Optimal Prediction

Cosma Shalizi

15 June 2010
Complex Systems Summer School

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}^{∞}

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 $X_{-\infty}^{t}$

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 **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 (need information theory to be precise about this)

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)

Excuse for not worrying about particular loss functions

Causal States

Crutchfield and Young (1989)
Histories *a* and *b* are equivalent iff

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

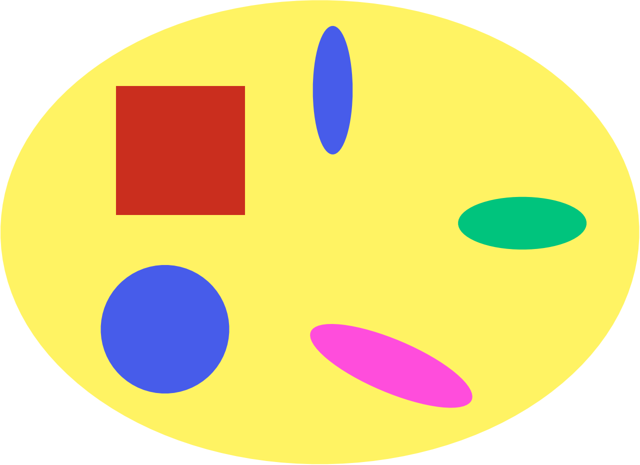
$[a] \equiv$ all histories equivalent to a
The statistic of interest, the **causal state**, is

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

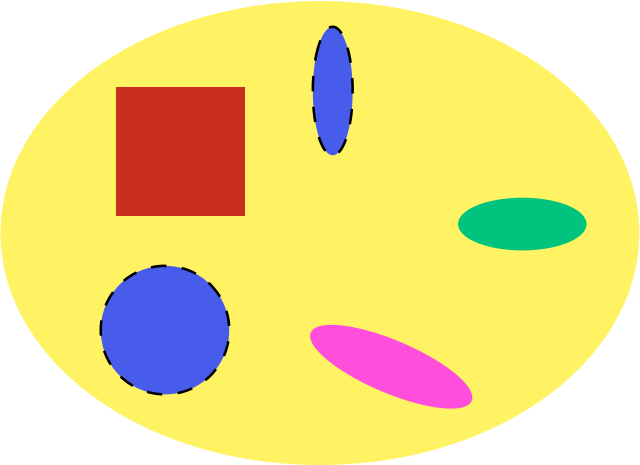
Each state is an equivalence class of histories
Each state is a conditional distribution over future events
IID = 1 state, periodic = p states

About "Causal"

Term introduced by Crutchfield and Young (1989)
For statistics, "causal" $\approx$ conditional independence *under manipulation* (Spirtes *et al.*, 2001; Pearl, 2009)
These states give us conditional independence but no guarantees about counterfactuals; *candidates* for causal models (Shalizi and Moore, 2003)



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



Partitioning histories into causal states

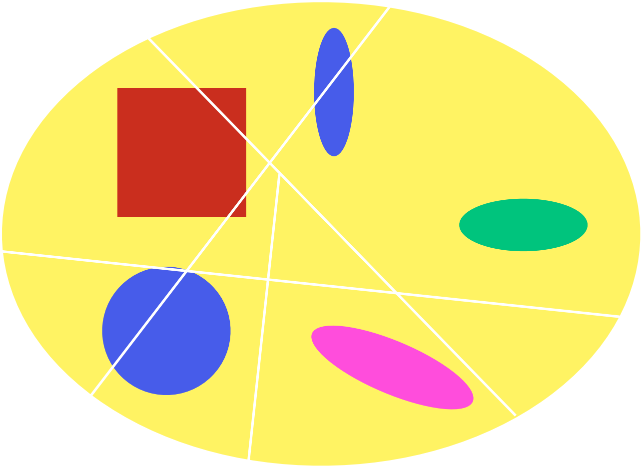
Sufficiency

Shalizi and Crutchfield (2001)

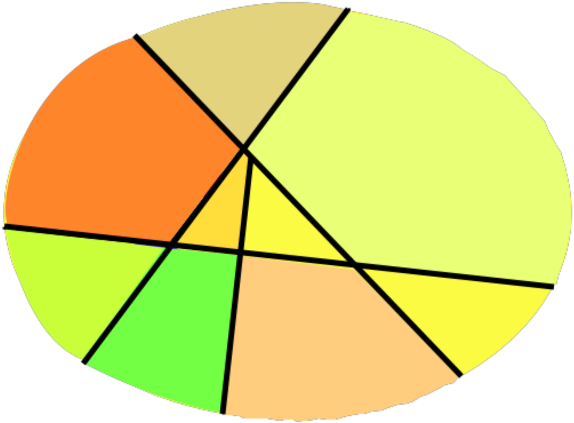
$$I[X_{t+1}^{\infty}; X_{-\infty}^{t}] = I[X_{t+1}^{\infty}; \epsilon(X_{-\infty}^{t})]$$

