Chaos, Complexity, and Inference (36-462)

Lecture 6

Cosma Shalizi

31 January 2008

The Story So Far

Deterministic dynamics can produce stable distributions of behavior
Discretizing with partitions gives symbol sequences
These need a statistical description

Inference for Markov chains

Inference for higher-order Markov chains

Inference for stochastic machines

Likelihood for Markov chains

Basic case: m states/symbols, transition matrix p^0 unknown
Parameters: matrix entries p_{ij}
observe $x_1^n \equiv x_1, x_2, \ldots x_n$
Th probability of this sequence is

$$\Pr\left(X_1^n = x_1^n\right) = \Pr\left(X_1 = x_1\right) \prod_{t=2}^{n} \Pr\left(X_t = x_t | X_{t-1} = x_{t-1}\right)$$

(by Markov property)

Re-write in terms of p_{ij}

$$L(P) = \Pr(X_1 = x_1) \prod_{t=2}^{n} p_{x_{t-1} x_t}$$

Define $N_{ij} \equiv$ number of times i is followed by j in X_1^n

$$L(P) = \Pr(X_1 = x_1) \prod_{i=1}^{m} \prod_{j=1}^{m} p_{ij}^{n_{ij}}$$

$$\mathcal{L}(P) = \log \Pr(X_1 = x_1) + \sum_{i,j} n_{ij} \log p_{ij}$$

Maximize as a function of all the p_{ij}

Solution:

$$\hat{p}_{ij} = \frac{n_{ij}}{\sum_j n_{ij}}$$

What about x_1? Use conditional likelihood to ignore it!
By the ergodic theorem,

$$\frac{N_{ij}}{n} \rightarrow p_i^0 p_{ij}^0$$

(where did p_i^0 come from?) also

$$\sum_j \frac{N_{ij}}{n} \rightarrow p_i^0$$

so

$$\hat{p}_{ij} \rightarrow p_{ij}^0$$

as we'd like

Parametrized Markov Chains

- May not be able to vary all the transition probabilities separately
- May have an actual theory about how the transition probabilities are functions of underlying parameters

In both cases, P is really $P(\theta)$, with θ the r-dimensional vector of parameters
Again, maximize the likelihood:

$$\frac{\partial \mathcal{L}}{\partial \theta_u} = \sum_{ij} \frac{\partial \mathcal{L}}{\partial p_{ij}} \frac{\partial p_{ij}}{\partial \theta_u}$$

For this to work, we need Guttorp's "Conditions A"

which he got from [1, p. 23]

1. The allowed transitions are the same for all θ
   technical convenience
2. $p_{ij}(\theta)$ has continuous θ-derivatives up to order 3
   authorizes Taylor expansions to 2nd order
   can sometimes get away with just 2nd partials
3. The matrix $\partial p_{ij} / \partial \theta_u$ always has rank r
   no redundancy in the parameter space
4. The chain is ergodic without transients for all θ
   trajectories are representative samples

Assume all this; also, $\theta^0 =$ true parameter value
Then:

1. MLE $\hat{\theta}$ exists
2. $\hat{\theta} \rightarrow \theta^0$ (consistency)
3. Asymptotic normality:

$$\sqrt{n}\left(\hat{\theta} - \theta^0\right) \rightsquigarrow \mathcal{N}(0, I^{-1}(\theta^0))$$

with **expected (Fisher) information**

$$I_{uv}(\theta) = \sum_{ij} \frac{p_i(\theta)}{p_{ij}(\theta)} \frac{\partial p_{ij}}{\partial \theta_u} \frac{\partial p_{ij}}{\partial \theta_v} = -\sum_{ij} p_i(\theta) p_{ij}(\theta) \frac{\partial^2 \log p_{ij}(\theta)}{\partial \theta_u \partial \theta_v}$$

(2nd equality is not obvious)

Error estimates based on $I(\theta^0)$ are weird: if you knew θ^0, why would you be calculating errors?
Option 1: use $I(\hat{\theta})$
Option 2: use the **observed information**

$$J_{uv} = -\sum_{ij} \frac{n_{ij}}{n} \frac{\partial^2 \log p_{ij}(\hat{\theta})}{\partial\theta_u \partial\theta_v}$$

