Inferring effective state variables and dynamics from data

Grenoble AI for Physics workshop
May 29, 2024

Nicolas Brodu

Information on the project https://team.inria.fr/comcausa/
Context / Goal: Modeling from observations

**Observations** → **Model** → **Predictions**

- From data
- Principled construction

**Parameters**
- Interpretable (= characteristic scales)
- Few meta-parameters

**Inferred**
- Effective state variables
- Dynamics at given scales

**Use case = complex system**
- Example: a forest
- No access to all micro. parameters
- No known global states / equations
- Macro. data can be measured
Modeling panorama

Intro

First principles

- Neural Networks
- Model of data dynamics
- Data simulations (ODE, PDE, SDE...)
- Discrete elements (Active matter, mechanics...)
- Evolution equations

ML guided by physics

- Make dynamics linear
  - Koopman operator
  - Hidden Markov Models

Hidden / latent variables

- Observable = function

Function dictionary

- Data assimilation

Parametric models (Gaussian Mixture...)

Statistics

- Time series analysis (auto-regressive, moving average...)
- Neural Networks
- Neural Networks
- Hidden Markov Models

- Regression
- Bayes rule
- Training
- Inference

Causal states

- Make dynamics Markovian

∞ Markov order, ∞ states

Markov order, states
Notion of causal states

Causal state = \{ \text{all pasts: same } P(\text{future} \mid \text{past}) \}

Same causes, same consequences
Intrinsic properties

States do not depend on the frame of reference
⇒ **intrinsic property of the physical process**

No new observation can distinguish two past sequences in the same causal state
⇒ **finest building blocks for modeling**

No dependency left on histories
⇒ **Markovian dynamics**

---

Known generative model

Ensemble of observations

- States do not depend on the frame of reference
  ⇒ **intrinsic property of the physical process**

- No new observation can distinguish two past sequences in the same causal state
  ⇒ **finest building blocks for modeling**

- No dependency left on histories
  ⇒ **Markovian dynamics**

**Causal state** = \{all pasts: same P(future | past)\}
Dissipative pendulum

\[
\frac{d\theta}{dt} = \dot{\theta} \\
\frac{d\dot{\theta}}{dt} = -\frac{b}{m} \dot{\theta} - \frac{g}{L} \sin \theta
\]

Causal states = points in phase space

– Unique trajectory = future
– No dependency on the past
Effective state variables, recovering laws of motion

Dissipative pendulum

\[
\begin{align*}
\frac{d\theta}{dt} &= \dot{\theta} \\
\frac{d\dot{\theta}}{dt} &= -\frac{b}{m} \dot{\theta} - \frac{g}{L} \sin \theta
\end{align*}
\]

Causal states = points in phase space
- Unique trajectory = future
- No dependency on the past

Causal states are independent from referential
- Build one from data \(\Rightarrow\) other state variables
- Equivalently valid law of motion
Effective state variables, recovering laws of motion

Dissipative pendulum

\[ \frac{d\theta}{dt} = \dot{\theta} \]
\[ \frac{d\dot{\theta}}{dt} = -\frac{b}{m} \dot{\theta} - \frac{g}{L} \sin \theta \]

Causal states = points in phase space
- Unique trajectory = future
- No dependency on the past

Causal states are independent from referential
- Build one from data \( \Rightarrow \) other state variables
- Equivalently valid law of motion

Reconstruction from measured data
- Model interpretation = effective state variables + evolution at that data scale
- Capturing effective physical law \( \Rightarrow \) may generalize better out of observed samples
Case study on real data: monthly sunspots observations

Data: SILSO
Inferred state variables
- 11-years cycle (x and y)
- Amplitude modulations (z)

Trajectories on a structure resembling an attractor embedding
Inferred state variables
- 11-years cycle (x and y)
- Amplitude modulations (z)

Trajectories on a structure resembling an attractor embedding

Predictions
- Trajectory constrained on the structure
- Linear operator converging to the average state
Sunspots – Predictions in the data space

Measured data

Prediction (trajectory)
Prediction (linear model)
Noisy Lorenz system (SDE)

\[
\begin{align*}
du &= -a (u - v) \, dt + \eta dW \\
dv &= (bu - v - uw) \, dt + \eta dW \\
dw &= (-cw + uv) \, dt + \eta dW \\
(a, b, c) &= (10, 28, 8/3)
\end{align*}
\]

Data

- Simulation: \( V(t) = [u, v, w](t) \)
- Addition of Gaussian noise, var. \( \sigma^2 \)
- Measurements: \( M(t) = V(t) + G(\sigma^2) \)

