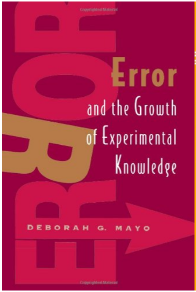
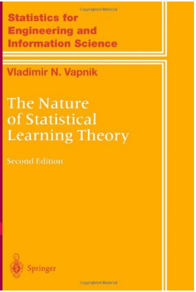
Complexity, Prediction, and Inference

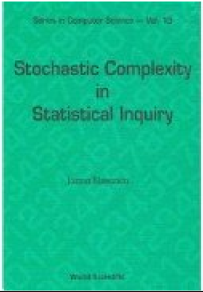
Cosma Shalizi

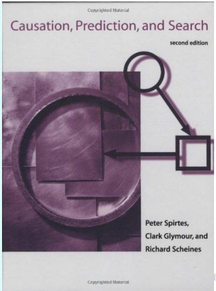
Statistics Dept., Carnegie Mellon University & Santa Fe Institute

23 June 2012
Ockham Workshop


Error
and the Growth
of Experimental
Knowledge
DEBORAH G. MAYO


Statistics for Engineering and Information Science
Vladimir N. Vapnik
The Nature of Statistical Learning Theory
Second Edition
Springer


Stochastic Complexity
in
Statistical Inquiry


Causation, Prediction, and Search
second edition
Peter Spirtes,
Clark Glymour, and
Richard Scheines

What Is Statistics, and How Does It Help Scientists?

Computer science, operations research, statistics, etc. as **mathematical engineering**

What Is Statistics, and How Does It Help Scientists?

Computer science, operations research, statistics, etc. as **mathematical engineering**

Statistics: design and analyze methods of inference from imperfect data

What Is Statistics, and How Does It Help Scientists?

Computer science, operations research, statistics, etc. as **mathematical engineering**

Statistics: design and analyze methods of inference from imperfect data

ML: design and analyze methods of automatic prediction

Not the same, but not totally alien either

Classical Statistics

Applied Statistics

Scientist (or brewer, etc.): has a concrete inferential problem about the world, plus data
Statistician: builds an abstract machine to turn data into an answer, with honesty about uncertainty

Classical Statistics

Applied Statistics

Scientist (or brewer, etc.): has a concrete inferential problem about the world, plus data
Statistician: builds an abstract machine to turn data into an answer, with honesty about uncertainty

Theoretical Statistics

Advice to applied statisticians about what tools work when

What Statisticians Care About

"Will this method be reliable enough to be useful?"

What Statisticians Care About

"Will this method be reliable enough to be useful?"
Articulated: accuracy, precision, error rates, rate of convergence, quantification of uncertainty through confidence ("how *unlucky* would we have to be to be wrong?"), bias-variance trade-offs, data reductions ("statistics", sufficiency, necessity, . . .), identification, residual diagnostics, . . .

Why Classical Statistics Used to Be So Boring

A very general theory of inference…

Why Classical Statistics Used to Be So Boring

A very general theory of inference...
...and powerful methods it applied to (non-parametric regression, non-parametric density estimation) ...

Why Classical Statistics Used to Be So Boring

A very general theory of inference...
...and powerful methods it applied to (non-parametric regression, non-parametric density estimation) ...
...and yet used hardly any of it

Why Classical Statistics Used to Be So Boring

A very general theory of inference...
...and powerful methods it applied to (non-parametric regression, non-parametric density estimation) ...
...and yet used hardly any of it
Data was hard, expensive and slow

Why Classical Statistics Used to Be So Boring

A very general theory of inference. . .
. . . and powerful methods it applied to (non-parametric regression, non-parametric density estimation) . . .
. . . and yet used hardly any of it
Data was hard, expensive and slow
Calculations were hard, expensive and slow

Why Classical Statistics Used to Be So Boring

A very general theory of inference. . .
. . . and powerful methods it applied to (non-parametric regression, non-parametric density estimation) . . .
. . . and yet used hardly any of it
Data was hard, expensive and slow
Calculations were hard, expensive and slow
$\therefore$ low-dimensional data
$+$ low-dimensional parametric models
$+$ modeling assumptions to short-cut long calculations
$\therefore$ boring

Why Classical Statistics Used to Be So Boring

A very general theory of inference...

...and powerful methods it applied to (non-parametric regression, non-parametric density estimation) ...