(Guttorp's Eq. 2.207, but he's missing the sum over state pairs.)
Notice that

$$J_{uv} = -\frac{1}{n}\frac{\partial^2 \mathcal{L}(\hat{\theta})}{\partial\theta_u \partial\theta_v}$$

nJ is how much the likelihood changes with a small change in parameters from the maximum; J^{-1} is how much we can change the parameters before the change in likelihood is noticeable

Alternative error estimates

Can get standard errors and confidence intervals from these Gaussian distributions

but they're asymptotic

Generally no simple formulas for the finite-sample distributions

This doesn't matter (much) because we can simulate

Parametric bootstrapping

1. Have real data x_1^n, get parameter estimate $\hat{\theta}$
2. Simulate from $\hat{\theta}$, get fake data Y_1^n ("bootstrap")
3. Estimate from faked data, get $\tilde{\theta}$

Approximately,

$$(\hat{\theta} - \theta^0) \sim (\tilde{\theta} - \hat{\theta})$$

We want the distribution on the left; we can get arbitrarily close to the distribution on the right, by repeating steps 2 and 3 as many times as we want

(Connections between bootstrap and maximum likelihood: [2])

Higher-order Markov Chains

Markov property: for all t,

$$\Pr\left(X_t|X_1^{t-1}\right) = \Pr\left(X_t|X_{t-1}\right)$$

k^{th}-order Markov: for all t,

$$\Pr\left(X_t|X_1^{t-1}\right) = \Pr\left(X_t|X_{t-k}^{t-1}\right)$$

In a Markov chain, the *immediate* state determines the distribution of future trajectories

Extended chain device: Define $Y_t = X_t^{t+k-1}$

Y_1^t is a Markov chain

The likelihood theory is thus exactly the same, only we need to condition on the first k observations

Hypothesis Testing

Likelihood-ratio testing is simple, for nested hypotheses
$\hat{\theta}_{\text{small}}$ = MLE under the smaller, more restricted hypothesis, d_{small} degrees of freedom
$\hat{\theta}_{\text{big}}$ = MLE under larger hypothesis, d.o.f. d_{big}
If the smaller hypothesis is true,

$$2[\mathcal{L}(\hat{\theta}_{\text{big}}) - \mathcal{L}(\hat{\theta}_{\text{small}})] \rightsquigarrow \chi^2_{d_{\text{big}} - d_{\text{small}}}$$

Everything is nested inside the non-parameterized estimate; it has $m(m-1)$ degrees of freedom for a first-order chain, $m^k(m-1)$ for a k-order chain.
fixed transition matrix, or fixed value of θ^0, has 0 d.o.f.
lower-order chains are nested inside higher-order chains, so you can test for order restrictions

Partially-observable Markov chain process where we observe a random function of a Markov chain

$$X_t = f(S_t, N_t),\ S_t \text{ Markov},\ N_t \perp S_t$$

Hidden Markov model observation X_t independent of everything else given state S_t

Stochastic finite automaton X_t plus S_t uniquely determine S_{t+1} a.k.a. **chain with complete connections**

HMMs and SFAs are both special cases of POMCs
HMMs are more common in signal processing
SFAs are more useful for dynamics, and easier to analyze: stochastic counterparts to the machines from last lecture
Good intros to HMMs: [3, 4]
Good advanced reference on HMMs: [5]

Specification of an SFA:

1. Set of states $\mathcal{S}$, alphabet of symbols $\mathcal{A}$
2. Transition function $T(i,j) =$ state reached starting from i on symbol j
3. Emission probabilities $Q_{ij} =$ probability of state i producing symbol j
4. Initial distribution over states

Graph: circles and arrows, as before; add probabilities Q_{ij} to the arrows
Skeleton or **structure** of SFA: just (1) and (2)

