Empirical Dynamic Inference: A methodology for using data to model uncertainty about complex dynamics

D.J. Passey

September 28, 2023

Posing the Question

As we seek to understand the world, that is, identify its forms and comprehend the complex web of connections that underlies the rich and varying patterns of life, we use models.

Most often, these models are efficient mental constructs that we can call up in our minds in order to simulate, intervene and make predictions about the systems in which we live.

It has been the goal of many philosophers, mathematicians and statisticians to formalize these natural cognitive processes. The precise logic of these formalisms enabled us to translate these operations into algorithms that leverage the power of modern compute.

Three Classes of Problems

Before proceeding, we make a loose distinction about three classes of problems.

1. There is a class of problems (and accompanying formal models) that seek to investigate the question, "holding the system (or data generating process) constant, how can I leverage past observations to predict what will happen next?" Some examples of this are recommender systems, classification problems, time series forecasting, or text completion. Many of the formal models used to solve this class of problem fall under the umbrella of machine learning but some, such as probabilistic programming models, do not.

2. A second class of problems seeks to investigate the question, "holding the system constant, how will it respond to a previously unobserved intervention?". This problem occurs when evaluating the effects of proposed government policies, estimating the impact of different pharmaceuticals on the population, or understanding how to intervene in an ecosystem in order to maintain the survival of a species. The methodologies for studying

this class of problem include causal inference, system dynamics, structural equation models and generalized regression.

3. The third class of problems, (likely the most challenging) may have the greatest potential to benefit society. It investigates the question "how can I change the system, or devise a new system that achieves a desired outcome?". Some examples of this are product design, invention, questions such as "how should we structure society?" or "can I create an ecosystem that produces more food than conventional farming". The toolkit for these particular problems remains ill-defined. From the vantage point of this author, it appears that these problems are solved via years of intensive investigation culminating in expert knowledge of *how things work*. This is often the domain of theory, theory that is not always formalized into mathematics, such as Foucault's analysis of the subject and power, the nitrogen cycle, or the macro-deleveraging process.

At this point, an attentive reader has likely thought of problem areas that span multiple of the above classes, or tools that solve multiple problems. It is correct that these problems and their accompanying toolkits are *highly interrelated*. Advanced tools from a certain class of problem can lead to advances in another class. For example, DragonNet is a neural network designed to estimate treatment effects from observational data [6].

DragonNet also illustrates how the problems themselves are interrelated. The model's loss function optimizes both prediction error and treatment assignment error to achieve better causal effect estimates. This illustrates how prediction, as described in item one above, is related to causal interventions as described in problem two.

These concepts are related in the sense that a model which can perfectly predict everything, could also accurately predict treatment effects. However, in practice, algorithms that only minimize prediction error are typically insufficient to elucidate causal effect in the presence of confounders. Analogously, models that accurately estimate causal effect (such as regressions) have high prediction error. (Note that this does not illustrate some universal tradeoff beteen causal effect and prediction—it is possible that someone will invent an algorithm that can do both.)

However, this distinction is still useful because it can guide problem solvers to methodologies that are likely to be helpful as they seek to answer important questions. While it is possible to turn many problems into a machine learning problems, it may be more profitable to address some questions with causal analysis or PDE simulation. Even the AI giants like Facebook and Google employ Bayesian techniques and AB testing to evaluate new features. Similarly engineering firms all over the world use PDEs to simulate products before they are fabricated.

The research presented in this work will focus primarily on problem two and as much as possible attempt to extend results to problem three. The goal of this work is to begin the development of a formal methodology for leveraging

data to effectively quantify uncertainty during to process of creating, assessing and comparing explanatory dynamic models that seek to accurately describe real world relationships through time.

How to Evaluate Success

The problem described above is a near universal problem in the sciences, and whether considered directly or indirectly, it is confronted in many fields of work. There is an extensive suite of methods for studying time series data and drawing conclusions about variable relationships. Yet the kinds of conclusions generated and the assumptions of the techniques vary. In addition, the lack of uncertainty quantification makes many of them ill-suited for scientific debate.