...and yet used hardly any of it

Data was hard, expensive and slow

Calculations were hard, expensive and slow

$\therefore$ low-dimensional data

$+$ low-dimensional parametric models

$+$ modeling assumptions to short-cut long calculations

$\therefore$ boring

Computing was the binding constraint

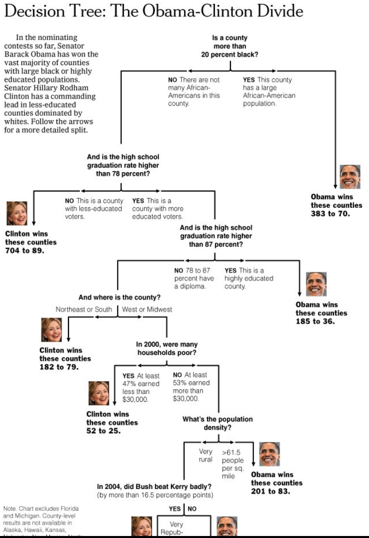
How Computing Saved Statistics

Computation became easy, cheap and fast

How Computing Saved Statistics

Computation became easy, cheap and fast
$\therefore$ fit and use non-parametric (but interpretable) models: splines, kernels, CART...
$+$ evaluate models with sub-sampling (cross-validation)
$+$ find uncertainty with re-sampling (bootstrap)
$+$ model-building by penalized optimization (lasso etc.)
$+$ model-discovery by constraint satisfaction (PC, FCI, etc.)
$+$ simulation-based inference

Putting the CART before the Horse Race



Mis-Specification

Good estimator + well-specified model $\Rightarrow$ converge to truth
Good estimator + mis-specified model $\Rightarrow$ converge to *closest approximation* to truth ("pseudo-truth")
(even true with Bayesian inference)
e.g., additive regression converges to the best additive approximation
this may or may not be a problem for *scientific* inference

Parsimony?

Computation is always a consideration
Restrictions (dimensions, penalties, . . .) help convergence
Do we want to converge quickly to the wrong answer?

Parsimony?

Computation is always a consideration

Restrictions (dimensions, penalties, . . .) help convergence

Do we want to converge quickly to the wrong answer?

Parsimony for scientists is more about mechanisms than fixing parameters or imposing linearity

Parsimony?

Computation is always a consideration
Restrictions (dimensions, penalties, ...) help convergence
Do we want to converge quickly to the wrong answer?
Parsimony for scientists is more about mechanisms than fixing parameters or imposing linearity
Let's try to articulate *system* complexity

The Guiding Idea

The behavior of complex systems is hard to describe

The Guiding Idea

The behavior of complex systems is hard to describe

. . . even if you know what you're doing

The Guiding Idea

The behavior of complex systems is hard to describe
. . . even if you know what you're doing
von Neumann: a cat is complex because it has no model simpler than the cat itself

The Guiding Idea

The behavior of complex systems is hard to describe
. . . even if you know what you're doing
von Neumann: a cat is complex because it has no model simpler than the cat itself
Complexity $\approx$ resources needed for optimal description or prediction

Three Kinds of Complexity

1. *Prediction* of the system, in the optimal model (units: bits)
   Wiener, von Neumann, Kolmogorov, Pagels and Lloyd, . . .

Three Kinds of Complexity

1. *Prediction* of the system, in the optimal model (units: bits)
   Wiener, von Neumann, Kolmogorov, Pagels and Lloyd, ...
2. *Learning* that model (units: samples)
   Fisher, Vapnik and Chervonenkis, Valiant, ...

Three Kinds of Complexity

1. *Prediction* of the system, in the optimal model (units: bits)
   Wiener, von Neumann, Kolmogorov, Pagels and Lloyd, . . .
2. *Learning* that model (units: samples)
   Fisher, Vapnik and Chervonenkis, Valiant, . . .
3. *Computational* complexity of running the model (units: ops)

Three Kinds of Complexity

1. *Prediction* of the system, in the optimal model (units: bits)
   Wiener, von Neumann, Kolmogorov, Pagels and Lloyd, ...
2. *Learning* that model (units: samples)
   Fisher, Vapnik and Chervonenkis, Valiant, ...
3. *Computational* complexity of running the model (units: ops)

Stick to predicting

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 optional

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 $X_{-\infty}^{t}$
What's a good statistic to use?

Predictive Sufficiency

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

Why Care About Sufficiency?

Why Care About Sufficiency?

Optimal strategy, under any loss function, only needs a sufficient statistic (Blackwell & Girshick)

Why Care About Sufficiency?

Optimal strategy, under any loss function, only needs a sufficient statistic (Blackwell & Girshick)
Strategies using insufficient statistics can generally be improved (Blackwell & Rao)

Why Care About Sufficiency?

