Discrete-time Markov chains

Consider a sequence of random variables $S_1, S_2, S_3, \ldots$ with common state space $\mathcal{S} = \{1, 2, \ldots, |\mathcal{S}|\}$.

This sequence forms a Markov chain if $\{S_1, S_2, \ldots, S_{t-1}\}$ and $\{S_{t+1}, S_{t+2}, \ldots\}$ are conditionally independent, given S_t, for all t.
Equivalently, $\Pr(S_t | S_{t-1}, S_{t-2}, \ldots) = \Pr(S_t | S_{t-1})$.

The matrix $p_{tij} = \Pr(S_{t+1} = j | S_t = i)$ (for fixed t) is called the transition matrix (at step t). Note that $p_{tij} \geq 0$ and $\sum_j p_{tij} = 1$.

Often we assume $p_{tij} = p_{1ij} \equiv p_{ij}$ for all t, in which case the Markov chain is said to be *homogeneous*. Let P denote the matrix (p_{ij}).

Let's write down a few results for *homogeneous* Markov chains; we won't be using these today, but may return to them next week.

The two-step transition matrix is obtained by matrix multiplication:

$$p_{ij}^{(2)} = \Pr(S_3 = j | S_1 = i) = \sum_k p_{ik} p_{kj} = (P^2)_{ij}$$

Markov chains (continued)

Of special interest is the limiting form, $\lim_{n\to\infty} P^n$.

A distribution π (on $\mathcal{S}$) is a stationary (or equilibrium) distribution of the chain if $\pi' = \pi' P$.

For finite-state chains, such π always exist.

A few definitions:

A chain is *aperiodic* if $\text{gcd}\{n \geq 1 : p_{ii}^{(n)} > 0\} = 1$ for all i (e.g., $p_{ii} > 0$ for all i).

A chain is *irreducible* if for every pair of states (i, j), there exists $n_{ij} > 0$ such that $p_{ij}^{(n_{ij})} > 0$.

A finite-state Markov chain is aperiodic and irreducible iff there is some power n so that P^n has all entries positive.

For aperiodic, irreducible Markov chains, the stationary distribution, π is unique and $\lim_{n\to\infty} P^n = (\pi' \ldots \pi')'$. All entries of π are necessarily positive.

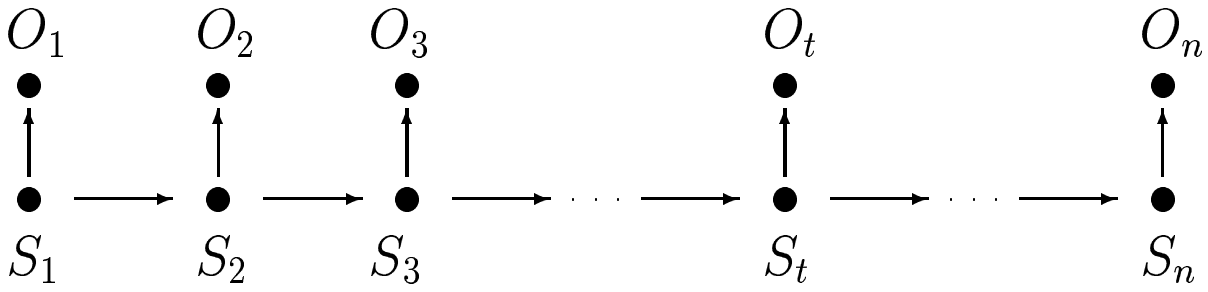
References: Lange (ch 23), Feller vol 1 (ch 15)

Hidden Markov models

Consider an inhomogeneous Markov chain $S_1, S_2, \ldots S_n$ with common state space $\mathcal{S} = \{1, 2, \ldots, |\mathcal{S}|\}$, initial distribution $\pi_i = \mathsf{Pr}(S_1 = i)$ and transition probabilities $p_{ij}^{(t)} = \mathsf{Pr}(S_{t+1} = j | S_t = i)$. [Note the change in notation!]

Suppose we don't observe the S_t, but rather observe random variables $O_t \in \mathcal{O}$, with O_t conditionally independent of everything else, given S_t, and $q_i^{(t)}(O_t) = \mathsf{Pr}(O_t | S_t = i)$.

