Markovian, Predictive, and Conceivably Causal Representations of Stochastic Processes

Cosma Shalizi

Statistics Dept., Carnegie Mellon University & Santa Fe Institute

22 October 2010
RSS “Complexity and Statistics”

Predictability and Complexity

The common intuition: complex systems are ones which are hard to describe

Predictability and Complexity

The common intuition: complex systems are ones which are hard to describe

Noise is easy to describe — if you know statistics

Predictability and Complexity

The common intuition: complex systems are ones which are hard to describe

Noise is easy to describe — if you know statistics

Describe by predicting, to check ourselves

Predictability and Complexity

The common intuition: complex systems are ones which are hard to describe

Noise is easy to describe — if you know statistics

Describe by predicting, to check ourselves

What resources do we need to predict?

Predictability and Complexity

The common intuition: complex systems are ones which are hard to describe

Noise is easy to describe — if you know statistics

Describe by predicting, to check ourselves

What resources do we need to predict?

1. Data points to fit a model (sample complexity)

Predictability and Complexity

The common intuition: complex systems are ones which are hard to describe

Noise is easy to describe — if you know statistics

Describe by predicting, to check ourselves

What resources do we need to predict?

1. Data points to fit a model (sample complexity)
2. History to extrapolate from (forecasting complexity)

Predictability and Complexity

The common intuition: complex systems are ones which are hard to describe

Noise is easy to describe — if you know statistics

Describe by predicting, to check ourselves

What resources do we need to predict?

1. Data points to fit a model (sample complexity)
2. History to extrapolate from (forecasting complexity)
3. Computing time to calculate (computational complexity)

Predictability and Complexity

The common intuition: complex systems are ones which are hard to describe

Noise is easy to describe — if you know statistics

Describe by predicting, to check ourselves

What resources do we need to predict?

1. Data points to fit a model (sample complexity)
2. History to extrapolate from (forecasting complexity)
3. Computing time to calculate (computational complexity)

(2) brings together dynamics, statistics and information theory

“State”

Dynamics: “state” = variable now which fixes all future observables

“State”

Dynamics: “state” = variable now which fixes all future observables
Allow indeterminism: state fixes *distribution* of future observables

"State"

Dynamics: "state" = variable now which fixes all future observables
Allow indeterminism: state fixes *distribution* of future observables
Would like the state to be and well-behaved

e.g., homogeneous Markov

"State"

Dynamics: "state" = variable now which fixes all future observables

Allow indeterminism: state fixes *distribution* of future observables

Would like the state to be and well-behaved

e.g., homogeneous Markov

Try construct states by constructing predictions

How many do we need? Organized how?

Notation etc.

Upper-case letters are random variables, lower-case their realizations

Stochastic process $\ldots, X_{-1}, X_0, X_1, X_2, \ldots$

$X_s^t = (X_s, X_{s+1}, \ldots X_{t-1}, X_t)$

Past up to and including t is $X_{-\infty}^t$, future is X_{t+1}^{∞}

Discrete time is *not* required but cuts down on measure theory

Making a Prediction

Look at $X_{-\infty}^{t}$, make a guess about X_{t+1}^{∞}
Most general guess is a probability distribution
Only ever attend to selected aspects of $X_{-\infty}^{t}$
mean, variance, phase of 1st three Fourier modes, ...
$\therefore$ guess is a function or **statistic** of $X_{-\infty}^{t}$
Good statistic: summarize as much as possible while keeping predictive power

Predictive Sufficiency

Data-processing inequality: For any statistic σ,

$$I[X_{t+1}^{\infty}; X_{-\infty}^{t}] \geq I[X_{t+1}^{\infty}; \sigma(X_{-\infty}^{t})]$$

σ is **predictively sufficient** iff

$$I[X_{t+1}^{\infty}; X_{-\infty}^{t}] = I[X_{t+1}^{\infty}; \sigma(X_{-\infty}^{t})]$$

Sufficient statistics retain all predictive information in the data

Predictive Sufficiency

Data-processing inequality: For any statistic σ,

$$I[X_{t+1}^{\infty}; X_{-\infty}^{t}] \geq I[X_{t+1}^{\infty}; \sigma(X_{-\infty}^{t})]$$

σ is **predictively sufficient** iff

$$I[X_{t+1}^{\infty}; X_{-\infty}^{t}] = I[X_{t+1}^{\infty}; \sigma(X_{-\infty}^{t})]$$

Sufficient statistics retain all predictive information in the data
$\therefore$ Minimizing any loss function only needs a sufficient statistic (Blackwell & Girshick)

Excuse for not worrying about particular loss functions

“Causal” States

(Crutchfield and Young, 1989)

Histories a and b are equivalent iff

$$\Pr\left(X_{t+1}^{\infty} | X_{-\infty}^{t} = a\right) = \Pr\left(X_{t+1}^{\infty} | X_{-\infty}^{t} = b\right)$$

"Causal" States

(Crutchfield and Young, 1989)

Histories a and b are equivalent iff

$$\Pr\left(X_{t+1}^{\infty} | X_{-\infty}^{t} = a\right) = \Pr\left(X_{t+1}^{\infty} | X_{-\infty}^{t} = b\right)$$

$[a] \equiv$ all histories equivalent to a

"Causal" States

(Crutchfield and Young, 1989)

Histories a and b are equivalent iff

$$\Pr\left(X_{t+1}^{\infty} | X_{-\infty}^{t} = a\right) = \Pr\left(X_{t+1}^{\infty} | X_{-\infty}^{t} = b\right)$$

$[a] \equiv$ all histories equivalent to a
The statistic of interest, the **causal state**, is

$$\epsilon(x_{-\infty}^{t}) = [x_{-\infty}^{t}]$$

Set $s_t = \epsilon(x_{-\infty}^{t})$

"Causal" States

(Crutchfield and Young, 1989)