Optimal strategy, under any loss function, only needs a sufficient statistic (Blackwell & Girshick)

Strategies using insufficient statistics can generally be improved (Blackwell & Rao)

$\therefore$ Don't worry 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 $s_t = \epsilon(x_{-\infty}^{t-1})$

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

"Causal" States

(Crutchfield and Young, 1989)

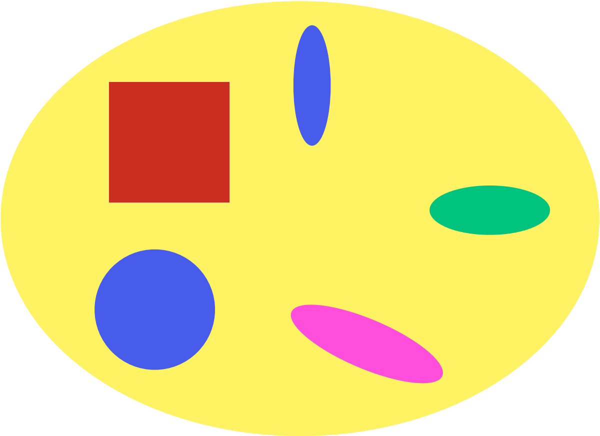
Histories a and b are equivalent iff

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

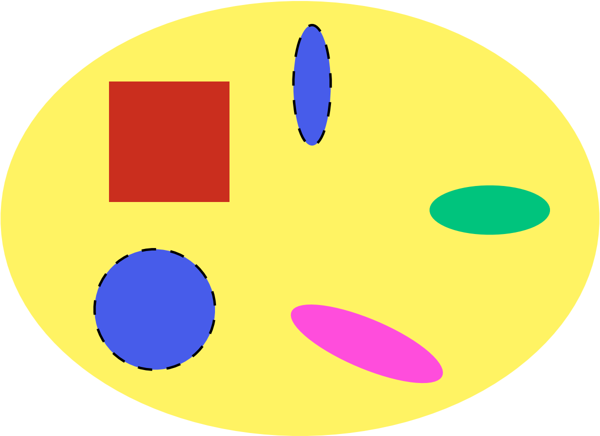
$[a] \equiv$ all histories equivalent to a
The statistic of interest, the **causal state**, is

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

Set $s_t = \epsilon(x_{-\infty}^{t-1})$
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)

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

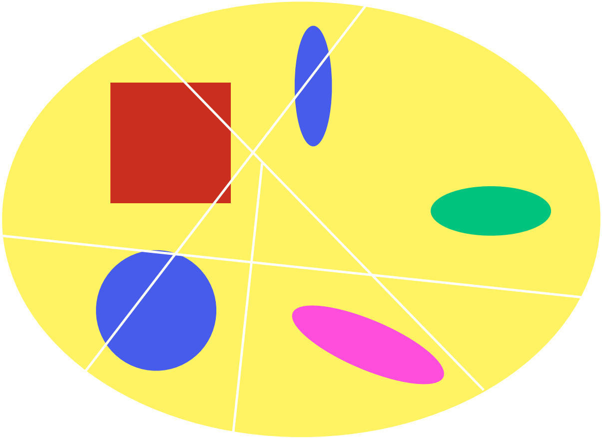
Sufficiency

(Shalizi and Crutchfield, 2001)

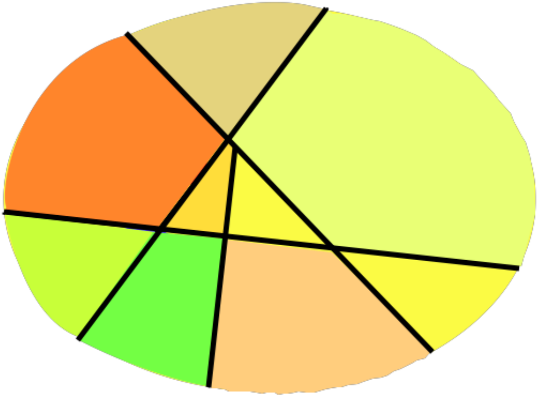
$$I[X_{t+1}^{\infty}; X_{-\infty}^{t}] = I[X_{t+1}^{\infty}; \epsilon(X_{-\infty}^{t})]$$

because

$$\begin{aligned} & \Pr\left(X_{t+1}^{\infty} | S_t = \epsilon(x_{-\infty}^{t})\right) \\ = & \int_{y \in [x_{-\infty}^{t}]} \Pr\left(X_{t+1}^{\infty} | X_{-\infty}^{t} = y\right) \Pr\left(X_{-\infty}^{t} = y | S_t = \epsilon(x_{-\infty}^{t})\right) dy \\ = & \Pr\left(X_{t+1}^{\infty} | X_{-\infty}^{t} = x_{-\infty}^{t}\right) \end{aligned}$$



