-1}||_\infty > tol$ **do**

$$\mathcal{C}^{k+1} = S_h\left((\Phi(X^\circ))^{-1}(\dot{X}^\circ - \mathcal{E}^k - b^k), \lambda\right)$$

$$\mathcal{E}^{k+1} = S_2\left(\dot{X}^\circ - b^k - \Phi(X^\circ)\mathcal{C}^{k+1}, \mu\right)$$

$$b^{k+1} = b^k + \Phi(X^\circ)\mathcal{C}^{k+1} + \mathcal{E}^{k+1} - \dot{X}^\circ$$

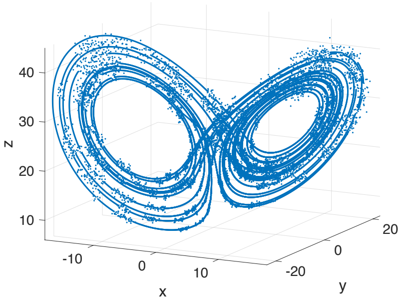
end while

where

$$S_h(u,\gamma) := u \cdot I_{|u|\geq\gamma} = \begin{cases} u & \text{if} \quad |u| \geq \gamma \\ 0 & \text{otherwise,} \end{cases}$$

$$S_2(u_j,\gamma) = \max\left(1 - \frac{1}{\gamma\|u\|_2}, 0\right) u_j, \quad \text{for each row } u_j \text{ of } u.$$

Numerical Results: Lorenz System

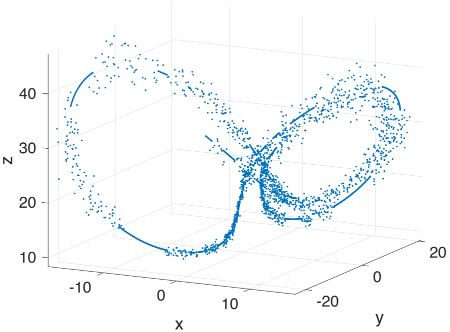


	$\dot{x}$	$\dot{y}$	$\dot{z}$
1	0	0	0
x	-9.9999	27.9995	0
y	9.9999	-0.9999	0
z	0	0	-2.6666
x^2	0	0	0
xy	0	0	0.9999
xz	0	-0.9999	0
y^2	0	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
z^4	0	0	0

Lorenz system $\dot{x} = -10x + 10y$, $\dot{y} = 28x - y - xz$, $\dot{z} = -2.66z + xy$, with 19.19% corrupted data, 40000 measurements, $\Delta t = 0.0005$. The model recovers the coefficients with 0.0096% error and detect exactly the locations of the outliers after 24 iterations.

Lorenz System - Small Sample Size

- 5000 measurements, $\Delta t = 0.0005$, with 71.89% corruption. The model detects exactly the locations of the outliers and recovers the coefficients with 0.0477% error.



- If $T \leq 2$ (4000 measurements), the scheme doesn't work well
  - It's important to have sufficient amount of measurements.

Lorenz Data with Noise

- Add Gaussian noise to the entire data
- Build the dictionary and approximate the time derivative from noisy + corrupted data

Standard Deviation of Noise	# Times Detect Exactly Outliers (over 100)	Coefficient Error (%)
$0.4\Delta t$	89	$\text{min} = 0.0009, \text{max} = 0.0525$
$0.6\Delta t$	87	$\text{min} = 0.0006, \text{max} = 0.9395$
$0.8\Delta t$	65	$\text{min} = 0.0012, \text{max} = 1.57$

Table: Different noise levels and the recovery results associated with the Lorenz system, $\Delta t = 0.0005$, 40000 measurements, and around 20% corrupted

Reconstruction Guarantee Analysis

Theorem (**T'** and Ward, 2016)

Suppose we observe corrupted measurements of the time-1 map

$$X^{\circ,t} = x^t + \Theta_{1,t}, \quad \dot{X}^{\circ,t} = \dot{\mathbf{x}}^t + \Theta_{2,t}, \quad t = 1, 2, \ldots, m,$$

where $\mathbf{x}^t = x(t)$ is the flow generated by a strongly ergodic vector field whose time-1 map satisfies the Central Limit Theorem. Assume the governing equations are multivariate polynomials of degree at most p.