Histories a and b are equivalent iff

$$\Pr\left(X_{t+1}^{\infty} | X_{-\infty}^{t} = a\right) = \Pr\left(X_{t+1}^{\infty} | X_{-\infty}^{t} = b\right)$$

$[a] \equiv$ all histories equivalent to a
The statistic of interest, the **causal state**, is

$$\epsilon(x_{-\infty}^{t}) = [x_{-\infty}^{t}]$$

Set $s_t = \epsilon(x_{-\infty}^{t})$
A state is an equivalence class of histories *and* a distribution over future events

"Causal" States

(Crutchfield and Young, 1989)

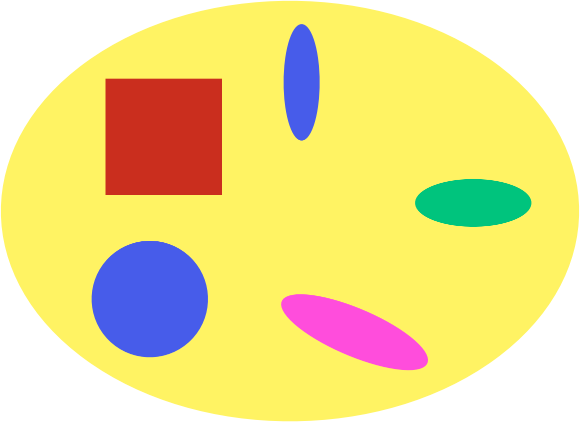
Histories *a* and *b* are equivalent iff

$$\Pr\left(X_{t+1}^{\infty} | X_{-\infty}^{t} = a\right) = \Pr\left(X_{t+1}^{\infty} | X_{-\infty}^{t} = b\right)$$

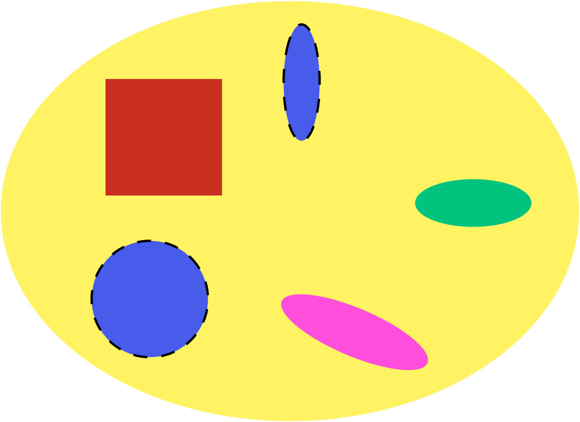
$[a] \equiv$ all histories equivalent to *a*
The statistic of interest, the **causal state**, is

$$\epsilon(x_{-\infty}^{t}) = [x_{-\infty}^{t}]$$

Set $s_t = \epsilon(x_{-\infty}^{t})$
A state is an equivalence class of histories *and* a distribution over future events
IID = 1 state, periodic = *p* states



set of histories, color-coded by conditional distribution of futures



Partitioning histories into causal states

Sufficiency

(Shalizi and Crutchfield, 2001)

$$I[X_{t+1}^{\infty}; X_{-\infty}^{t}] = I[X_{t+1}^{\infty}; \epsilon(X_{-\infty}^{t})]$$

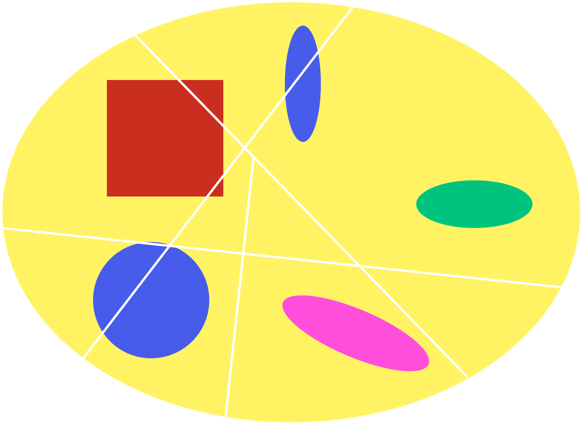
Sufficiency

(Shalizi and Crutchfield, 2001)

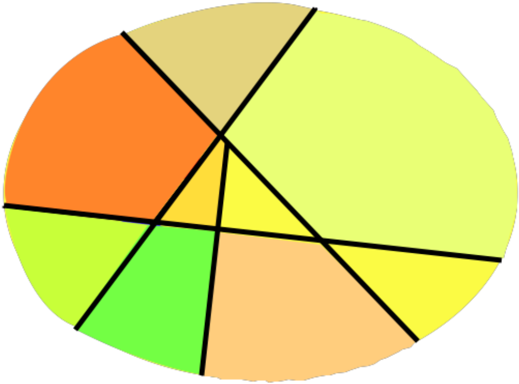
$$I[X_{t+1}^{\infty}; X_{-\infty}^{t}] = I[X_{t+1}^{\infty}; \epsilon(X_{-\infty}^{t})]$$

because

$$\Pr\left(X_{t+1}^{\infty} | S_t = \epsilon(x_{-\infty}^{t})\right) = \Pr\left(X_{t+1}^{\infty} | X_{-\infty}^{t} = x_{-\infty}^{t}\right)$$



A non-sufficient partition of histories



Effect of insufficiency on predictive distributions

Markov Properties

Sufficiency ⇒ current state screens off future from past:

$$X_{t+1}^{\infty} \perp\!\!\!\perp X_{-\infty}^{t} | S_t$$

Markov Properties

Sufficiency $\Rightarrow$ current state screens off future from past:

$$X_{t+1}^{\infty} \perp\!\!\!\perp X_{-\infty}^{t} | S_t$$

Unconditional predictions imply conditional ones $\Rightarrow$ recursive transitions for states:

$$\epsilon(x_{-\infty}^{t+1}) = T(\epsilon(x_{-\infty}^{t}), x_{t+1})$$

"Algebraically transitive" statistic

Automata theory: "deterministic transitions" (even though there are probabilities)