A non-sufficient partition of histories



Effect of insufficiency on predictive distributions

Group x and y together when they have the same *consequences*
not when they have the same *appearance*
"Lebesgue smoothing" instead of "Riemann smoothing"
Learn the predictive geometry, not the original geometry

Markov Properties

Future observations are independent of the past given the causal state:

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

Markov Properties

Future observations are independent of the past given the causal state:

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

by sufficiency:

$$\begin{aligned} & \Pr\left(X_{t+1}^{\infty} | X_{-\infty}^{t} = x_{-\infty}^{t}, S_{t+1} = \epsilon(x_{-\infty}^{t})\right) \\ = & \Pr\left(X_{t+1}^{\infty} | X_{-\infty}^{t} = x_{-\infty}^{t}\right) \\ = & \Pr\left(X_{t+1}^{\infty} | S_{t+1} = \epsilon(x_{-\infty}^{t})\right) \end{aligned}$$

Recursive Updating/Deterministic Transitions

Recursive transitions for states:

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

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

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

Causal States are Markovian

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

Causal States are Markovian

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

because

$$S_{t+1}^{\infty} = T(S_t, X_t^{\infty})$$

Causal States are Markovian

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

because

$$S_{t+1}^{\infty} = T(S_t, X_t^{\infty})$$

and

$$X_t^{\infty} \perp\!\!\!\perp \left\{X_{-\infty}^{t-1}, S_{-\infty}^{t-1}\right\} | S_t$$

Causal States are Markovian

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

because

$$S_{t+1}^{\infty} = T(S_t, X_t^{\infty})$$

and

$$X_t^{\infty} \perp\!\!\!\perp \left\{ X_{-\infty}^{t-1}, S_{-\infty}^{t-1} \right\} | S_t$$

Also, the transitions are homogeneous

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

$$\epsilon(X_{-\infty}^{t}) = g(\eta(X_{-\infty}^{t}))$$

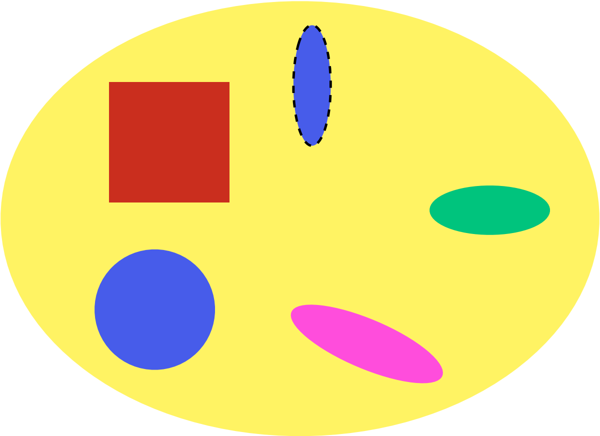
Minimality

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

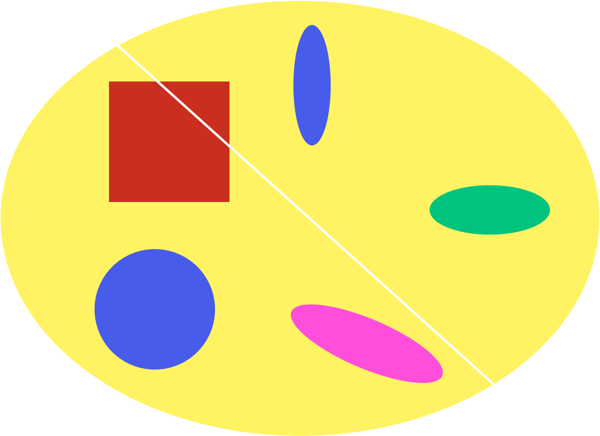
$$\epsilon(X_{-\infty}^{t}) = g(\eta(X_{-\infty}^{t}))$$

Therefore, if η is sufficient

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



Sufficient, but not minimal, partition of histories



Coarser than the causal states, but not sufficient

Uniqueness

There is really no other minimal sufficient statistic

Uniqueness

There is really no other minimal sufficient statistic
If η is minimal, there is an h such that

$$\eta = h(\epsilon) \text{ a.s.}$$

Uniqueness

There is really no other minimal sufficient statistic
If η is minimal, there is an *h* such that

$$\eta = h(\epsilon) \text{ a.s.}$$

but $\epsilon = g(\eta)$ (a.s.)