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

Markov Properties

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

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

because of sufficiency:

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

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)

If $a \sim b$, any future event F, and single observation f

$$
\begin{aligned}
\Pr\left(X_{t+1}^{\infty} \in fF | X_{-\infty}^{t} = a\right) &= \Pr\left(X_{t+1}^{\infty} \in fF | X_{-\infty}^{t} = b\right) \\
\Pr\left(X_{t+1} = f, X_{t+2}^{\infty} \in F | X_{-\infty}^{t} = a\right) &= \Pr\left(X_{t+1} = f, X_{t+2}^{\infty} \in F | X_{-\infty}^{t} = b\right) \\
& \ldots \\
\Pr\left(X_{t+2}^{\infty} \in F | X_{-\infty}^{t} = a, X_{t+1}^{\infty} = f\right) &= \Pr\left(X_{t+2}^{\infty} \in F | X_{-\infty}^{t} = b, X_{t+1}^{\infty} = f\right) \\
\Pr\left(X_{t+2}^{\infty} \in F | X_{-\infty}^{t+1} = af\right) &= \Pr\left(X_{t+2}^{\infty} \in F | X_{-\infty}^{t+1} = bf\right) \\
af &\sim bf
\end{aligned}
$$

EXERCISE: Filling in the missing step

Causal States are Markovian

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

because
S_{t+1}^{∞} is a function of S_t and X_{t+1}^{∞}
and
X_{t+1}^{∞} is independent of *all* of the past given S_t

Markovian Representation

The observed process (X_t) is non-Markovian and ugly
But it is generated from a homogeneous Markov process (S_t)
Not the usual sort of hidden Markov model because of the deterministic transitions
(An advantage, HMMs need complicated calculations to estimate distributions over their states)

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

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

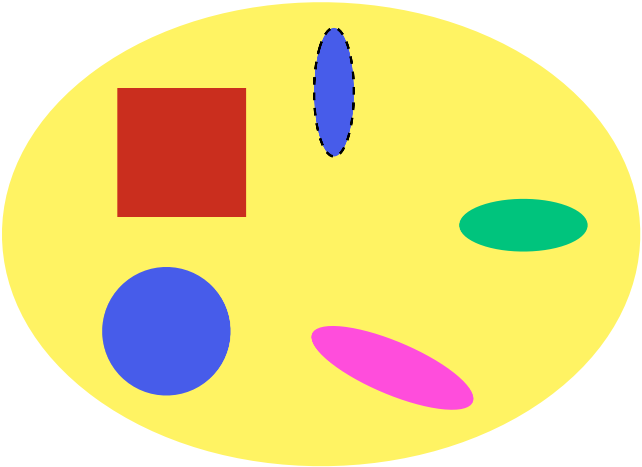
Minimality

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

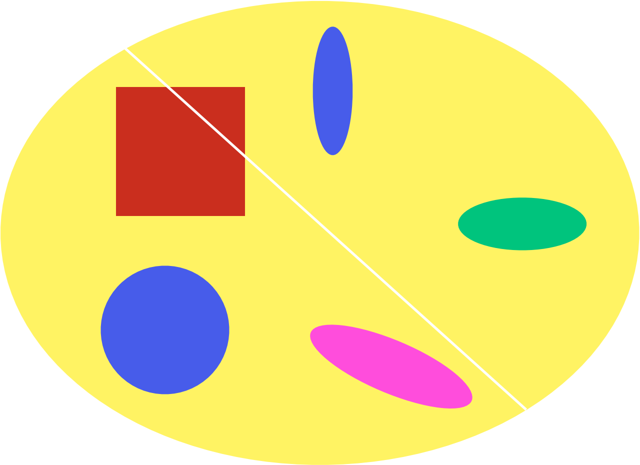
$$\epsilon(X^t_{-\infty}) = g(\eta(X^t_{-\infty}))$$