Markov Properties

Sufficiency $\Rightarrow$ current state screens off future from past:

$$X_{t+1}^{\infty} \perp\!\!\!\perp X_{-\infty}^{t} | S_t$$

Unconditional predictions imply conditional ones $\Rightarrow$ recursive transitions for states:

$$\epsilon(x_{-\infty}^{t+1}) = T(\epsilon(x_{-\infty}^{t}), x_{t+1})$$

"Algebraically transitive" statistic

Automata theory: "deterministic transitions" (even though there are probabilities)

$\therefore$ Causal states are Markovian:

$$S_{t+1}^{\infty} \perp\!\!\!\perp S_{-\infty}^{t-1} | S_t$$

homogeneous transition rates

Minimality

ϵ is **minimal sufficient**
$=$ can be computed from any other sufficient statistic

Minimality

ϵ is **minimal sufficient**
$=$ can be computed from any other sufficient statistic
$=$ for any sufficient η, exists a function g such that

$$\epsilon(X^t_{-\infty}) = g(\eta(X^t_{-\infty}))$$

so m.s.s. are all equivalent

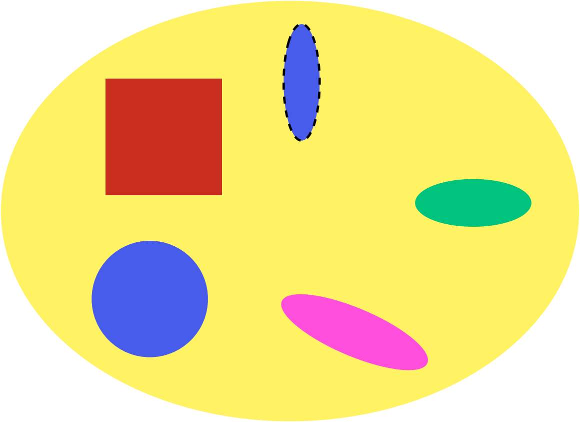
Minimality

ϵ is **minimal sufficient**
$=$ can be computed from any other sufficient statistic
$=$ for any sufficient η, exists a function g such that

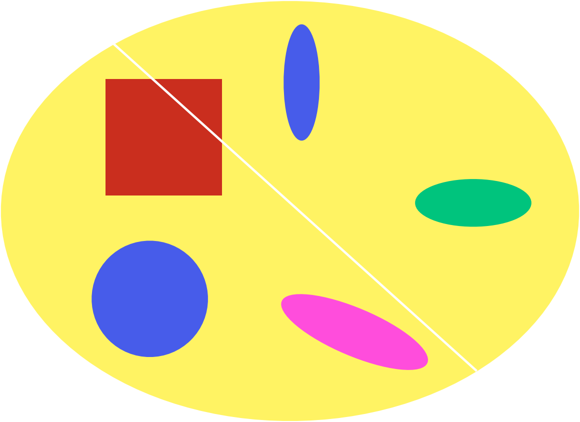
$$\epsilon(X^t_{-\infty}) = g(\eta(X^t_{-\infty}))$$

so m.s.s. are all equivalent
Therefore, if η is sufficient

$$I[\epsilon(X^t_{-\infty}); X^t_{-\infty}] \leq I[\eta(X^t_{-\infty}); X^t_{-\infty}]$$



Sufficient, but not minimal, partition of histories



Coarser than the causal states, but not sufficient

Minimal stochasticity

If $R_t = \eta(X_{-\infty}^{t-1})$ is also sufficient, then

$$H[R_{t+1}|R_t] \geq H[S_{t+1}|S_t]$$

Minimal stochasticity

If $R_t = \eta(X_{-\infty}^{t-1})$ is also sufficient, then

$$H[R_{t+1}|R_t] \geq H[S_{t+1}|S_t]$$

$\therefore$ the predictive states are the closest we get to a deterministic model, without losing power

Entropy Rate

$$
\begin{aligned}
h_1 \equiv \lim_{n\rightarrow\infty} H[X_n|X_1^{n-1}] &= \lim_{n\rightarrow\infty} H[X_n|S_n] \\
&= H[X_1|S_1]
\end{aligned}
$$

Entropy Rate

$$
\begin{aligned}
h_1 \equiv \lim_{n\to\infty} H[X_n|X_1^{n-1}] &= \lim_{n\to\infty} H[X_n|S_n] \\
&= H[X_1|S_1]
\end{aligned}
$$

so the predictive states lets us calculate the entropy rate

Entropy Rate

$$
\begin{aligned}
h_1 \equiv \lim_{n\to\infty} H[X_n | X_1^{n-1}] &= \lim_{n\to\infty} H[X_n | S_n] \\
&= H[X_1 | S_1]
\end{aligned}
$$

so the predictive states lets us calculate the entropy rate and do source coding

Minimal Markovian Representation

The observed process (X_t) is non-Markovian and ugly

Minimal Markovian Representation

The observed process (X_t) is non-Markovian and ugly
But it is generated from a homogeneous Markov process (S_t)

Minimal Markovian Representation

The observed process (X_t) is non-Markovian and ugly
But it is generated from a homogeneous Markov process (S_t)
After minimization, this representation is (essentially) unique

Minimal Markovian Representation

The observed process (X_t) is non-Markovian and ugly
But it is generated from a homogeneous Markov process (S_t)
After minimization, this representation is (essentially) unique
Can exist smaller Markovian representations, but then always have distributions over those states…

Minimal Markovian Representation

The observed process (X_t) is non-Markovian and ugly
But it is generated from a homogeneous Markov process (S_t)
After minimization, this representation is (essentially) unique
Can exist smaller Markovian representations, but then always have distributions over those states...
...and those distributions correspond to predictive states

What Sort of Markov Model?

