

Chaos, Complexity, and Inference (36-462)

Lecture 6: Statistical Inference for Discrete Stochastic Processes

Cosma Shalizi

29 January 2009

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](#)

FURTHER READING: *Everyone* steals the theory of likelihood inference for Markov chains from Billingsley (1961). But Guttorp (1995) is easier reading. For hidden Markov models, see Fraser (2008).

Likelihood for Markov chains

Basic case: m states/symbols, transition matrix p^0 unknown

Parameters: matrix entries p_{ij}

observe $x_1^n \equiv x_1, x_2, \dots, x_n$

The probability of this sequence is

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

(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}

The Maximum Likelihood Estimator

Solution to constrained maximization problem (see handout):

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

What about x_1 ? Use conditional likelihood to ignore it!

Why the MLE Works

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

Either way, 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 Billingsley (1961, 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, $\theta^0 = \text{true parameter value}$

Then:

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

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

with **expected (Fisher) information**

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

(2nd equality is not obvious)

Error of the MLE

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 = how much the likelihood changes with a small change in parameters from the maximum
- J^{-1} = how much we can change the parameters before the change in likelihood is noticeable
- If the model is right, then $J \rightarrow I(\hat{\theta})$, because both $\rightarrow I(\theta^0)$
∴ use $J - I(\hat{\theta})$ to test for mis-specification (White, 1994)

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 needed

(Connections between bootstrap and maximum likelihood: Efron (1982))

Higher-order Markov Chains

Markov property: for all t ,

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

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

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

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,

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

which gives significance level

- If the bigger hypothesis is true, a non-central χ^2 distribution
(can give power)

What about small n ? Distribution won't have converged

Use parametric bootstrapping again:

- 1 Calculate log likelihood ratio Λ on real data, call this λ
- 2 Simulate from $\hat{\theta}_{\text{small}}$, get fake data Y_1^n
- 3 Estimate $\tilde{\theta}_{\text{small}}, \tilde{\theta}_{\text{big}}$ from Y_1^n
- 4 Calculate $\tilde{\Lambda}$ from $\tilde{\theta}_{\text{small}}, \tilde{\theta}_{\text{big}}$
- 5 Repeat (2)–(4) B times to get sample of $\tilde{\Lambda}$
- 6 $p\text{-value} = \# \{ \tilde{\Lambda} \geq \lambda \} / B$

Getting power is similar but simulate from $\hat{\theta}_{\text{big}}$

Everything is nested inside the non-parameterized estimate

d.o.f. = $m(m - 1)$ for a first-order chain

d.o.f. = $m^k(m - 1)$ for a k -order chain

fixed transition matrix, or fixed θ^0 , has d.o.f. = 0

lower-order chains are nested inside higher-order chains, so
you can test for order restrictions

A Little Bit Beyond Markov Chains

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: Rabiner (1989); Charniak (1993), and especially Fraser (2008)

Good advanced reference on HMMs: Cappé *et al.* (2005)

Specifying 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 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

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

If the initial state is not known:

Likelihood = weighted sum of state-conditional likelihoods

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 (τ = 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 (Csiszár and Shields, 2000)

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

Model Discovery/Construction

Systematically build a model to match the data

Basic idea goes back to the JANET algorithm (Foulkes, 1959)

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

Each state has a conditional distribution $\Pr(X_t|s)$.

Each state also has $\Pr(X_t|as)$, 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 Chain

Tree representation

Equivalent to higher-order Markov chain

order = length of longest path from root to leaf

So why bother with VLMCs?

Computation and comprehensibility

Use the tree to predict, and to see what the states “mean” in terms of history

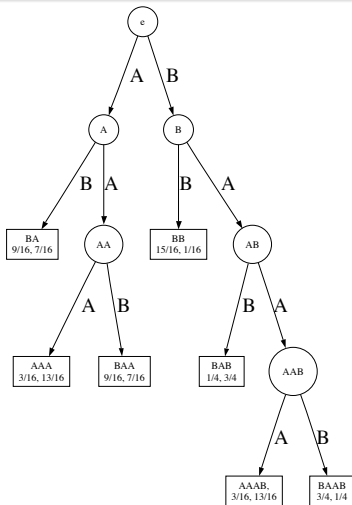
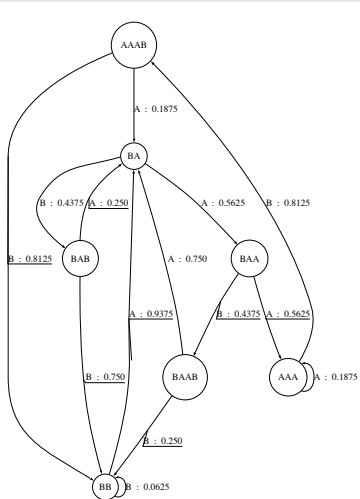
Statistical efficiency