Uniqueness

There is really no other minimal sufficient statistic
If η is minimal, there is an h such that

$$\eta = h(\epsilon) \text{ a.s.}$$

but $\epsilon = g(\eta)$ (a.s.) so

$$\begin{aligned} g(h(\epsilon)) &= \epsilon \\ h(g(\eta)) &= \eta \end{aligned}$$

ϵ and η partition histories in the same way (a.s.)

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

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

What Sort of Markov Model?

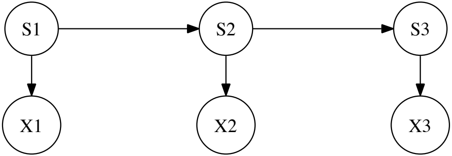
Common-or-garden HMM:

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

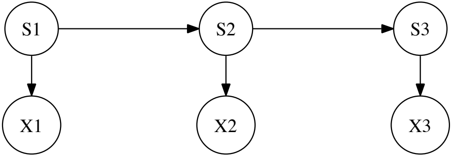
But here

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

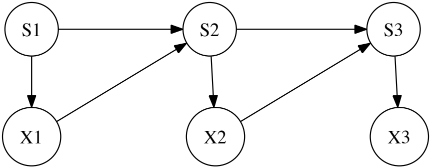
This is a **chain with complete connections** (Onicescu and Mihoc, 1935; Iosifescu and Grigorescu, 1990)



HMM

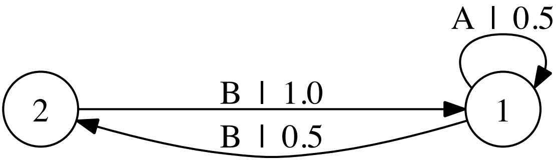


HMM

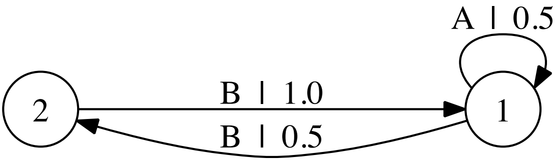


CCC

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)

Inventions

- Statistical relevance basis (Salmon, 1971, 1984)
- Measure-theoretic prediction process (Knight, 1975, 1992)

Inventions

- Statistical relevance basis (Salmon, 1971, 1984)
- Measure-theoretic prediction process (Knight, 1975, 1992)
- Forecasting/true measure complexity (Grassberger, 1986)

Inventions

- Statistical relevance basis (Salmon, 1971, 1984)
- Measure-theoretic prediction process (Knight, 1975, 1992)
- Forecasting/true measure complexity (Grassberger, 1986)
- Causal states, ϵ machine (Crutchfield and Young, 1989)

Inventions

- Statistical relevance basis (Salmon, 1971, 1984)
- 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)

Inventions

- Statistical relevance basis (Salmon, 1971, 1984)
- 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)

Inventions

- Statistical relevance basis (Salmon, 1971, 1984)
- 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)

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

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)

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 some topology on infinite-dimensional distributions)

Versions for input-output systems, spatial and network dynamics (Shalizi, 2001, 2003; Shalizi *et al.*, 2004)

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)

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

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

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

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
0 for IID sources
$\log p$ for periodic sources

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

$= H[\epsilon(X^t_{-\infty})]$ for discrete causal states

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

$= H[\epsilon(X^t_{-\infty})]$ for discrete causal states
$=$ expected algorithmic sophistication (Gács *et al.*, 2001)

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

$= H[\epsilon(X^t_{-\infty})]$ for discrete causal states
$=$ expected algorithmic sophistication (Gács *et al.*, 2001)
$=$ log(geometric mean(recurrence time)) for stationary processes

Predictive Information

Predictive information:

$$I_{\mathrm{pred}} \equiv I[X_{t+1}^{\infty}; X_{-\infty}^{t}]$$

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

Predictive Information

Predictive information:

$$I_{\mathrm{pred}} \equiv I[X_{t+1}^{\infty}; X_{-\infty}^{t}]$$

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

You need at least *m* bits of state to get *m* bits of prediction

More on the Statistical Complexity

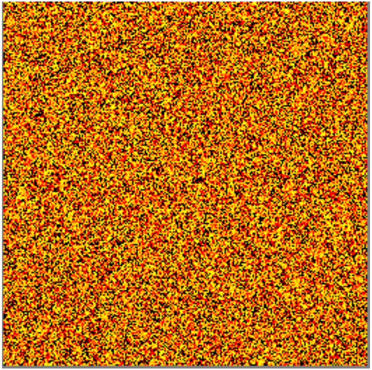
Property *of the process*, not learning problem