Let $O = (O_1, O_2, \ldots, O_n)$.



This is a *hidden Markov model* (HMM).

Note that

$$\mathsf{Pr}(O, S_1 = i_1, \ldots S_n = i_n) = \pi_{i_1} \prod_{t=1}^{n-1} p_{i_t i_{t+1}}^{(t)} \prod_{t=1}^{n} q_{i_t}^{(t)}(O_t)$$

Ref: Rabiner, Proceedings of the IEEE 77:257–286, 1989

Example 1

Genetic mapping in experimental crosses

Consider a single F_2 offspring from an intercross between inbred mouse strains, typed at a set of n markers (in known order) along a chromosome.

Let r_t denote the recombination fraction between markers t and $t+1$. Let $S_t \in \{AA, AB, BA, BB\}$ denote the individual's *phase-known genotype* at marker t. Let $O_t \in \{A, H, B, C, D, -\}$ denote its *observed* genotype at the marker, where $C = \{H \text{ or } B\} =$ not A, $D = \{A \text{ or } H\} =$ not B, and $- =$ missing

Under the assumption of no crossover interference, the S_t for an inhomogeneous Markov chain with transition matrix

$$P_t = \begin{pmatrix} (1-r_t)^2 & r_t(1-r_t) & r_t(1-r_t) & r_t^2 \\ r_t(1-r_t) & (1-r_t)^2 & r_t^2 & r_t(1-r_t) \\ r_t(1-r_t) & r_t^2 & (1-r_t)^2 & r_t(1-r_t) \\ r_t^2 & r_t(1-r_t) & r_t(1-r_t) & (1-r_t)^2 \end{pmatrix}$$

The initial distribution is $\pi = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})'$

The O_t form an HMM.

Example 1 (continued)

The matrix $q_i^{(t)}(O_t)$ is as follows:

	O_t					
S_t	A	H	B	C	D	$-$
AA	$1-\epsilon$	$\frac{1}{2}\epsilon$	$\frac{1}{2}\epsilon$	ϵ	$1-\frac{1}{2}\epsilon$	1
AB	$\frac{1}{2}\epsilon$	$1-\epsilon$	$\frac{1}{2}\epsilon$	$1-\frac{1}{2}\epsilon$	$1-\frac{1}{2}\epsilon$	1
BA	$\frac{1}{2}\epsilon$	$1-\epsilon$	$\frac{1}{2}\epsilon$	$1-\frac{1}{2}\epsilon$	$1-\frac{1}{2}\epsilon$	1
BB	$\frac{1}{2}\epsilon$	$\frac{1}{2}\epsilon$	$1-\epsilon$	$1-\frac{1}{2}\epsilon$	$\frac{1}{2}\epsilon$	1

where ϵ is the genotyping error rate.

Major problems:

- Find the MLEs of the recombination fractions r_t

  $\hat{r} = \arg\max_r \Pr(O|r)$

- Calculate the log likelihood $l(\hat{r}|O) = \log \Pr(O|\hat{r})$ for different marker orders

- Reconstruct the underlying sequence $\{S_t\}$:
  - Find $\hat{S} = \arg\max_S \Pr(S|O, \hat{r})$
  - Simulate from $\Pr(S|O, \hat{r})$

Example 2

Estimating pairwise relationships

Refs: Boehnke and Cox, AJHG 61:423–9, 1997;
Broman and Weber, AJHG 63:1563–4, 1998.

Consider a pair of individuals genotyped at n linked markers in a known order, with known inter-marker distances, and with known allele frequencies.

We assume no crossover interference and that the markers are in linkage equilibrium, and we wish to calculate the likelihood, given the observed genotype data, of the five relationships MZ twins, full siblings, parent/offspring, half siblings and unrelated.

Let $S_t \in \{0, 1, 2\}$ denote the number of alleles shared *identical by descent* (IBD) by the two individuals at marker t. Let O_t denote the genotypes of the two individuals at marker t.

Then (at least for these five relationships) the S_t form a Markov chain, and the O_t follow an HMM.