Common-or-garden HMM:

$$S_{t+1} \perp\!\!\!\perp X_t | S_t$$

What Sort of Markov Model?

Common-or-garden HMM:

$$S_{t+1} \perp\!\!\!\perp X_t | S_t$$

But here

$$S_{t+1} = T(S_t, X_{t+1})$$

What Sort of Markov Model?

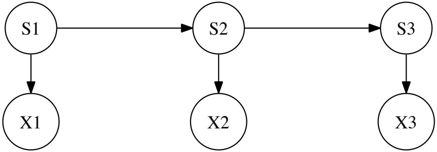
Common-or-garden HMM:

$$S_{t+1} \perp\!\!\!\perp X_t | S_t$$

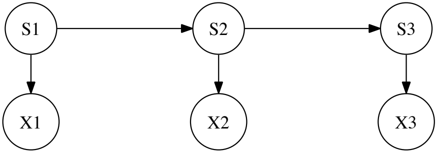
But here

$$S_{t+1} = T(S_t, X_{t+1})$$

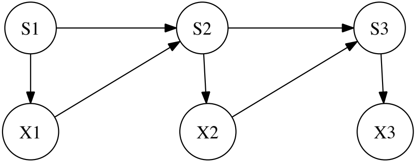
This is a **chain with complete connections** (Onicescu and Mihoc, 1935; Iosifescu and Grigorescu, 1990)



HMM

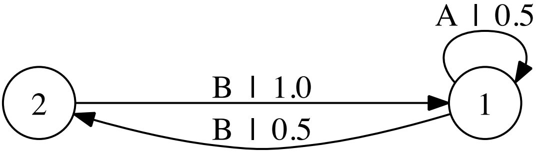


HMM

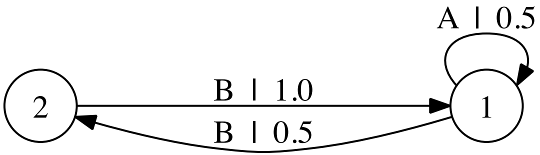


CCC

Example of a CCC: Even Process



Example of a CCC: Even Process



Blocks of As of any length, separated by even-length blocks of Bs

Not Markov at any order

Inventions

- Statistical relevance basis (Salmon, 1971, 1984)
- "Totally sufficient" statistic for a parametric family (Lauritzen, 1974, 1988)
- Measure-theoretic prediction process (Knight, 1975, 1992)
- Forecasting/true measure complexity (Grassberger, 1986)
- Causal states, ϵ machine (Crutchfield and Young, 1989)
- Observable operator model (Jaeger, 2000)
- Predictive state representations (Littman *et al.*, 2002)
- Sufficient posterior representation (Langford *et al.*, 2009)

Extension 1: Input-Output

(Littman *et al.*, 2002; Shalizi, 2001, ch. 7)

System output (X_t), input (Y_t)
Histories $x^t_{-\infty}, y^t_{-\infty}$ have distributions of output x_{t+1} for each further input y_{t+1}
Equivalence class these distributions and enforce recursive updating
Internal states of the system, not trying to predict future inputs

Extension 2: Space and Time

(Shalizi, 2003; Shalizi *et al.*, 2004, 2006; Jänicke *et al.*, 2007)

Dynamic random field $X(\vec{r}, t)$
Past cone: points in space-time which could matter to $X(\vec{r}, t)$
Future cone: points in space-time for which $X(\vec{r}, t)$ could matter

Extension 2: Space and Time

(Shalizi, 2003; Shalizi *et al.*, 2004, 2006; Jänicke *et al.*, 2007)

Dynamic random field $X(\vec{r}, t)$
Past cone: points in space-time which could matter to $X(\vec{r}, t)$
Future cone: points in space-time for which $X(\vec{r}, t)$ could matter

Extension 2: Space and Time

(Shalizi, 2003; Shalizi *et al.*, 2004, 2006; Jänicke *et al.*, 2007)

Dynamic random field $X(\vec{r}, t)$
Past cone: points in space-time which could matter to $X(\vec{r}, t)$
Future cone: points in space-time for which $X(\vec{r}, t)$ could matter

Equivalence-class past cone configurations by conditional distributions over future cones
$S(\vec{r}, t)$ is a Markov field
Minimal sufficiency, recursive updating, etc., all go through

Statistical Complexity

Definition (Grassberger, 1986; Crutchfield and Young, 1989)

$C \equiv I[\epsilon(X_{-\infty}^{t}); X_{-\infty}^{t}]$ is the **statistical forecasting complexity** of the process

Statistical Complexity

Definition (Grassberger, 1986; Crutchfield and Young, 1989)

$C \equiv I[\epsilon(X_{-\infty}^{t}); X_{-\infty}^{t}]$ is the **statistical forecasting complexity** of the process

$=$ amount of information about the past needed for optimal prediction

Statistical Complexity

Definition (Grassberger, 1986; Crutchfield and Young, 1989)

$C \equiv I[\epsilon(X^t_{-\infty}); X^t_{-\infty}]$ is the **statistical forecasting complexity** of the process

$=$ amount of information about the past needed for optimal prediction
$= H[\epsilon(X^t_{-\infty})]$ for discrete causal states

Statistical Complexity

Definition (Grassberger, 1986; Crutchfield and Young, 1989)

$C \equiv I[\epsilon(X^t_{-\infty}); X^t_{-\infty}]$ is the **statistical forecasting complexity** of the process

$=$ amount of information about the past needed for optimal prediction
$= H[\epsilon(X^t_{-\infty})]$ for discrete causal states
$=$ expected algorithmic sophistication (Gács *et al.*, 2001)

Statistical Complexity

Definition (Grassberger, 1986; Crutchfield and Young, 1989)