More on the Statistical Complexity

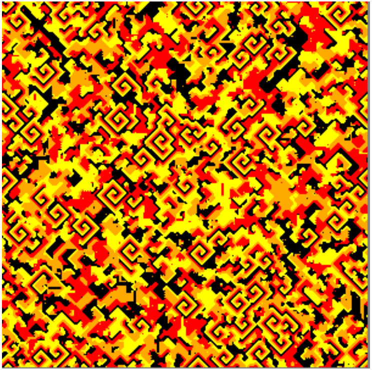
Property *of the process*, not learning problem
How much structure do we absolutely need to posit?

More on the Statistical Complexity

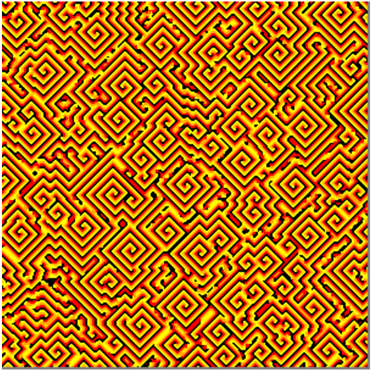
Property *of the process*, not learning problem
How much structure do we absolutely need to posit?
Relative to level of description/coarse-graining
thermodynamic vs. hydrodynamic vs. molecular description. . .
C = information about microstate in macrostate (sometimes; Shalizi and Moore (2003))



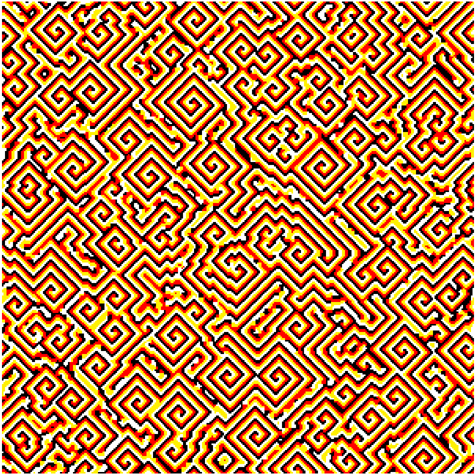
Initial configuration



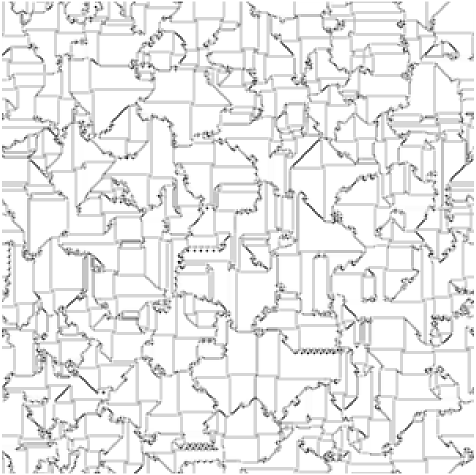
Intermediate time configuration



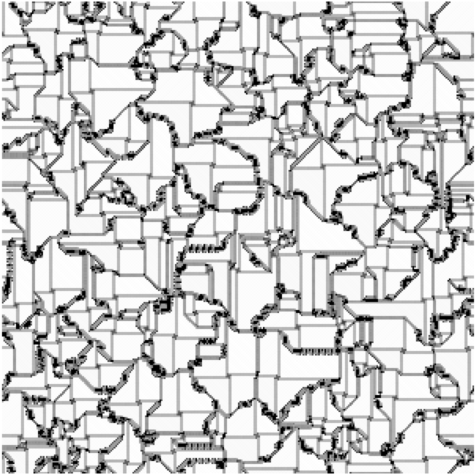
Asymptotic configuration, rotating spirals



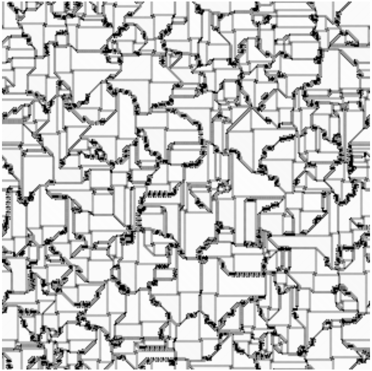
Typical long-time configuration



Hand-crafted order parameter field



Local complexity field



Order parameter (broken symmetry, physical insight, tradition, trial and error, current configuration) vs. local statistical complexity (prediction, automatic, time evolution) (Shalizi *et al.*, 2006)