Reconstruction Guarantee Analysis

Theorem (**T'** and Ward, 2016)

Suppose we observe corrupted measurements of the time-1 map

$$X^{\circ,t} = x^t + \Theta_{1,t}, \quad \dot{X}^{\circ,t} = \dot{\mathbf{x}}^t + \Theta_{2,t}, \quad t = 1, 2, \ldots, m,$$

where $\mathbf{x}^t = x(t)$ is the flow generated by a strongly ergodic vector field whose time-1 map satisfies the Central Limit Theorem. Assume the governing equations are multivariate polynomials of degree at most p.

There are constants C, C' (depending on $\Lambda, \|\Theta\|_\infty$, and ε) such that if $m \geq CN$ and $s \leq C'm^{.9}$, then with probability exceeding $1 - \varepsilon$ with respect to $\mathbf{x}_0 \sim d\mu$, the polynomial coefficients and locations of the outliers can be exactly recovered as the solution to the ℓ_1-minimization problem

$$\min_{\mathcal{C},\mathcal{E}} \|\mathcal{E}\|_1 \quad \textit{subject to} \quad \Phi(X^\circ)\mathcal{C} + \mathcal{E} = \dot{X}^\circ$$

Sketch of the Proof

Result from statistical properties of Lorenz-like systems:

- Lorenz equations support a compact, connected attractor Λ and the flow $\dot{\mathbf{x}}(t) = f(\mathbf{x}(t))$ admits a *physical* measure μ.
- Central Limit Theorem for geometric Lorenz attractors: Fix $\eta > 0$. Let $\varphi : \mathbb{R}^3 \to \mathbb{R}$ be a $C^{1+\eta}$ function, and let $Z \sim \mathcal{N}(0, \sigma^2)$. Then there are constants $C_1 > 0$ and $C_{x,\varphi} \geq 0$ such that

$$\left| \frac{1}{m} \sum_{j=0}^{m-1} \varphi(\mathbf{x}^j) - \int_\Lambda \varphi \, d\mu - \frac{\sigma}{\sqrt{m}} Z \right| \leq C_{x,\varphi} m^{-3/4} (\log(m))^{1/2} (\log\log(m))^{1/4}$$

for μ-almost all $x \in \Lambda$, and $\sigma^2 \leq C_1 \|\varphi\|^2_{C^{1+\eta}}$.

---

[4] Tucker, ''The Lorenz attractor exists'', *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, 1999.

[5] Arajo, Melbourne, and Varandas, ''Rapid mixing for the Lorenz attractor and statistical limit laws for their time-1 maps'', *Comm. in Mathematical Physics*, 2015.

Sketch of the Proof

$$\min_{\mathcal{C},\mathcal{E}} \|\mathcal{E}\|_1 \quad \text{subject to} \quad \Phi(X^\circ)\mathcal{C} + \mathcal{E} = \dot{X}^\circ \quad (*)$$

Result from compressed sensing[6]:

- Every $(\mathcal{C}, \mathcal{E})$, satisfying $A\mathcal{C} + \mathcal{E} = y$ and $\mathcal{E} \in \mathbb{R}^m$ is s-sparse, is the unique solution to (*) **if and only if** A is *full column rank* and for every $v \in \mathcal{R}(A) \setminus \{0\}$, the following holds

$$\sum_{j=1}^{2s} |v_{(j)}| < \frac{1}{2} \|v\|_1$$

- Generally, only random A shown to reach optimal sparsity level $s \asymp \frac{m}{\log(m)}$
- We will show that this is also in the setting where $A = \Phi$ is constructed via data from certain chaotic systems!