$C \equiv I[\epsilon(X^t_{-\infty}); X^t_{-\infty}]$ is the **statistical forecasting complexity** of the process

$=$ amount of information about the past needed for optimal prediction
$= H[\epsilon(X^t_{-\infty})]$ for discrete causal states
$=$ expected algorithmic sophistication (Gács *et al.*, 2001)
$=$ log(period) for period processes

Statistical Complexity

Definition (Grassberger, 1986; Crutchfield and Young, 1989)

$C \equiv I[\epsilon(X^t_{-\infty}); X^t_{-\infty}]$ is the **statistical forecasting complexity** of the process

$=$ amount of information about the past needed for optimal prediction
$= H[\epsilon(X^t_{-\infty})]$ for discrete causal states
$=$ expected algorithmic sophistication (Gács *et al.*, 2001)
$=$ log(period) for period processes
$=$ log(geometric mean(recurrence time)) for stationary processes

Statistical Complexity

Definition (Grassberger, 1986; Crutchfield and Young, 1989)

$C \equiv I[\epsilon(X_{-\infty}^{t}); X_{-\infty}^{t}]$ is the **statistical forecasting complexity** of the process

$=$ amount of information about the past needed for optimal prediction
$= H[\epsilon(X_{-\infty}^{t})]$ for discrete causal states
$=$ expected algorithmic sophistication (Gács *et al.*, 2001)
$=$ log(period) for period processes
$=$ log(geometric mean(recurrence time)) for stationary processes
Property *of the process*, not learning problem

Connecting to Data

Everything so far has been math/probability

Connecting to Data

Everything so far has been math/probability

(The Oracle tells us the infinite-dimensional distribution of X)

Connecting to Data

Everything so far has been math/probability

(The Oracle tells us the infinite-dimensional distribution of X)

Can we do some statistics and find the states?

Two senses of “find”: learn in a fixed model vs. discover the right model

Learning

Given states and transitions (ϵ, T), realization x_1^n
Estimate $\Pr(X_{t+1} = x | S_t = s)$

Learning

Given states and transitions (ϵ, T), realization x_1^n
Estimate $\Pr\left(X_{t+1} = x | S_t = s\right)$

- Just estimation for stochastic processes
- Easier than ordinary HMMs because S_t is a function of trajectory
- Exponential families in the all-discrete case, very tractable

Discovery

Given x_1^n
Estimate $\epsilon, T, \Pr(X_{t+1} = x | S_t = s)$

Discovery

Given x_1^n
Estimate $\epsilon, T, \Pr(X_{t+1} = x | S_t = s)$

- Inspiration: "geometry from a time series" in nonlinear dynamics
- Inspiration: PC algorithm for learning causal structure by testing conditional independence
- Function learning approach (Langford *et al.*, 2009)
- Nobody seems to have tried non-parametric Bayes

CSSR: Causal State Splitting Reconstruction

Key observation: Recursion + one-step-ahead predictive sufficiency $\Rightarrow$ general predictive sufficiency

- Get next-step distribution right by independence testing
- Then make states recursive

Assumes discrete observations, discrete time, finite causal states

Paper: Shalizi and Klinkner (2004); C++ code, `http://bactra.org/CSSR/`

One-Step Ahead Prediction

Start with all histories in the same state

One-Step Ahead Prediction

Start with all histories in the same state
Given current partition of histories into states, test whether going one step further back into the past changes the next-step conditional distribution

One-Step Ahead Prediction

Start with all histories in the same state

Given current partition of histories into states, test whether going one step further back into the past changes the next-step conditional distribution

Use a hypothesis test to control false positive rate

One-Step Ahead Prediction

Start with all histories in the same state

Given current partition of histories into states, test whether going one step further back into the past changes the next-step conditional distribution

Use a hypothesis test to control false positive rate

If yes, split that cell of the partition, but see if it matches an existing distribution

Must allow this merging or else lose minimality

If no match, add new cell to the partition

Stop when no more divisions can be made or a maximum history length Λ is reached

For consistency, $\Lambda < \frac{\log n}{h_1 + \iota}$ for some ι

Convergence

$\mathcal{S}$ = true causal state structure
$\widehat{\mathcal{S}}_n$ = structure reconstructed from n data points
Assume: finite # of states, every state has a finite history, using long enough histories, technicalities:

$$\Pr\left(\widehat{\mathcal{S}}_n \neq \mathcal{S}\right) \rightarrow 0$$

Convergence

$\mathcal{S}$ = true causal state structure
$\widehat{\mathcal{S}}_n$ = structure reconstructed from n data points
Assume: finite # of states, every state has a finite history, using long enough histories, technicalities:

$$\Pr\left(\widehat{\mathcal{S}}_n \neq \mathcal{S}\right) \rightarrow 0$$

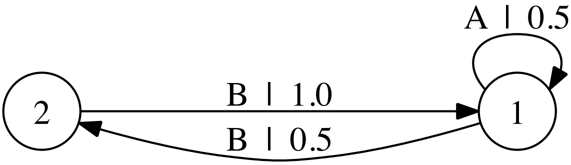
$\mathcal{D}$ = true distribution, $\widehat{\mathcal{D}}_n$ = inferred
Error scales like independent samples

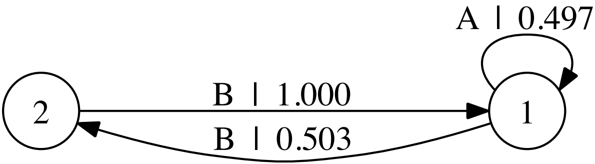
$$\mathbf{E}\left[\|\widehat{\mathcal{D}}_n - \mathcal{D}\|_{TV}\right] = O(n^{-1/2})$$

