Sparse reconstruction of ordinary differential equations with inference

Sara Venkatraman, Sumanta Basu, Martin T. Wells
Cornell University, Department of Statistics and Data Science

Scan for arXiv!

Summary

- Sparse regression has become a popular technique for learning systems of ODEs/PDEs from time series or spatiotemporal data.
- This methodology involves regressing the time derivative of a process on a large set of possible functions, and identifying the (sparse) subset of functions that accurately describe the dynamics of the process.
- Quantifying the uncertainty inherent in the learned differential equations, e.g. via confidence intervals, remains an open problem.
- We propose leveraging recent advances in high-dimensional inference to obtain hypothesis tests and confidence intervals for individual terms comprising a system of ODEs learned from data.
- This significance-driven approach recovers the functional form of a differential equation more accurately than existing methods.

Motivation

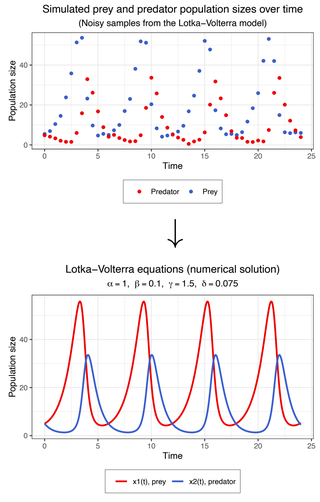
How can we learn the form of a differential equation dx/dt from time series data $x(t_1), ..., x(t_n)$?

- Example – The Lotka-Volterra (LV) equations:
  Let $x_1(t)$ and $x_2(t)$ be the population sizes of a prey species and a predator species, respectively:

$$\frac{dx_1}{dt} = \alpha x_1 - \beta x_1 x_2,$$

$$\frac{dx_2}{dt} = \delta x_1 x_2 - \gamma x_2$$

  where $\alpha, \beta, \gamma, \delta$ are rates of increase/decrease.
- What if we did not know these equations and wanted to learn them to understand the population dynamics, given measurements of $x_1(t)$ and $x_2(t)$ at $t_1, ..., t_n$?



- Purpose: Learned ODEs can be numerically solved for simulation/forecasting, identifying critical parameters, etc.

Problem setup

- Consider a d-dimensional variable $\mathbf{x}$ at time t:

$$\mathbf{x}(t) = [x_1(t), ..., x_d(t)]^T \in \mathbb{R}^d,$$

  whose temporal evolution is governed by:

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}(t)), \text{ for some } \mathbf{f} : \mathbb{R}^d \to \mathbb{R}^d.$$

  Given temporal data $\mathbf{x}(t_1), ..., \mathbf{x}(t_n)$, we want to learn $\mathbf{f}$ in closed form.

- Assumptions: Assume $\mathbf{f}$ has a sparse representation in some basis, e.g. polynomials up to degree k.
  - Why: Many ODEs can be written as a linear combination of components $x_1, ..., x_d$ (or products of them).
  - Example: Lotka-Volterra equations can be written as:

$$\begin{bmatrix} \frac{dx_1}{dt} & \frac{dx_2}{dt} \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_1x_2 \end{bmatrix} \begin{bmatrix} \alpha & 0 \\ 0 & -\gamma \\ -\beta & \delta \end{bmatrix}$$

- Approach: Suppose $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^d$ governs a 2-D ODE system (i.e. $d = 2$) and we assume it is in the span of degree-2 polynomials. Then $d\mathbf{x}/dt = \mathbf{f}(\mathbf{x}(t))$ becomes:

$$\begin{bmatrix} \frac{dx_1}{dt} & \frac{dx_2}{dt} \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_2 & x_1^2 & x_2^2 & x_1x_2 \end{bmatrix} \begin{bmatrix} \beta_{0,1} & \beta_{0,2} \\ \beta_{1,1} & \beta_{2,2} \\ ... & ... \\ \beta_{5,1} & \beta_{5,2} \end{bmatrix}$$

  or more compactly, $\dot{\mathbf{x}} = \mathbf{\Theta}(\mathbf{x})\mathbf{B}$.

- Objective: Estimate the *sparse* coefficient matrix $\mathbf{B}$ from data $\mathbf{x}(t_1), ..., \mathbf{x}(t_n)$ using sparse regression.

- The nonzero entries of the estimated $\mathbf{B}$ indicate which polynomial terms belong in $\mathbf{f}$.