[6] Bandeira, Scheinberg, and Vincent, ''On partial sparse recovery'', IEEE Signal Processing Letters 2013.

- Indeed, we can prove that the matrix $A = [\Phi_{m\times r}; I_{m\times m}]$ satisfies the null space property:
    - For every $w \in \ker A \setminus \{\vec{0}\}$ and every set $S \subset \{1,\dots,m+r\}$ of cardinality s, the following holds

$$\|w_S\|_1 < \frac{1}{2}\|w\|_1.$$

- The central limit theorem for chaotic systems can be replaced by other concentration inequalities.
- Extend the proof's technique of [T. and Ward, 2017] to a wider class of data that are not required to be independent.

Learning Functions from Stationary Processes

- Suppose we observe corrupted measurements

$$\left(\mathbf{U}^{(i)} = \mathbf{X}^{(i)} + \mathbf{\Theta}^{(i)}, \mathbf{Y}^{(i)} = f(\mathbf{X}^{(i)})\right)_{i=1}^{m} \subset \mathbb{R}^d \times \mathbb{R},$$

- Assume

$$f(x_1, \cdots, x_d) = \sum_{|\alpha| = \alpha_1 + \ldots + \alpha_d \le p} c^{\alpha} x_1^{\alpha_1} \ldots x_d^{\alpha_d}, \quad j = 1, \ldots, n,$$

- Let $y = [Y^{(1)} \ldots Y^{(m)}]^T$ then $y = \Phi\, c + e$.
- Adding sparsity constraints to the solution:

$$\min_{c,e} \|c\|_1 + \|e\|_1 \quad \text{subject to} \quad y = \Phi\, c + e.$$

Assumptions:

- $\max_i \|\Theta^{(i)}\|_\infty \leq B_\Theta, \quad \text{and} \quad \max_i \|X^{(i)}\|_\infty \leq B_{\mathbf{X}}$
- Assume the common distribution μ of $\{X^{(i)}\}_{i=1}^m$ is non-degenerated, i.e., $\mu(X^{(1)} \in A) = 1$ implies A contains infinitely many elements.
- $\{X^{(i)}\}_{i=1}^m$ satisfies the following concentration inequality

$$\Pr\left(\left|\sum_{i=1}^m \varphi(\mathbf{X}_i) - m\mathbb{E}[\varphi(\mathbf{X})]\right| \geq \zeta\right) \leq C_1 \exp\left(-\frac{\zeta^2}{C_2\omega_m + C_3\zeta\kappa_m}\right),$$

  for any bounded Borel function φ.
- $\sqrt{\omega_m \log m} + \kappa_m \log m = o(m)$

Theorem (Ho, T', and Ward, 2018)

Under previous assumptions, there are constants C', C'' depending only on B_Θ, B_X, p such that if

$$m \geq C', \quad s \leq C'' m$$

then the polynomial coefficients of f as well as the outlier vector e can be exactly recovered from the unique solution to the ℓ_1-minimization problem

$$\min_{c,e} \|e\|_1 + \|c\|_1 \quad \textit{subject to } e + \Phi c = y.$$

Sparse Recovery for i.i.d. Random Variables

- **Bernstein inequality for i.i.d. random variables** $(\mathbf{X}_i)$ (Modha and Masry, 1996): Suppose $|\psi(X_1) - \mathbb{E}(\psi(X_1))| \leq d_1$ *a.s.*, then we have

$$\Pr\left(\left|\sum_{i=1}^{m} \psi(\mathbf{X}_i) - m\mathbb{E}[\psi(\mathbf{X})]\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{C_2 m + C_3 t}\right)$$

  where

$$C_2 = 2\mathbb{E}(\psi^2(X_1)) - 2(\mathbb{E}(\psi(X_1)))^2,\ C_3 = \frac{2}{3}d_1$$

  and ψ is any bounded Borel function.