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 $\Pr\left(X_{t+1} = x | S_t = s\right)$

Learning

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

- 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: PC algorithm for learning graphical models by testing conditional independence
- Alternative: Function learning approach (Langford *et al.*, 2009)
- Nobody seems to have tried non-parametric Bayes (though (Pfau *et al.*, 2010) is a step in that direction)

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 hold false positive rate at α

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 hold false positive rate at α

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

Must allow this merging or else no minimality

If no match, add new cell to the partition

Recursive Transitions

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

For consistency, $\Lambda < \frac{\log n}{h+\iota}$ for some ι (Marton and Shields, 1994)

Recursive Transitions

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

For consistency, $\Lambda < \frac{\log n}{h+\iota}$ for some ι (Marton and Shields, 1994)

Ensure recursive transitions

Equivalent to: determinize a non-deterministic stochastic automaton

Recursive Transitions

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

For consistency, $\Lambda < \frac{\log n}{h+\iota}$ for some ι (Marton and Shields, 1994)

Ensure recursive transitions

Equivalent to: determinize a non-deterministic stochastic automaton

technical; boring; can influence finite-sample behavior

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, $\alpha \rightarrow 0$ slowly:

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, $\alpha \rightarrow 0$ slowly:

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

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, $\alpha \rightarrow 0$ slowly:

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

Empirical conditional distributions for histories converge

(large deviations principle for Markov chains)

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, $\alpha \rightarrow 0$ slowly:

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

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

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

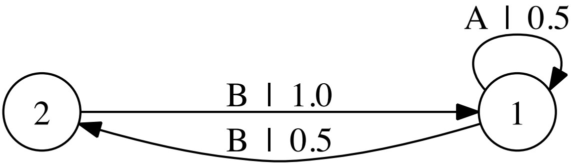
$$\mathbf{E}\left[\|\widehat{\mathcal{D}}_n - \mathcal{D}\|_{TV}\right] = O(n^{-1/2})$$

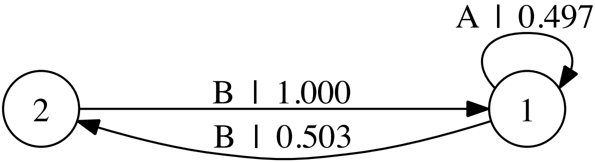
$\mathcal{D} =$ true distribution, $\widehat{\mathcal{D}}_n$ = inferred
Error scales like independent samples

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

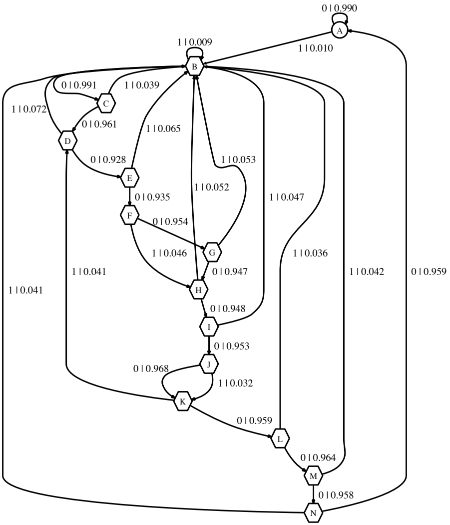
Each state's predictive distribution converges $O(n^{-1/2})$
(from LDP again, take mixture)

Example: The Even Process





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



Causal states reconstructed from rat barrel cortex neuron during spontaneous firing; state *A* is the resting state, the rest “implement” a combination of decaying firing rate and refractory periods (Haslinger *et al.*, 2010)

Occam?

CSSR: start with a small model, expand when forced to
Seems to converge faster than state-merging algorithms
Is this Occam? Should we care?

Summary

- Your stochastic process has a unique, minimal Markovian representation

Summary

- Your stochastic process has a unique, minimal Markovian representation
- This representation has nice predictive properties

Summary

- Your stochastic process has a unique, minimal Markovian representation
- This representation has nice predictive properties
- Can reconstruct from sample data in some cases...

Summary

- Your stochastic process has a unique, minimal Markovian representation
- This representation has nice predictive properties
- 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 has nice predictive properties
- Can reconstruct from sample data in some cases. . .
  and a lot more could be done in this line
- Both the representation and the reconstruction have an Occam flavor

I'm Glad You Asked That Question!