Methodology

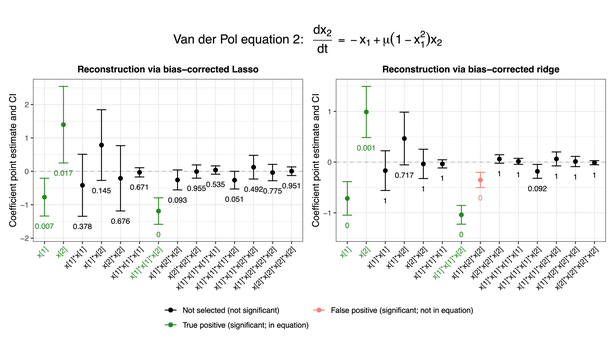
- Regression model setup: Given data $\mathbf{x}(t_1), ..., \mathbf{x}(t_n)$, we have $\dot{\mathbf{X}} = \mathbf{\Theta}(\mathbf{X})\mathbf{B} + \boldsymbol{\varepsilon}$, which we can write out as:

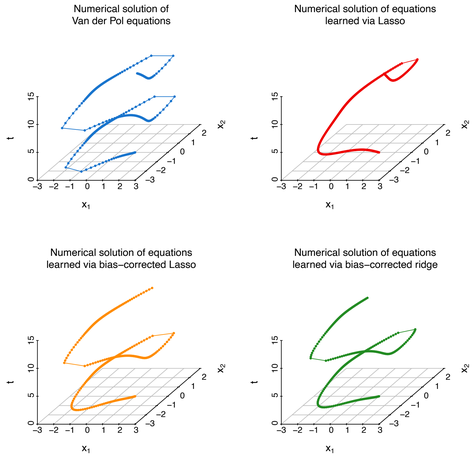
$$\begin{bmatrix} \dot{x}_1(t_1) & \dot{x}_2(t_1) \\ ... & ... \\ \dot{x}_1(t_n) & \dot{x}_2(t_n) \end{bmatrix} = \begin{bmatrix} 1 & x_1(t_1) & x_2(t_1) & x_1^2(t_1) & x_2^2(t_1) & x_1(t_1)x_2(t_1) \\ ... & ... & ... & ... & ... & ... \\ 1 & x_1(t_n) & x_2(t_n) & x_1^2(t_n) & x_2^2(t_n) & x_1(t_n)x_2(t_n) \end{bmatrix} \begin{bmatrix} \beta_{0,1} & \beta_{0,2} \\ ... & ... \\ \beta_{5,1} & \beta_{5,2} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ ... \\ \varepsilon_n \end{bmatrix}$$

  where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$. $\dot{\mathbf{X}}$ is computed via finite differences and $\mathbf{\Theta}(\mathbf{X})$ is computed from the observed data.
- Existing approaches: Estimate each column of the sparse matrix $\mathbf{B}$ via the Lasso or other sparse optimization methods [2, 5, 6]. Identify the relevant polynomial terms as those with non-zero coefficients.
- Limitations of existing methods: Many spurious, high-order polynomial terms tend to be included in the learned function $\mathbf{f}$, with very small coefficients.
  - Need a way to quantify coefficients' uncertainty, e.g. by significance/confidence intervals or inclusion probabilities.
  - Existing uncertainty quantification methods rely on intensive Bayesian inference or resampling [3, 4].
- Our approach: Instead of the Lasso, use new high-dimensional inference techniques that provide hypothesis tests/confidence intervals for regularized regression. These are free of tuning parameters and are computationally efficient.
  - Asymptotic normality of bias-corrected Lasso and ridge estimators yield hypothesis tests for each $H_0 : \beta_{i,j} \neq 0$.
  - We also use SEMMS [1], an empirical Bayes variable selection method for high-dimensional GLMs.
  - We select the polynomial terms that are statistically significant or have high posterior probabilities of being non-zero.

Simulation results

We generate data from the Van der Pol ODE system, $dx_1/dt = x_2, \quad dx_2/dt = -x_1 + \mu(1 - x_1^2)x_2$ by adding noise to the numerical solution with $\mu = 2$, and then try to learn the dx_2/dt equation.





Correct polynomial terms (x_1, x_2, and $x_1^2x_2$) are identified with bias-corrected regression methods as statistically significant.

References

[1] H. Y. Bar, J. G. Booth, and M. T. Wells. A scalable empirical Bayes approach to variable selection in generalized linear models. *Journal of Computational and Graphical Statistics*, 2020.

[2] S. L. Brunton, J. L. Proctor, and J. N. Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, April 2016.

[3] U. Fasel, J. N. Kutz, B. W Brunton, and S. L. Brunton. Ensemble-SINDy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control. *Proceedings of the Royal Society A*, 2022.

[4] S. M. Hirsh, D. A. Barajas-Solano, and J. N. Kutz. Sparsifying priors for bayesian uncertainty quantification in model discovery. *Royal Society Open Science*, 2022.

[5] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 2017.

[6] H. Schaeffer, G. Tran, and R. Ward. Extracting sparse high-dimensional dynamics from limited data. *SIAM Journal on Applied Mathematics*, 2018.