Causal states

- Adding noise does not change equiv. classes \( \Rightarrow \) robust to measurement noise!
- Intrinsic noise: SDE details preserved
Observables = strings $\in \{A,B\}^*$

Example: Even process

$$s = \ldots B A A B B A A A A B B B A A A B A B \ldots$$

even numbers of As

[Crutchfield 1988]

Minimal edge-emitting & unifilar HMM $\Rightarrow$ States are causal states

NOT a state-emitting HMM

Finite length artifact (sequences of all A)

Formal system

Inferred from data

Cluster
Generic HMM, IFS

Generic HMM $\Rightarrow$ Iterated Function System

Symbols $i \in \{A, B, C\}$

$s_{n+1} = f^i(s_n)$ with $p(i|s)$

Observables = strings $\in \{A, B, C\}^*$

Infer...BACBBC...ACABBC...

Jump dynamics

Generate

Ensemble of observations past $\rightarrow$ future

...BACBBCACBCAB...
...ACABBCBACCCA...
...CBACBAACBABC...

Using 25k samples

Infer

Number of samples 8313 8313 8303

[Causal states]

Using 25k samples

Observables = strings $\in \{A, B, C\}^*$

Number of samples 8313 8313 8303

[Causal states]

[Causal states]

[Causal states]

[Causal states]
Step 1: A functional space for representing conditional distributions

Step 2: Geometric structure of the causal states set

Step 3: Parametrize that set structure $\Rightarrow$ coordinates = effective variables
A functional space for representing conditional distributions

Causal state = \{\text{same } P(\text{future} | \text{past})\} \Rightarrow \text{whole class mapped to the same point}

Some possibilities: explicit functional basis, neural networks, information geometry…

Easier and convenient: Reproducing Kernel Hilbert Spaces (RKHS)

[Brodu 2022]
[Loomis 2023]
RKHS embedding of distributions - Some intuition

Density estimation

\[ \hat{p}(y') = \frac{1}{N} \sum_{i} k(y_{i}, y') \]

kernel centered on \(y_{i}\)

sample values
Theory

RKHS embedding of distributions - Some intuition

Density estimation

\[ \hat{p}(y') = \frac{1}{N} \sum_{i} k(y_i, y') \]

Geometric view = barycenter

\[
\begin{align*}
k(y_1, \cdot) & \times k(y_6, \cdot) & \times k(y_3, \cdot) \\
k(y_2, \cdot) & \times \hat{p} & \times k(y_5, \cdot) \\
k(y_8, \cdot) & \times & \times k(y_7, \cdot) \\
& & k(y_4, \cdot)
\end{align*}
\]
RKHS embedding of distributions - Some intuition

Density estimation

\[ \hat{p}(y') = \frac{1}{N} \sum_i k(y_i, y') \]

Kernel centered on \( y_i \)

Conditional distribution = weighted average

\[ \hat{p}(Y|X = x) = \sum_i \alpha_i(x)k(y_i, \cdot) \]

\( \alpha_i(x) = \text{proximity with } x_i, \text{ using } k'(x_i, \cdot) \)

Sample values = pairs \((x_i, y_i)\)

Geometric view = barycenter

Kernel centered on \( y_i \)

Sample values = pairs \((x_i, y_i)\)

Conditional density estimation

\[
\begin{align*}
\hat{p}(Y|X = x) &= \sum_i \alpha_i(x)k(y_i, \cdot) \\
\alpha_i(x) &= \text{proximity with } x_i, \text{ using } k'(x_i, \cdot) \\
\text{Sample values} &= \text{pairs } (x_i, y_i)
\end{align*}
\]
Reproducing kernel property

\[ \langle k(x, \cdot), f \rangle_{\mathcal{H}} = f(x) \]

Kernels can be combined, scaled

\[ \mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2 \]

\[ k^V(v, v') = k^T(t, t') \cdot k^P(p, p') \cdot k^E(e, e') \]

\[ \mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2 \]

\[ k_{\text{seq}} = \sum_{\tau=0}^{L} w_{\tau} k(x_{t-\tau}, \cdot) \]

Many known kernels, many data types (scalars, vectors, graphs, strings…)

Kernels act on dimensionless data

\[ k^V\left(\frac{\nu}{\lambda}, \frac{\nu'}{\lambda}\right) = \exp\left(-\frac{\|\nu - \nu'\|^2}{\lambda^2}\right) \]

⇒ Consistently merge heterogeneous data sources

E.g., T = temperature, P = precipitations, E = evapotranspiration

⇒ A scale is needed for each source
Hypothesis: system described by M state parameters

⇒ causal states set dimension at most M (manifold, fractal…)

Physical property: does not depend on data size N

States lying on a reduced dimension set $M \ll N$

$\hat{p}(Y|X = x_i)$

Trajectory

RKHS embedding of dimension $N$
Theory

Steps 3: Coordinates ⇒ Effective state variables

Hypothesis: system described by $M$ state parameters

⇒ causal states set dimension at most $M$ (manifold, fractal...)