$m - 1$ degrees of freedom $\leq m^k(m - 1)$ d.o.f. for full chain

fewer d.o.f. \Rightarrow less variance in estimates

weaker “curse of dimensionality”

BIC works for selecting VLMCs (Csiszár and Talata, 2006)



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

Periodic re-discoveries of Foulkes's idea

(Rissanen, 1983; Ron *et al.*, 1996; Bühlmann and Wyner, 1999; Kennel and Mees, 2002)

Check out the `VLMC` package from CRAN

Some evidence that people (or at least mid-1960s undergrads in Michigan) do something like this (Feldman and Hanna, 1966)

More exactly, people seem to learn the states, but don't make the right predictions in those states

This would be a nice topic for someone to re-visit

What about sofic processes?

Learning strictly sofic machines is more tricky

One approach is CSSR (“causal state-splitting reconstruction”) (Shalizi and Klinkner, 2004)

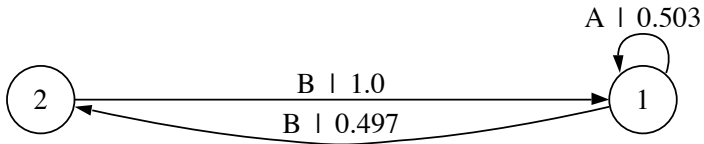
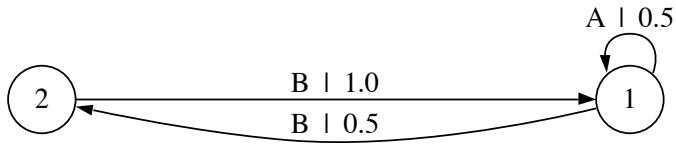
- 1 Learn states (tree-like) which predict one step ahead, much like Janet

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

- 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)

- Billingsley, Patrick (1961). *Statistical Inference for Markov Processes*. Chicago: University of Chicago Press.
- Bühlmann, Peter and Abraham J. Wyner (1999). “Variable Length Markov Chains.” *The Annals of Statistics*, **27**: 480–513. URL <http://projecteuclid.org/euclid.aos/1018031204>.
- Cappé, Olivier, Eric Moulines and Tobias Rydén (2005). *Inference in Hidden Markov Models*. New York: Springer.
- Charniak, Eugene (1993). *Statistical Language Learning*. Cambridge, Massachusetts: MIT Press.
- Csiszár, Imre and Paul C. Shields (2000). “The Consistency of the BIC Markov order estimator.” *Annals of Statistics*, **28**: 1601–1619. URL <http://projecteuclid.org/euclid.aos/1015957472>.
- Csiszár, Imre and Zsolt Talata (2006). “Context Tree Estimation for Not Necessarily Finite Memory Processes, Via BIC and

MDL.” *IEEE Transactions on Information Theory*, **52**: 1007–1016. doi:10.1109/TIT.2005.864431.


Efron, Bradley (1982). “Maximum Likelihood and Decision Theory.” *The Annals of Statistics*, **10**: 340–356. URL <http://projecteuclid.org/euclid.aos/1176345778>.

Feldman, Julian and Joe F. Hanna (1966). “The Structure of Responses to a Sequence of Binary Events.” *Journal of Mathematical Psychology*, **3**: 371–387.

Foulkes, J. D. (1959). “A Class of Machines which Determine the Statistical Structure of a Sequence of Characters.” In *Wescon Convention Record*, vol. 4, pp. 66–73. Institute of Radio Engineers.

Fraser, Andrew M. (2008). *Hidden Markov Models and Dynamical Systems*. Philadelphia: SIAM Press.

Guttorp, Peter (1995). *Stochastic Modeling of Scientific Data*. Stochastic Modeling. London: Chapman and Hall.

- Kennel, Matthew B. and Alistair I. Mees (2002). “Context-tree Modeling of Observed Symbolic Dynamics.” *Physical Review E*, **66**: 056209.
- Rabiner, Lawrence R. (1989). “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.” *Proceedings of the IEEE*, **77**: 257–286.
doi:10.1109/5.18626.
- Rissanen, Jorma (1983). “A Universal Data Compression System.” *IEEE Transactions on Information Theory*, **29**: 656–664.
- Ron, Dana, Yoram Singer and Naftali Tishby (1996). “The Power of Amnesia: Learning Probabilistic Automata with Variable Memory Length.” *Machine Learning*, **25**: 117–149.
- 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>.

White, Halbert (1994). *Estimation, Inference and Specification Analysis*. Cambridge, England: Cambridge University Press.