The primary goal of this work to is to increase our ability to explain and understand reality. This goal rests on the assumption that there is a logic to the world that can be uncovered and understood. Therefore, we evaluate methodologies on the quality of the conclusions that can be drawn from their analyses. This evaluation will include the following:

1. A discussion of the assumptions made by the method and the difficulty of testing the assumptions.
2. Benchmarking the method on datasets that meet its assumptions.
3. How well the methodology surrounding the method resolves competing explanations.
4. An analysis of the ability of the method to make convincing arguments about the structure of the natural world.

Discussion of Problem 2: Extrapolation

Relevance and Importance

We'll begin with a few examples of situations where problem 2 is important. For each example, assume that no similar intervention has been observed.

- (Ecology) How would coral reef growth be affected if the government increased the length of the fishing season?
- (Molecular Biology) What would happen to the cell wall if we inhibited the expression of a particular gene?
- (Economics) How would small businesses be affected if we increased the minimum wage in a particular city?
- (Social Science) How would increasing the number of black teachers affect future income of black students?

- (Systems Biology) What would happen to the interaction between gut microbiome and mood in response to a specific diet?

It is important to note that machine learning models are not typically trusted to answer these sorts of questions. This may be because these questions are strictly concerned with out of distribution data and machine learning algorithms have no measure of uncertainty about their predictions. Bayesian models for causal inference estimate causal effect and also provide a measure of how well particular causal question could be answered by the available data. While it may be the case that on average ML models make more accurate predictions, there is no way of assessing the risk profile of basing policy on the predictions of a particular ML model whereas standard statistical techniques lay bare the assumptions and the uncertainty surrounding their predictions.

Important Concepts

- **State space exploration:** An important concept, especially for non-parametric methods is the degree to which the trajectory or set of trajectories explores the space. For the problem of extrapolation, we are concerned with an area of the state space that has not been explored, so we are most interested in understanding how much the observed data tells us about the effects of our posed intervention.

- **Incorporating known mechanisms:** Often, we have knowledge about the world that is relevant to extrapolation. For instance, we know that a change in shark populations will not impact ocean salinity, or that a person with a home loan must make payments on that loan or default. The causal inference methodology uses causal graphs to incorporate known mechanisms into their analyses, similarly, physics informed ML builds physical assumptions into the loss function. It is important to remember that a computational model knows far less about the world than the researcher. All it knows is its structure and the data it sees. Incorporating mechanisms into a model ensures that when it makes out of distribution predictions, they will at least satisfy known mechanisms.

- **Causality and dynamics:** The school of Pearlian causal inference has a strong and precise definition of causality and it's practicioners can be eager to cast doubt on alternate formulations-often rightfully so. However, the Pearlian school mainly restricts it's analyses to static data and assumes that the direction and magnitude of causation is independent of system state (Is this really correct?). Many systems are known to exhibit state dependent causal effects and even show the reversal of direction of the effect.

  Additionally, the Pearlian school approaches data with the maxim, "The causes are not in the data" illustrating that you must approach analyses with strong causal assumptions in order to correctly uncover causal effects.

While this is a good maxim for static data, time series data contains information about the **sequence of events** which empowers algorithms to make inferences about causes.

Clearly, in the space of dynamics, we need new definitions of causality to meet our needs.

A Survey of Related Work

Detecting Causality in Complex Ecosystems (Convergent Cross Mapping/Empirical Dynamic Modeling)

A 2012 paper that presented a method for identifying causality in non-linear dynamical systems [8]. It solves three problems with Granger causality:

1. Granger causality assumes separable variables, that is, by removing a variable, X from the data, all information about X is lost and not included in another variable
2. Granger causality fails to identify weakly coupled variables
3. Granger causality cannot distinguish interactions from external forces

It uses a dynamical systems definition of causality, where two time series variables are causally linked if they belong to the same system. (I assume that means that you can't decompose the system into a smaller system that excludes one of the variables.)