Physical property: does not depend on data size $N$

States lying on a reduced dimension set $M \ll N$

Choice of basis = Generalized Fourier modes of minimally embedding manifold

- Causal states = distributions over future ⇒ Additional modes refine predictive info
- Estimated with diffusion maps ⇒ $M$ inferred from an eigenspectrum decay profile

[Coifman 2006, Berry 2020]

[RKHS embedding of dimension $N$]

$\hat{p}(Y|X = x_i)$

[Trajectory]

[Brodu 2022 + in prep.]
Data

– Experimental field, Grignon FR
– Temperature, soil humidity, sun illumination, évapotranspiration, precipitations
– ICOS data (flux tower + in situ sensors), 11 years of daily observations

Dynamics

– Culture of corn then ⇒ plant response is ≠
– Impossible to distinguish visually on raw data

Practice CO₂ Flux and evapotranspiration : Grignon site (INRAE)
Molecular dynamics of n-butane

**Inputs**
- Positions $x,y,z$ of atoms sampled every 200 fs
- Local frame of reference

**Dynamics**
- Clusters = meta-stable conformations
- Sub-clusters = hydrogen position (chemically equivalent, distinct in data)
- Slow clusters + fast transitions \(\Rightarrow\) discrete dynamics at large $\Delta t$

Data: Stefan Klus

Dihedral angle

360 300 240 180 120 60

Gauche

Anti

Eclipsed A

Eclipsed B

Eclipsed B

Eclipsed B
Data

- 50 years of measurements, very high quality
- 49 watersheds, Peruvian coast
- Pacific Ocean: 4 indices of sea surface temperature
- Per watershed: precipitations, runoff, evapotranspiration, temperature

Project with: Luc Bourrel (France), Pedro Rau (Peru)
**Inferred state variables**

– Seasonal cycle
– General amplitude of the oscillation
– Much more than 3 dimensions… difficult to visualize

**Anomaly detection**

– 2016 weaker ⇔ closer to the regular structure
– Conditions Niña at the center of the structure
– Deviations of trajectories months before

**Work in progress - plans**

– Trying various predictive models
– Regionalization: inference of local models for each hydroclimatically consistent region
Diversity of cases to model:

- ODE / SDE: Causal states = 1-to-1 mapping with state space
- Finite states, edge-emitting unifilar HMM
- General HMM, IFS, infinite states

Data-based reconstruction could be anything in between!

- Propose an encoding of causal states \( s_t \equiv (e_t^1, e_t^2, \ldots, e_t^M) \)
- Fit the transition dynamics with that encoding

\[
s_{t+1} - s_t \sim P(\Delta s, a \mid s_t) \quad \text{with density} \quad p(\Delta s, a) = f(e_t^1, e_t^2, \ldots, e_t^M)
\]

\[
ds = a(s)dt + b(s)dW \quad \text{if} \quad dJ(s) = 0 \quad \text{otherwise} \quad s' \sim P(s' \mid s)
\]

- Work in progress: many ML tools and possibilities
Theory

Quantifying the diffusion of information

Measured data

Initial causal state or distribution \( Q_0 \)

\[ T^\tau [Q_0] \]

SDE \( \Leftrightarrow \) Diffusion: distribution of states

Predictions

Limit distribution of the SDE

\( Q_\tau \)

\( Q_{\tau_1} \)

\( Q_\infty \)

Converging in the limit to the data mean

Initial information (latest measurements) completely lost

\( T = \) transfer operator

(in theory…) This model specifies how predictive information diffuses through time!
## Modeling
- Complex Systems
- Machine learning approach (≠ neural nets)
- Modeling the dynamics at the scale of data

## Understanding
- Principled construction
- Generalizes ODE / SDE
- Effective state variables + dynamics

---

## Heterogeneous data
- Temporal data (spatial extension possible)
- (quasi-)arbitrary, (non-numerical possible)
- Multiple sources combined properly (RKHS)

## Few parameters, interpretable
- Length of the past / future causal dependencies
- Characteristic scales of each data source
- A few more internal meta-parameters (kernel)

---

## Inputs

**Effective state variables**
- Condense predictive info. at data scale
- Link with known mechanisms?

**Predictive model**
- Dynamics in reduced dimensions
- Work in progress…

---

**Outputs**

**Application example: Anomaly detection**

- Condense predictive info. at data scale
- Link with known mechanisms?

**Inferred dynamics**

**Anomalies**

- Work in progress…