Handwaving

Empirical conditional distributions for histories converge
(large deviations principle for Markov chains)
Histories in the same state become harder to accidentally separate
Histories in different states become harder to confuse
Each state's predictive distribution converges $O(n^{-1/2})$
(from LDP again, take mixture)

Example: The Even Process





reconstruction with $\Lambda = 3$, $n = 1000$, $\alpha = 0.005$

About “Causal”

Term “causal states” introduced by Crutchfield and Young (1989)

About “Causal”

Term “causal states” introduced by Crutchfield and Young (1989) without too much consideration

About “Causal”

Term “causal states” introduced by Crutchfield and Young (1989) without too much consideration

All about probabilistic prediction, not counterfactuals

About "Causal"

Term "causal states" introduced by Crutchfield and Young (1989) without too much consideration

All about probabilistic prediction, not counterfactuals

selecting sub-ensembles of naturally-occurring trajectories vs. *enforcing* certain trajectories

About “Causal”

Term “causal states” introduced by Crutchfield and Young (1989) without too much consideration

All about probabilistic prediction, not counterfactuals

selecting sub-ensembles of naturally-occurring trajectories vs. *enforcing* certain trajectories

Still, those screening-off properties are *really suggestive*

Back to Physics

(Shalizi and Moore, 2003)

Assume: Microscopic state $Z_t \in \mathcal{Z}$, with an evolution operator f

Back to Physics

(Shalizi and Moore, 2003)

Assume: Microscopic state $Z_t \in \mathcal{Z}$, with an evolution operator f
Assume: Micro-states support counterfactuals

Back to Physics

(Shalizi and Moore, 2003)

Assume: Microscopic state $Z_t \in \mathcal{Z}$, with an evolution operator f
Assume: Micro-states support counterfactuals
Assume: Never get to see Z_t, instead deal with $X_t = \gamma(Z_t)$

Back to Physics

(Shalizi and Moore, 2003)

Assume: Microscopic state $Z_t \in \mathcal{Z}$, with an evolution operator f
Assume: Micro-states support counterfactuals
Assume: Never get to see Z_t, instead deal with $X_t = \gamma(Z_t)$
X_t are **coarse-grained**, **macroscopic** variables
Each macrovariable gives a partition Γ of $\mathcal{Z}$

Sequences of X_t values refine Γ

$$\Gamma^{(T)} = \bigwedge_{t=1}^{T} f^{-t}\Gamma$$

Sequences of X_t values refine Γ

$$\Gamma^{(T)} = \bigwedge_{t=1}^{T} f^{-t}\Gamma$$

ϵ partitions histories of X

Sequences of X_t values refine Γ

$$\Gamma^{(T)} = \bigwedge_{t=1}^{T} f^{-t}\Gamma$$

ϵ partitions histories of X
$\therefore$ ϵ joins cells of $\Gamma^{(\infty)}$

Sequences of X_t values refine Γ

$$\Gamma^{(T)} = \bigwedge_{t=1}^{T} f^{-t}\Gamma$$

ϵ partitions histories of X
$\therefore \epsilon$ joins cells of $\Gamma^{(\infty)}$
$\therefore \epsilon$ induces a partition Δ of $\mathcal{Z}$

Sequences of X_t values refine Γ

$$\Gamma^{(T)} = \bigwedge_{t=1}^{T} f^{-t}\Gamma$$

ϵ partitions histories of X
$\therefore \epsilon$ joins cells of $\Gamma^{(\infty)}$
$\therefore \epsilon$ induces a partition Δ of $\mathcal{Z}$
This is a new, Markovian coarse-grained variable

Connecting to Causality

Interventions moving z from one cell of Δ to another changes the distribution of X_{t+1}^{∞}

Connecting to Causality

Interventions moving z from one cell of Δ to another changes the distribution of X_{t+1}^{∞}

Changing z inside a cell of Δ might still make a difference

Connecting to Causality

Interventions moving z from one cell of Δ to another changes the distribution of X_{t+1}^{∞}
Changing z inside a cell of Δ might still make a difference
“There must be at least this much structure”

Summary

- Your stochastic process has a unique, minimal Markovian representation

Summary

- Your stochastic process has a unique, minimal Markovian representation
- This representation is the optimal predictor

Summary

- Your stochastic process has a unique, minimal Markovian representation
- This representation is the optimal predictor
- Can reconstruct from sample data in some cases…

Summary

- Your stochastic process has a unique, minimal Markovian representation
- This representation is the optimal predictor
- Can reconstruct from sample data in some cases…
  and a lot more could be done in this line

Summary

- Your stochastic process has a unique, minimal Markovian representation
- This representation is the optimal predictor
- Can reconstruct from sample data in some cases… and a lot more could be done in this line
- The Markov states have the right screening-off properties for causal models

How Broad Are These Results?

Knight (1975, 1992) gave most general constructions

- Non-stationary X
- t continuous (but discrete works as special case)
- X_t with values in a Lusin space (= image of a complete separable metrizable space under a measurable bijection)
- S_t is a homogeneous strong Markov process with deterministic updating
- S_t has cadlag sample paths (in appropriate topologies on infinite-dimensional distributions)

A Cousin: The Information Bottleneck

(Tishby *et al.*, 1999)

For inputs X and outputs Y, fix $\beta > 0$, find $\eta(X)$, the **bottleneck variable**, maximizing

$$I[\eta(X); Y] - \beta I[\eta(X); X]$$

give up 1 bit of predictive information for β bits of memory
Predictive sufficiency comes as $\beta \to \infty$, unwilling to lose *any* predictive power

The “I’m Glad You Asked That Question” Slides

"Geometry from a Time Series"

Deterministic dynamical system with state z_t on a smooth manifold of dimension m, $z_{t+1} = f(z_t)$

Only identified up to a smooth, invertible change of coordinates (diffeomorphism)

Observe a time series of a single smooth, instantaneous function of state $x_t = g(z_t)$