- The samples $\{X^{(t)}\}$ satisfy the concentration inequality with $\omega_m = m$ and $\kappa_m = 1$,

$$\sqrt{\omega_m \log m} + \kappa_m \log m = o(m).$$

- So with probability $\left(1 - \frac{1}{m^\delta}\right)$ for some constant $\delta > 0$, m large enough, the associated ℓ_1-minimization problem has a unique solution.

Sparse Recovery for Exponentially Strongly α-Mixing Processes

- Recall: for a stationary stochastic process $\{X_t\}$, define

$$\alpha(s) = \sup_{\substack{-\infty<t<\infty \\ A\in\sigma(X_t^-), B\in\sigma(X_{t+s}^+)}} |\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)|$$

- The stochastic process is said to be exponentially strongly α-mixing if

$$\alpha(s) \leq \overline{\alpha}\exp(-c_\alpha s^\beta), \quad s \geq 1,$$

  for some $\overline{\alpha} > 0, \beta > 0$, and $c > 0$, where the constants β and c are assumed to be known.

Sparse Recovery for Exponentially Strongly α-Mixing Processes

- Exponentially strongly α-mixing satisfies the following concentration inequality (Modha and Masry, 1996)

$$\Pr\left(\left|\sum_{i=1}^{m}\psi(\mathbf{X}_i) - m\mathbb{E}[\psi(\mathbf{X})]\right| \geq t\right) \leq C_1 \exp\left(-\frac{t^2}{(C_2 m^2 + C_3 tm)/m_\alpha}\right)$$

  where

$$m_\alpha := \left\lfloor \frac{m}{\lceil (8m/c_\alpha)^{1/(\beta+1)} \rceil} \right\rfloor = \mathcal{O}(m^{\beta/(\beta+1)}),$$

$$C_1 = 2(1+4e^{-2}\overline{\alpha}),\ C_2 = 2\mathbb{E}(\psi^2(X_1)) - 2(\mathbb{E}(\psi(X_1)))^2,\ C_3 = \frac{2}{3}d_1$$

  and ψ is any bounded Borel function.
- The samples $\{X^{(t)}\}$ satisfy the concentration inequality with $\omega_m = m^2/m_\alpha,\ \kappa_m = m/m_\alpha,$

$$\sqrt{\omega_m \log m} + \kappa_m \log m = o(m).$$

- So with probability $\left(1 - \frac{1}{m^\delta}\right)$ for some constant $\delta > 0$, m large enough, the associated ℓ_1-minimization problem has a unique solution.

Conclusions and Future Directions

Conclusions

- Shown that the dictionary matrix generated from the polynomial space satisfies range space property, coercivity,...
- Proved the sparse recovery for high dimensional nonlinear systems, 3D chaotic systems, and stationary process with concentration inequalities
- Presented several numerical examples to validate the proposed methods.

Future directions

- Simulate numerical experiments to validate sparse recovery for stationary processes
- Analyze reconstruct guarantees for other dictionary matrices
- Look for real applications

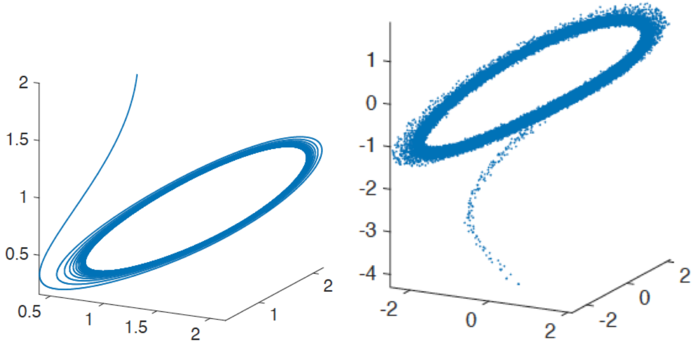


Figure: An example where the state space quickly approaches a limit cycle, which almost stays on a hypersurface of degree 2. The state space is generated from the Lorenz system with $\mu = 7.73$, initialization $U_0 = [1, 1, 2]$, and time step $dt = 0.005$.