Main problem: Calculate $l(R) = \log \mathsf{Pr}(O|R)$ for each of our five relationships R.

Example 3

A DNA sequence analysis example

Ref: Churchill, Bull Math Biol, 51:79–94, 1989.

Imagine that DNA is composed of GC-rich regions and GC-poor regions.

Let $S_t \in \{R, P\}$ indicate whether nucleotide t is within a GC-rich or GC-poor region, and imagine that the S_t form a homogeneous Markov chain with transition matrix

$$P = \begin{pmatrix} 1-p_2 & p_2 \\ p_1 & 1-p_1 \end{pmatrix}$$

and initial (stationary) distribution $\pi' = (p_1 \;\; p_2)/(p_1+p_2)$.

Let $O_t \in \{A, C, G, T\}$ be the observed nucleotide at site t. Let the $q_i^{(t)}(O_t)$ matrix be something like the following:

	O_t			
S_t	A	C	G	T
R	$\frac{1}{2}q_R$	$\frac{1}{2}(1-q_R)$	$\frac{1}{2}(1-q_R)$	$\frac{1}{2}q_R$
P	$\frac{1}{2}q_P$	$\frac{1}{2}(1-q_P)$	$\frac{1}{2}(1-q_P)$	$\frac{1}{2}q_P$

Problems: Estimate parameters; reconstruct $\{S_t\}$.

HMM technology

We wish to solve four problems:

- Calculate the likelihood $\Pr(O)$.
- Simulate from the joint distribution $\Pr(S|O)$.
- Find the most likely underlying sequence $\hat{S} = \arg\max_S \Pr(S|O)$.
- Find the MLEs of the parameters $\theta = [\pi_i, p_{ij}^{(t)}, q_i^{(t)}(O_t)]$; $\hat{\theta} = \arg\max_\theta \Pr(S|O, \theta)$.

Baum-Welch algorithm

Forward equations

Define $\alpha_t(i) = \mathsf{Pr}(O_1, \ldots, O_t, S_t = i)$.

$$\begin{aligned}
\alpha_1(i) &= \mathsf{Pr}(O_1, S_1 = i) \\
&= \pi_i \; q_i^{(1)}(O_1)
\end{aligned}$$

$$\begin{aligned}
\alpha_{t+1}(i) &= \mathsf{Pr}(O_1, \ldots, O_{t+1}, S_{t+1} = i) \\
&= \sum_j \mathsf{Pr}(O_1, \ldots, O_t, O_{t+1}, S_t = j, S_{t+1} = i) \\
&= \sum_j \mathsf{Pr}(O_1, \ldots, O_t, S_t = j) \times \mathsf{Pr}(S_{t+1} = i | S_t = j) \\
&\qquad \times \mathsf{Pr}(O_{t+1} | S_{t+1} = i) \\
&= q_i^{(t+1)}(O_{t+1}) \sum_j \alpha_t(j) \; p_{ji}^{(t)}
\end{aligned}$$

We can then get $\mathsf{Pr}(O) = \sum_i \alpha_n(i)$.

Baum-Welch algorithm (continued)

Backward equations

(an alternative, plus these are used later)

Define $\beta_t(i) = \mathsf{Pr}(O_{t+1}, \ldots, O_n | S_t = i)$.

$$\beta_n(i) \equiv 1$$

$$\begin{aligned}
\beta_{t-1}(i) &= \mathsf{Pr}(O_t, \ldots, O_n | S_{t-1} = i) \\
&= \sum_j \mathsf{Pr}(O_t, \ldots, O_n, S_t = j | S_{t-1} = i) \\
&= \sum_j \mathsf{Pr}(S_t = j | S_{t-1} = i) \times \mathsf{Pr}(O_t | S_t = j) \\
&\qquad \times \mathsf{Pr}(O_{t+1}, \ldots, O_n | S_t = j) \\
&= \sum_j \beta_t(j)\ q_j^{(t)}(O_t)\ p_{ij}^{(t-1)}
\end{aligned}$$

Note that $\mathsf{Pr}(O, S_t = i) = \alpha_t(i)\ \beta_t(i)$.

Thus $\mathsf{Pr}(O) = \sum_i \alpha_t(i)\ \beta_t(i)$, for any t.