Set $s_t = (x_t, x_{t-1}, \ldots x_{t-k+1})$

Generically, if $k \geq 2m + 1$, then $z_t = \phi(s_t)$

ϕ is smooth and invertible

ϕ commutes with time evolution, $\phi(s_{t+1}) = f(\phi(s_t))$

Regressing s_{t+1} on s_t gives $\phi^{-1} \circ f$

Idea due to Packard *et al.* (1980); Takens (1981), modern review in Kantz and Schreiber (2004)

Ensuring Recursive Transitions

Need to determinize a probabilistic automaton
Several ways of doing this; technical and not worth going into here
Trickiest part of the algorithm and can influence the finite-sample behavior

Some Uses

Geomagnetic fluctuations (Clarke *et al.*, 2003)
Visualization for hydrodynamics and climate modeling (Jänicke *et al.*, 2007)
Natural language processing (Padró and Padró, 2005a,c,b, 2007a,b)
Anomaly detection (Friedlander *et al.*, 2003a,b; Ray, 2004)
Information sharing in networks (Klinkner *et al.*, 2006; Shalizi *et al.*, 2007)
Social media propagation (Cointet *et al.*, 2007)
Neural spike train analysis (Haslinger *et al.*, 2010)

Clarke, Richard W., Mervyn P. Freeman and Nicholas W. Watkins (2003). "Application of Computational Mechanics to the Analysis of Natural Data: An Example in Geomagnetism." *Physical Review E*, **67**: 0126203. URL http://arxiv.org/abs/cond-mat/0110228.

Cointet, Jean-Philippe, Emmanuel Faure and Camille Roth (2007). "Intertemporal topic correlations in online media." In *Proceedings of the International Conference on Weblogs and Social Media [ICWSM]*. Boulder, CO, USA. URL http://camille.roth.free.fr/travaux/cointetfaureroth-icwsm-cr4p.pdf.

Crutchfield, James P. and Karl Young (1989). "Inferring Statistical Complexity." *Physical Review Letters*, **63**: 105–108. URL http://www.santafe.edu/~cmg/compmech/pubs/ISCTitlePage.htm.

Friedlander, David S., Shashi Phoha and Richard Brooks (2003a). "Determination of Vehicle Behavior based on Distributed Sensor Network Data." In *Advanced Signal Processing Algorithms, Architectures, and Implementations XIII* (Franklin T. Luk, ed.), vol. 5205 of *Proceedings of the SPIE*. Bellingham, WA: SPIE. Presented at SPIE's 48th Annual Meeting, 3–8 August 2003, San Diego, CA.

Friedlander, Davis S., Isanu Chattopadhayay, Asok Ray, Shashi Phoha and Noah Jacobson (2003b). "Anomaly Prediction in Mechanical System Using Symbolic Dynamics." In *Proceedings of the 2003 American Control Conference, Denver, CO, 4–6 June 2003*.

Gács, Péter, John T. Tromp and Paul M. B. Vitanyi (2001). "Algorithmic Statistics." *IEEE Transactions on Information Theory*, **47**: 2443–2463. URL http://arxiv.org/abs/math.PR/0006233.

Grassberger, Peter (1986). "Toward a Quantitative Theory of Self-Generated Complexity." *International Journal of Theoretical Physics*, **25**: 907–938.

Haslinger, Robert, Kristina Lisa Klinkner and Cosma Rohilla Shalizi (2010). "The Computational Structure of Spike Trains." *Neural Computation*, **22**: 121–157. URL http://arxiv.org/abs/1001.0036. doi:10.1162/neco.2009.12-07-678.

Iosifescu, Marius and Serban Grigorescu (1990). *Dependence with Complete Connections and Its Applications*. Cambridge, England: Cambridge University Press. Revised paperback printing, 2009.

Jaeger, Herbert (2000). "Observable Operator Models for Discrete Stochastic Time Series." *Neural Computation*, **12**:

1371–1398. URL http://www.faculty.iu-bremen.de/hjaeger/pubs/oom_neco00.pdf.

Jänicke, Heike, Alexander Wiebel, Gerik Scheuermann and Wolfgang Kollmann (2007). "Multifield Visualization Using Local Statistical Complexity." *IEEE Transactions on Visualization and Computer Graphics*, **13**: 1384–1391. URL http://www.informatik.uni-leipzig.de/bsv/Jaenicke/Papers/vis07.pdf. doi:10.1109/TVCG.2007.70615.

Kantz, Holger and Thomas Schreiber (2004). *Nonlinear Time Series Analysis*. Cambridge, England: Cambridge University Press, 2nd edn.

Klinkner, Kristina Lisa, Cosma Rohilla Shalizi and Marcelo F. Camperi (2006). "Measuring Shared Information and Coordinated Activity in Neuronal Networks." In *Advances in*

Neural Information Processing Systems 18 (NIPS 2005) (Yair Weiss and Bernhard Schölkopf and John C. Platt, eds.), pp. 667–674. Cambridge, Massachusetts: MIT Press. URL http://arxiv.org/abs/q-bio.NC/0506009.

Knight, Frank B. (1975). "A Predictive View of Continuous Time Processes." *Annals of Probability*, **3**: 573–596. URL http://projecteuclid.org/euclid.aop/1176996302.

— (1992). *Foundations of the Prediction Process*. Oxford: Clarendon Press.

Langford, John, Ruslan Salakhutdinov and Tong Zhang (2009). "Learning Nonlinear Dynamic Models." Electronic preprint. URL http://arxiv.org/abs/0905.3369.

Lauritzen, Steffen L. (1974). "Sufficiency, Prediction and Extreme Models." *Scandinavian Journal of Statistics*, **1**: 128–134. URL http://www.jstor.org/pss/4615564.

— (1988). *Extremal Families and Systems of Sufficient Statistics*. Berlin: Springer-Verlag.