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

Uniqueness

There is no other minimal sufficient statistic

Uniqueness

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

$$\eta = h(\epsilon)$$

Uniqueness

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

$$\eta = h(\epsilon)$$

but $\epsilon = g(\eta)$

Uniqueness

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

$$\eta = h(\epsilon)$$

but $\epsilon = g(\eta)$ so

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

$g = h^{-1}$ and ϵ and η partition histories in the same way

Minimal stochasticity

If $R_t = \eta(X^t_{-\infty})$ is also sufficient, then

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

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

Entropy Rate

Recall $h_1 = \lim_{n\rightarrow\infty} H[X_n|X_1^{n-1}]$

Entropy Rate

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

$$\begin{aligned}\lim_{n\to\infty} H[X_n|X_1^{n-1}] &= \lim_{n\to\infty} H[X_n|S_{n-1}] \\ &= H[X_1|S_0]\end{aligned}$$

Entropy Rate

Recall $h_1 = \lim_{n\rightarrow\infty} H[X_n|X_1^{n-1}]$

$$\begin{aligned}\lim_{n\rightarrow\infty} H[X_n|X_1^{n-1}] &= \lim_{n\rightarrow\infty} H[X_n|S_{n-1}] \\ &= H[X_1|S_0]\end{aligned}$$

so knowing the causal states lets us calculate the entropy rate

History and Aliases

- Statistical relevance basis (Salmon, 1971, 1984)

History and Aliases

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

History and Aliases

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

History and Aliases

- 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)

History and Aliases

- 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)

History and Aliases

- 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)

History and Aliases

- 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)

Statistical Complexity

Definition

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

Statistical Complexity

Definition

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

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

Statistical Complexity

Definition

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

Statistical Complexity

Definition

$C \equiv I[\epsilon(X^t_{-\infty}); X^t_{-\infty}]$ 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
$=$ log(period) for period processes

Statistical Complexity

Definition

$C \equiv I[\epsilon(X^t_{-\infty}); X^t_{-\infty}]$ 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
$=$ log(period) for period processes
$=$ log(geometric mean(recurrence time)) for stationary processes

Statistical Complexity

Definition

$C \equiv I[\epsilon(X^t_{-\infty}); X^t_{-\infty}]$ 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
$=$ log(period) for period processes
$=$ log(geometric mean(recurrence time)) for stationary processes
$=$ information about microstate in macroscopic observations (sometimes)

Can We Find Causal State Models?

Depends on the meaning of “find”

- Parameter estimation with known structure (“learning”)
  - curved exponential families
  - maximum likelihood estimation is simple, consistent and efficient
- Reconstruct the structure from observed behavior (“discovery”)

CSSR: Causal State Splitting Reconstruction

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

- Get next-step distribution right
- 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 real 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 real 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 (from AEP)

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

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:

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

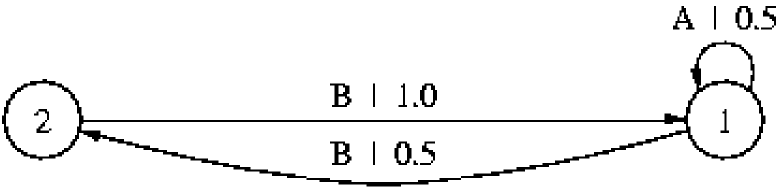
$\mathcal{D}$ = true distribution, $\widehat{\mathcal{D}}_n$ = inferred
Error (in L_1/total variation) scales like independent samples

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

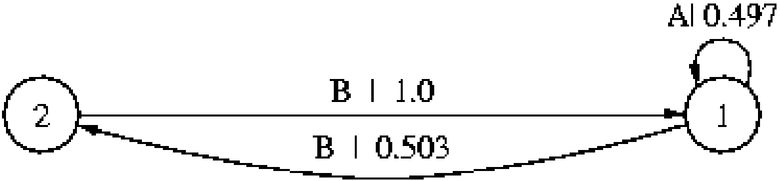
Handwaving

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



Blocks of As of any length, separated by even-length blocks of Bs
Infinite-range correlation (not Markov at any order)



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

Some Uses

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)
Neural spike train analysis (Haslinger *et al.*, 2010)
Spatio-temporal applications: next lecture!

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*.

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.

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.

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.

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.

Pearl, Judea (2009). *Causality: Models, Reasoning, and Inference*. Cambridge, England: Cambridge University Press, 2nd edn.

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, 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 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 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.

Spirtes, Peter, Clark Glymour and Richard Scheines (2001). *Causation, Prediction, and Search*. Cambridge, Massachusetts: MIT Press, 2nd edn.