Simulating from Pr$(S|O)$

$$\text{Note that } \Pr(S_1 = i|O) = \frac{\alpha_1(i)\ \beta_1(i)}{\sum_i \alpha_1(i)\ \beta_1(i)}$$

Also,

$$\begin{aligned}
\Pr(S_{t+1} = j|O, S_t = i) &= \frac{\Pr(O, S_t = i, S_{t+1} = j)}{\Pr(O, S_t = i)} \\
&= \frac{\alpha_t(i)\ p_{ij}^{(t)}\ q_j^{(t+1)}(O_{t+1})\ \beta_{t+1}(j)}{\alpha_t(i)\ \beta_t(i)} \\
&= \frac{p_{ij}^{(t)}\ q_j^{(t+1)}(O_{t+1})\ \beta_{t+1}(j)}{\beta_t(i)}
\end{aligned}$$

So simulating from $\Pr(S|O)$ is easy, once we've calculated the β's!

Viterbi algorithm

We wish to find $\hat{S} = \arg\max_S \Pr(S|O)$.

The following is an example of a *dynamic programming* algorithm.

Note that $\Pr(S_1 = i_1, \ldots, S_{t+1} = i_{t+1}|O) =$
$\Pr(S_1 = i_1, \ldots, S_t = i_t|O) \Pr(S_{t+1} = i_{t+1}|O, S_t = i_t)$.

(The calculation of $\Pr(S_{t+1} = i_{t+1}|O, S_t = i_t)$ was shown on the previous page.)

Define $\delta_1(i_1) = \Pr(S_1 = i_1|O)$ and

$$\delta_t(i_t) = \max_{i_1,\ldots,i_{t-1}} \Pr(S_1 = i_1, \ldots, S_t = i_t|O)$$

Note that

$$\delta_t(i_t) = \max_{i_{t-1}} \delta_{t-1}(i_{t-1}) \Pr(S_t = i_t|O, S_{t-1} = i_{t-1})$$

Define

$$\gamma_t(i_t) = \arg\max_{i_{t-1}} \delta_{t-1}(i_{t-1}) \Pr(S_t = i_t|O, S_{t-1} = i_{t-1}).$$

The algorithm:

1. Calculate $\delta_t(i_t)$ and $\gamma_t(i_t)$ for $t = 2, \ldots, n$.
2. Let $\hat{S}_n = \arg\max_{i_n} \delta_n(i_n)$.
3. Trace-back: let $\hat{S}_{t-1} = \gamma_t(\hat{S}_t)$.

Estimation

Given data on a set of N sequences $\{O_t\}$, we can obtain the MLEs of the parameters $\theta = [\pi_i, p_{ij}^{(t)}, q_i^{(t)}(O_t)]$ using a form of the EM algorithm. At each iteration we get something like the following.

$$\hat{\pi}_i = \frac{1}{N} \sum \Pr(S_1 = i|O)$$

$$\hat{p}_{ij}^{(t)} = \frac{\sum \Pr(S_t = i, S_{t+1} = j|O)}{\sum \Pr(S_t = i|O)}$$

$$\hat{q}_i^{(t)}(o) = \frac{\sum \Pr(S_t = i|O)\ 1\{O_t = o\}}{\sum \Pr(S_t = i|O)}$$

In the case of a *homogeneous* chain, we include sums over t in the numerators and denominators above.

We may get by, in the homogeneous case, with data on a single chain ($N = 1$), in which case we may assume that π is either known or is the stationary distribution corresponding to the transition matrix (p_{ij}) (or both).

The $p_{ij}^{(t)}$ and $q_i^{(t)}(\cdot)$ may be functions of some smaller number of parameters; the $q_i^{(t)}(\cdot)$ may be densities. Things are still (usually) straightforward and simple.

A practical issue

If n (the length of the $\{S_t\}$ chain) is large, α_n and β_1 will be susceptible to underflow.

Solutions:

Rabiner: Rescale the α's.

Churchill: Reformulate the problem.

Me: Calculate $\log \alpha_t$ and $\log \beta_t$ and use an `addlog` type function (as discussed in the last lecture).