The paper contains interesting ecology datasets such as a sardine anchovy dataset and a paramecium one.

Time Series Analysis (Hamilton)

Apparently a seminal text [1]. Focuses mainly on autoregressive models. Chapter on Kalman filters. The end of the book discusses concepts like cointegration and heteroskedasticity that could be interesting. Link to pdf.

Time Series Analysis Handbook

Notebooks with code on github. Compiled by PhDs in data science at the Asian Institute of Management in 2020-2023. It has convergent cross mapping and empirical dynamic modeling in it with lots of code and datasets.

From Ordinary Differential Equations to Structural Causal Models: The Deterministic Case

A theoretical work that connects differential equation to Pearlian notions of causality and suggests that structural equation models can describe a differential equation [3].

Causal inference for time series

A nature review paper of the structural causal model methodology [5]. Link to PDF.

Recent developments in empirical dynamic modelling

Taken from the abstract: "Recent extensions of EDM to multivariate time series substantially expand the range of applications and mechanistic questions that can be addressed, including detecting causal coupling, tracking changing interactions in real time, leveraging short time series from information shared in coupled variables, modelling dynamically changing stability, scenario exploration, and management applications involving optimal control" [4].

This makes me wonder if there is a way to incorporate Bayesian techniques into this approach.

Ecological Modeling from Time-Series Inference: Insight into Dynamics and Stability of Intestinal Microbiota

Uses a generalized Lotka-Voltera to model and assess stability of the intestinal mircobiome [7].

dynGENIE3: dynamical GENIE3 for the inference of gene networks from time series expression data

Comapres an ML based method for identifying gene regulatory networks with a number of other methods. Is only beaten by Gaussian processes (which is computationally complex) [2]. The paper compares 13 methods in total on a gene regulatory benchmark dataset. They classify the methods into five categories:

1. Tree Ensembles
2. Mutual information
3. Dynamic Bayesian networks
4. ODEs
5. Non-linear Dynamical Systems
6. Granger Causality

It would be very interesting to learn more about each of these classes of methods.

Experimentation

Each methodology was used to make inferences in three scenarios.

1. Simulations of models where the underlying system is completely characterized. (Stochastic differential equations, differential equations with measurement error introduced.)

2. Real world data, where the underlying system is well understood by the academic community. (Modeling the evolution of the nitrogen cycle or the impact of interest rates on inflation.)

3. Real world data where the true system is unknown. (While there is no "ground truth" in this scenario, the value here involves assessing how well the methodology provides the researcher with confidence about the validity of results.)

References

[1] James D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

[2] Vân Anh Huynh-Thu and Pierre Geurts. dyngenie3: dynamical genie3 for the inference of gene networks from time series expression data. *Scientific Reports*, 8(1):3384, 2018.

[3] Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. From ordinary differential equations to structural causal models: The deterministic case. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, page 440–448, Arlington, Virginia, USA, 2013. AUAI Press.

[4] Stephan B. Munch, Tanya L. Rogers, and George Sugihara. Recent developments in empirical dynamic modelling. *Methods in Ecology and Evolution*, 14(3):732–745, 2023.

[5] Jakob Runge, Andreas Gerhardus, Gherardo Varando, Veronika Eyring, and Gustau Camps-Valls. Causal inference for time series. *Nature Reviews Earth & Environment*, 4(7):487–505, Jul 2023.

[6] Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[7] Richard R. Stein, Vanni Bucci, Nora C. Toussaint, Charlie G. Buffie, Gunnar Rätsch, Eric G. Pamer, Chris Sander, and João B. Xavier. Ecological modeling from time-series inference: Insight into dynamics and stability of intestinal microbiota. *PLoS Comput Biol*, 9, 2012.

[8] George Sugihara, Robert May, Hao Ye, Chih hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. Detecting causality in complex ecosystems. *Science*, 338(6106):496–500, 2012.