If $u \sim v$, any future event F, and single observation a

$$\begin{aligned}
\Pr\left(X_{t+1}^{\infty} \in aF | X_{-\infty}^{t} = u\right) &= \Pr\left(X_{t+1}^{\infty} \in aF | X_{-\infty}^{t} = v\right) \\
\Pr\left(X_{t+1} = a, X_{t+2}^{\infty} \in F | X_{-\infty}^{t} = u\right) &= \Pr\left(X_{t+1} = a, X_{t+2}^{\infty} \in F | X_{-\infty}^{t} = v\right)
\end{aligned}$$

$$\begin{aligned}
& \Pr\left(X_{t+2}^{\infty} \in F | X_{-\infty}^{t+1} = ua\right) \Pr\left(X_{t+1} = a | X_{-\infty}^{t} = u\right) \\
&= \Pr\left(X_{t+2}^{\infty} \in F | X_{-\infty}^{t+1} = va\right) \Pr\left(X_{t+1} = a | X_{-\infty}^{t} = v\right)
\end{aligned}$$

$$\begin{aligned}
\Pr\left(X_{t+2}^{\infty} \in F | X_{-\infty}^{t+1} = ua\right) &= \Pr\left(X_{t+2}^{\infty} \in F | X_{-\infty}^{t+1} = va\right) \\
ua &\sim va
\end{aligned}$$

If $u \sim v$, any future event F, and single observation a

$$\begin{aligned}
\Pr\left(X_{t+1}^{\infty} \in aF | X_{-\infty}^{t} = u\right) &= \Pr\left(X_{t+1}^{\infty} \in aF | X_{-\infty}^{t} = v\right) \\
\Pr\left(X_{t+1} = a, X_{t+2}^{\infty} \in F | X_{-\infty}^{t} = u\right) &= \Pr\left(X_{t+1} = a, X_{t+2}^{\infty} \in F | X_{-\infty}^{t} = v\right)
\end{aligned}$$

$$\begin{aligned}
&\Pr\left(X_{t+2}^{\infty} \in F | X_{-\infty}^{t+1} = ua\right) \Pr\left(X_{t+1} = a | X_{-\infty}^{t} = u\right) \\
&= \Pr\left(X_{t+2}^{\infty} \in F | X_{-\infty}^{t+1} = va\right) \Pr\left(X_{t+1} = a | X_{-\infty}^{t} = v\right)
\end{aligned}$$

$$\begin{aligned}
\Pr\left(X_{t+2}^{\infty} \in F | X_{-\infty}^{t+1} = ua\right) &= \Pr\left(X_{t+2}^{\infty} \in F | X_{-\infty}^{t+1} = va\right) \\
ua &\sim va
\end{aligned}$$

(same for continuous values or time but need more measure theory)

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

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

Entropy Rate

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

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

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

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

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

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

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

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

"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 $\phi^{-1} \circ f$

"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)

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 precision

About “Causal”

Term “causal states” introduced by Crutchfield and Young (1989) without too much precision
All about probabilistic prediction, not counterfactuals

About “Causal”

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

All about probabilistic prediction, not counterfactuals

(selecting sub-ensembles of naturally-occurring trajectories, not *enforcing* certain trajectories)

About "Causal"

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

All about probabilistic prediction, not counterfactuals

(selecting sub-ensembles of naturally-occurring trajectories, not *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
$\therefore$ ϵ joins cells of $\Gamma^{(\infty)}$
$\therefore$ ϵ 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"

Some Uses

Neural spike train analysis (Haslinger *et al.*, 2010), fMRI analysis (Merriam, Genovese and Shalizi in prep.)
Geomagnetic fluctuations (Clarke *et al.*, 2003)
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)

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.

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.

Marton, Katalin and Paul C. Shields (1994). “Entropy and the Consistent Estimation of Joint Distributions.” *Annals of Probability*, **22**: 960–977. URL `http://projecteuclid.org/euclid.aop/1176988736`. Correction, *Annals of Probability*, **24** (1996): 541–545.

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`.

Pfau, David, Nicholas Bartlett and Frank Wood (2010). "Probabilistic Deterministic Infinite Automata." In *Advances in Neural Information Processing Systems 23 [NIPS 2010]* (J. Lafferty and C. K. I. Williams and J. Shawe-Taylor and R.S. Zemel and A. Culotta, eds.), pp. 1930–1938. Cambridge, Massachusetts: MIT Press. URL `http://books.nips.cc/papers/files/nips23/NIPS2010_1179.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`.