Likelihood theory for SFA Observe x_1^n
Assume skeleton is known, initial state s_1 is known
Then state sequence is known recursively: $s_{t+1} = T(s_t, x_t)$
Log-likelihood:

$$\mathcal{L}(Q) = \sum_{t=1}^{n} \log Q_{s_t x_t} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{A}} n_{ij} \log Q_{ij}$$

with N_{ij} = emission counts
Once again,

$$\hat{Q}_{ij} = \frac{n_{ij}}{\sum_{j \in \mathcal{A}} n_{ij}}$$

and once again

$$\hat{Q}_{ij} \rightarrow Q^0_{ij}$$

If the initial state is not known:
Likelihood becomes weighted sum of state-conditional likelihoods; somewhat ugly but numerically maximizable
Synchronization: Write $s_{t+1} = T(s_1, x_1^t)$ — abuse of notation
Skeleton **synchronizes** if, after some τ, $T(s_1, x_1^\tau) = T(s_1', x_1^\tau)$
or, x_1^τ is enough to pin down the state, never mind starting point
All finite-type processes synchronize ($\tau =$ order of process)
Many strictly sofic processes synchronize after a random time (e.g. all three examples from Lecture 5)
Can do likelihood conditional on synchronization

What do if the skeleton is not known?

1. Try multiple skeletons, cross-validate
2. Try multiple skeletons, use BIC

$$BIC = \mathcal{L}(\hat{\theta}) - \frac{d}{2} \log n$$

Hand-waving:
Large $n \Rightarrow \hat{\theta}$ Gaussian around θ^0, s.d. $\propto n^{-1/2}$
Parameters with more impact on likelihood more precisely estimated
$-\frac{d}{2} \log n$ comes out as expected over-fitting
BIC is consistent for estimating the order of Markov chains

3. Other model-selection tests/heuristics (e.g. bootstrap tests)

Model Discovery/Construction

Systematically build a model to match the data

Basic idea for Markov chains goes back to John Foulkes's Janet in the 1950s [6]

Each state contains a word s; a sequence of observations should land us in that state if they end with that word

For each state, keep track of the conditional distribution $\Pr(X_t|s)$.

Also keep track of $\Pr(X_t|as)$, for each one-symbol extension as.

If $\Pr(X_t|s)$ differs significantly from $\Pr(X_t|as)$, split into multiple states.

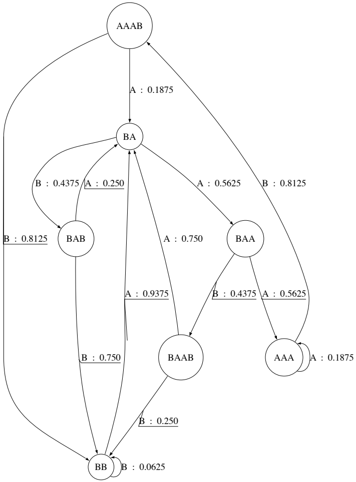
Keep going until no more splits are called for.

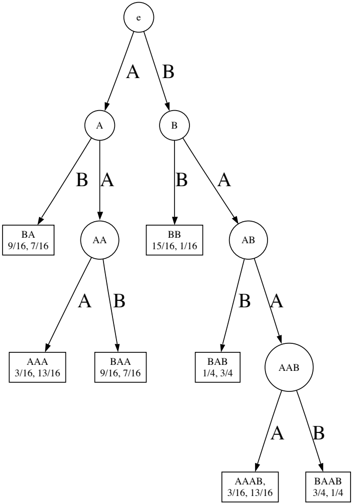
Result: **variable-length Markov chain**

Variable-length Markov chains are equivalent to higher-order Markov chains — why bother?

Computation and comprehensibility: tree representation

Statistics: fewer degrees of freedom ($m - 1$ per state), which means more efficient





Foulkes's example: 7 state machine, word length ≤ 4

- Periodic re-discoveries of Foulkes's idea [7, 8, 9, 10]
- Check out the `VLMC` package from CRAN
- Some evidence that people (or at least mid-1960s undergrads in Michigan) do something like this [11]
  - More exactly, people seem to learn the states, but don't make the right predictions in those states
  - This would be a nice topic to re-visit

What about sofic processes?