Littman, Michael L., Richard S. Sutton and Satinder Singh (2002). "Predictive Representations of State." In *Advances in Neural Information Processing Systems 14 (NIPS 2001)* (Thomas G. Dietterich and Suzanna Becker and Zoubin Ghahramani, eds.), pp. 1555–1561. Cambridge, Massachusetts: MIT Press. URL `http://www.eecs.umich.edu/~baveja/Papers/psr.pdf`.

Onicescu, Octav and Gheorghe Mihoc (1935). "Sur les chaînes de variables statistiques." *Comptes Rendus de l'Académie des Sciences de Paris*, **200**: 511–512.

Packard, Norman H., James P. Crutchfield, J. Doyne Farmer and Robert S. Shaw (1980). "Geometry from a Time Series." *Physical Review Letters*, **45**: 712–716.

Padró, Muntsa and Lluís Padró (2005a). "Applying a Finite Automata Acquisition Algorithm to Named Entity Recognition." In *Proceedings of 5th International Workshop on Finite-State Methods and Natural Language Processing (FSMNLP'05)*. URL http://www.lsi.upc.edu/~nlp/papers/2005/fsmnlp05-pp.pdf.

— (2005b). "Approaching Sequential NLP Tasks with an Automata Acquisition Algorithm." In *Proceedings of International Conference on Recent Advances in NLP (RANLP'05)*. URL http://www.lsi.upc.edu/~nlp/papers/2005/ranlp05-pp.pdf.

— (2005c). "A Named Entity Recognition System Based on a Finite Automata Acquisition Algorithm." *Procesamiento del Lenguaje Natural*, **35**: 319–326. URL http://www.lsi.upc.edu/~nlp/papers/2005/sepln05-pp.pdf.

— (2007a). "ME-CSSR: an Extension of CSSR using Maximum Entropy Models." In *Proceedings of Finite State Methods for Natural Language Processing (FSMNLP) 2007*. URL `http://www.lsi.upc.edu/%7Enlp/papers/2007/fsmnlp07-pp.pdf`.

— (2007b). "Studying CSSR Algorithm Applicability on NLP Tasks." *Procesamiento del Lenguaje Natural*, **39**: 89–96. URL `http://www.lsi.upc.edu/%7Enlp/papers/2007/sepln07-pp.pdf`.

Ray, Asok (2004). "Symbolic dynamic analysis of complex systems for anomaly detection." *Signal Processing*, **84**: 1115–1130.

Salmon, Wesley C. (1971). *Statistical Explanation and Statistical Relevance*. Pittsburgh: University of Pittsburgh

Press. With contributions by Richard C. Jeffrey and James G. Greeno.

— (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Shalizi, Cosma Rohilla (2001). *Causal Architecture, Complexity and Self-Organization in Time Series and Cellular Automata*. Ph.D. thesis, University of Wisconsin-Madison. URL `http://bactra.org/thesis/`.

— (2003). "Optimal Nonlinear Prediction of Random Fields on Networks." *Discrete Mathematics and Theoretical Computer Science*, **AB(DMCS)**: 11–30. URL `http://arxiv.org/abs/math.PR/0305160`.

Shalizi, Cosma Rohilla, Marcelo F. Camperi and Kristina Lisa Klinkner (2007). "Discovering Functional Communities in Dynamical Networks." In *Statistical Network Analysis:*

Models, Issues, and New Directions (Edo Airoldi and David M. Blei and Stephen E. Fienberg and Anna Goldenberg and Eric P. Xing and Alice X. Zheng, eds.), vol. 4503 of *Lecture Notes in Computer Science*, pp. 140–157. New York: Springer-Verlag. URL http://arxiv.org/abs/q-bio.NC/0609008.

Shalizi, Cosma Rohilla and James P. Crutchfield (2001). "Computational Mechanics: Pattern and Prediction, Structure and Simplicity." *Journal of Statistical Physics*, **104**: 817–879. URL http://arxiv.org/abs/cond-mat/9907176.

Shalizi, Cosma Rohilla, Robert Haslinger, Jean-Baptiste Rouquier, Kristina Lisa Klinkner and Cristopher Moore (2006). "Automatic Filters for the Detection of Coherent Structure in Spatiotemporal Systems." *Physical Review E*, **73**: 036104. URL http://arxiv.org/abs/nlin.CG/0508001.

Shalizi, Cosma Rohilla and Kristina Lisa Klinkner (2004). "Blind Construction of Optimal Nonlinear Recursive Predictors for Discrete Sequences." In *Uncertainty in Artificial Intelligence: Proceedings of the Twentieth Conference (UAI 2004)* (Max Chickering and Joseph Y. Halpern, eds.), pp. 504–511. Arlington, Virginia: AUAI Press. URL http://arxiv.org/abs/cs.LG/0406011.

Shalizi, Cosma Rohilla, Kristina Lisa Klinkner and Robert Haslinger (2004). "Quantifying Self-Organization with Optimal Predictors." *Physical Review Letters*, **93**: 118701. URL http://arxiv.org/abs/nlin.AO/0409024.

Shalizi, Cosma Rohilla and Cristopher Moore (2003). "What Is a Macrostate? From Subjective Measurements to Objective

Dynamics." Electronic pre-print. URL http://arxiv.org/abs/cond-mat/0303625.

Takens, Floris (1981). "Detecting Strange Attractors in Fluid Turbulence." In *Symposium on Dynamical Systems and Turbulence* (D. A. Rand and L. S. Young, eds.), pp. 366–381. Berlin: Springer-Verlag.

Tishby, Naftali, Fernando C. Pereira and William Bialek (1999). "The Information Bottleneck Method." In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing* (B. Hajek and R. S. Sreenivas, eds.), pp. 368–377. Urbana, Illinois: University of Illinois Press. URL http://arxiv.org/abs/physics/0004057.