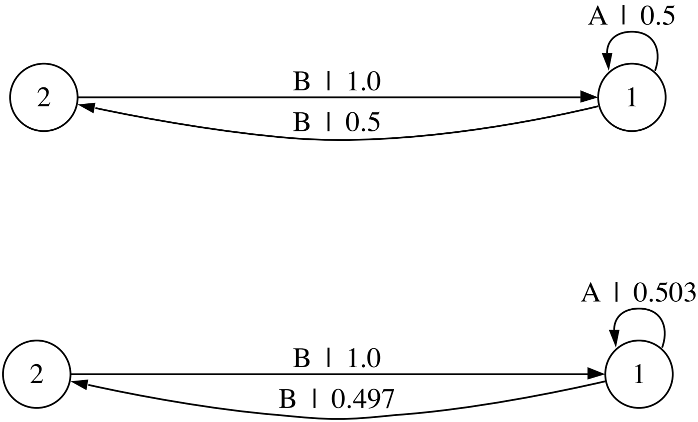
Learning strictly sofic machines is more tricky
One approach is CSSR ("causal state-splitting reconstruction") [12]

1. Learn states (tree-like) which predict one step ahead, much like Janet

$$\Pr\left(X_{t+1}|S_t\right) = \Pr\left(X_{t+1}|X_1^t\right)$$

2. Then sub-divide states until they are resolving, i.e. must have $R_{t+1} = T(R_t, X_t)$, and $S_t = f(R_t)$ for some T, f

Can learn even strictly sofic processes *if* they are synchronizing
Must not learn strict tree in (1), *and* must do (2)



exact even process vs. CSSR with $n = 10^4$

Error estimates: bootstrap
(paper in preparation on analytical theory but it is very tricky)

[1] Patrick Billingsley. *Statistical Inference for Markov Processes*. University of Chicago Press, Chicago, 1961.

[2] Bradley Efron. Maximum likelihood and decision theory. *The Annals of Statistics*, 10:340–356, 1982. URL http://links.jstor.org/sici?sici=0090-5364%28198206%2910%3A2%3C340%3AMLADT%3E2.0.CO%3B2-I.

[3] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286, 1989. URL http://dx.doi.org/10.1109/5.18626.

[4] Eugene Charniak. *Statistical Language Learning*. MIT Press, Cambridge, Massachusetts, 1993.

[5] Olivier Cappé, Eric Moulines, and Tobias Rydén. *Inference in Hidden Markov Models*. Springer, New York, 2005.

[6] J. D. Foulkes. A class of machines which determine the statistical structure of a sequence of characters. *Wescon Convention Record*, 4:66–73, 1959.

[7] Jorma Rissanen. A universal data compression system. *IEEE Transactions on Information Theory*, 29:656–664, 1983.

[8] Dana Ron, Yoram Singer, and Naftali Tishby. The power of amnesia: Learning probabilistic automata with variable memory length. *Machine Learning*, 25:117–149, 1996.

[9] Peter Bühlmann and Abraham J. Wyner. Variable length Markov chains. *The Annals of Statistics*, 27:480–513, 1999. URL `http://www.stat.berkeley.edu/tech-reports/479.abstract`.

[10] Matthew B. Kennel and Alistair I. Mees. Context-tree modeling of observed symbolic dynamics. *Physical Review E*, 66:056209, 2002.

[11] Julian Feldman and Joe F. Hanna. The structure of responses to a sequence of binary events. *Journal of Mathematical Psychology*, 3:371–387, 1966.

[12] Cosma Rohilla Shalizi and Kristina Lisa Klinkner. Blind construction of optimal nonlinear recursive predictors for discrete sequences. In Max Chickering and Joseph Y. Halpern, editors, *Uncertainty in Artificial Intelligence: Proceedings of the Twentieth Conference (UAI 2004)*, pages 504–511, Arlington, Virginia, 2004. AUAI Press. URL `http://arxiv.org/abs/cs.LG/0406011`.