HMM in crosses and small pedigrees

Lecture 8, Statistics 246,
February 17, 2004
Discrete-time Markov chains

Consider a sequence of random variables $X_1, X_2, X_3, \ldots$ with common finite state space $S$. This sequence forms a Markov chain if for all $t$, $\{X_1, \ldots, X_{t-1}\}$ and $\{X_{t+1}, X_{t+2}, \ldots\}$ are conditionally independent given $X_t$, equivalently,

$$pr(X_t \mid X_{t-1}, X_{t-2}, \ldots) = pr(X_t \mid X_{t-1}).$$

The matrix $p(i,j;t) = pr(X_t = j \mid X_{t-1} = i)$ is the transition matrix at step $t$. When $p(i,j;t) = p(i,j)$ for all $i$ and $j$, independent of $t$, we say the Markov chain is time-homogeneous, or has stationary transition probabilities. Many of the chains we’ll be meeting will be inhomogeneous, and $t$ will be in space, not time.

There are plenty of good books on elementary Markov chain theory, Feller vol 1 being my favourite, but they mostly concentrate on asymptotic behaviour in the homogeneous case. For the time being we don’t need this, or much else from the general theory, apart from the fact that multi-step transition matrices are products of 1-step transition matrices (Exercise).
Hidden Markov Models (HMM)

If \((X_t)\) is a Markov chain, and \(f\) is an arbitrary function on the state space, then \((f(X_t))\) will not in general be a Markov chain.

**Exercise**: Construct an example to demonstrate the last assertion.

It is sometimes the case that associated with a Markov chain \((X_t)\) is another process, \((Y_t)\) say, whose terms are conditionally independent given the chain \((X_t)\). This happens with so-called semi-Markov chains.

Both functions of Markov chains and this last situation are covered by the following useful definition, based on the work of L. E. Baum and colleagues around 1970. A bivariate Markov chain \((X_t, Y_t)\) is called a **Hidden Markov Model** if a) \((X_t)\) is a Markov chain, and b) the distribution of \(Y_t\) given \(X_t, X_{t-1}, X_{t-2} \ldots\) depends only on \(X_t\) and \(X_{t-1}\). In many examples, this dependence is only on \(X_t\), but in some, it can extend beyond \(X_{t-1}\), and/or include \(Y_t\). Once you see how the defining property is used in the calculations, you will get an idea of the possible extensions.

**Exercise**. Explain how functions of Markov chains are always HMM.
HMM, cont.

There are many suitable references on HMM, but two good ones for our purposes are the books by Timo Koski (HMM for bioinformatics, 2001) and Durbin et al (Biological sequence analysis, 1998).

The simplest specification of an HMM is via the transition probabilities \( p(i,j;t) \) for the underlying Markov chain \( (X_t) \), and the emission probabilities for the observations \( (Y_t) \), where these are given by

\[
q(i,j,k;t) = pr(Y_t = k \mid X_{t-1} = i, X_t = j).
\]

We also need an initial distribution \( \pi \) for the chain: \( \pi(i) = pr(X_0 = i) \).

In general we are not going to observe \( (X_t) \), which accounts for the word “hidden” in the name, but if we did, the probability of observing the state sequence \( x_0, x_1, x_2, \ldots, x_n \), and associated observations \( y_1, y_2, \ldots, y_n \) is

\[
\pi(x_0)p(x_0, x_1;1)q(x_0, x_1,y_1;1)\ldots p(x_{n-1},x_n;n )q(x_{n-1},x_n,y_n;n).
\]
HMM in experimental crosses

We are going to consider the chromosomes of offspring from crosses of inbred strains A and B of mice. Suppose that we have n markers along a chromosome, #1 say, in their correct order 1, 2, …n, say, with \( r_t \) being the recombination fraction between markers \( t \) and \( t+1 \). Consider the genotypes at these markers along chromosome 1 of an \( A \times H \) backcross mouse. Each genotype will be either \( A = aa \) or \( H = ab \), and you can do the Exercise. Under the assumption of no interference, the sequence of genotypes at markers 1, 2, …, \( n \) is a Markov chain with state space \{A, H\}, initial distribution \((1/2, 1/2)\), and 2×2 transition probability matrix \( R(r_t) \) having diagonal entries \( 1-r_t \) for no change \( (A\rightarrow A, H\rightarrow H) \), and off-diagonal entries \( r_t \) for change \( (A\rightarrow H, H\rightarrow A) \).

This Markov chain represents the crossover process along the chromosome passed by one parent, say the mother. In an \( F_2 \) intercross, \( H \times H \), there is a crossover process in both \( F_1 \) parents. If we consider the offspring’s possible ordered genotypes, that is, genotypes with known parental origin (also called known phase), then the sequence of ordered genotypes at markers 1, 2, …, \( n \) is also a Markov chain, with state space, \( \{a,b\} \otimes \{a,b\} = \{aa, ab, ba, bb\} \), initial distribution \( (1/2, 1/2) \otimes (1/2, 1/2) = (1/4, 1/4, 1/4. 1/4) \), and transition probability matrix \( P(t) = R(r_t) \otimes R(r_t) \).
HMM in experimental crosses, cont

between marker \( t \) and \( t+1 \). Here you need to know the notion of tensor product of matrices, also known as direct product or Kronecker product. This is defined in most books on matrices, but my favourite is Bellman’s book

Exercise. My notation suggests the idea of the product of two independent Markov chains. Define this notion carefully, and show we get a Markov chain.

Now, unlike in the backcross, the observed \( F_2 \) genotypes do not always tell us which parental strand had a recombination across an interval, and which didn’t, so we cannot always reconstruct the ordered genotypes.

With this 4-state Markov chain we have just 3 possible observed states, and we include some ambiguity states, giving us an observation space \( \{A, H, B, C, D, -\} \), where \( C = \{H, B\} = \textit{not} A \), \( D = \{A, H\} = \textit{not} B \), and \( - = \text{missing} \).

The resulting joint chain-observation process is an HMM, with emission probabilities which in a more general form can be written as the array on the next page. There \( \varepsilon \) is the error rate, which may in fact be marker specific, though for simplicity that is not indicated by the notation.
$F_2$ HMM emission probabilities

\[
\begin{array}{ccccccc}
X_t & A & H & B & C & D & \text{–} \\
Y_t & & & & & & \\
\begin{array}{ccccccc}
\text{aa} & 1 - \varepsilon & \frac{1}{2} \varepsilon & \frac{1}{2} \varepsilon & \varepsilon & 1 - \frac{1}{2} \varepsilon & 1 \\
\text{ab} & \frac{1}{2} \varepsilon & 1 - \varepsilon & \frac{1}{2} \varepsilon & 1 - \frac{1}{2} \varepsilon & 1 - \frac{1}{2} \varepsilon & 1 \\
\text{ba} & \frac{1}{2} \varepsilon & 1 - \varepsilon & \frac{1}{2} \varepsilon & 1 - \frac{1}{2} \varepsilon & 1 - \frac{1}{2} \varepsilon & 1 \\
\text{bb} & \frac{1}{2} \varepsilon & \frac{1}{2} \varepsilon & 1 - \varepsilon & 1 - \frac{1}{2} \varepsilon & \varepsilon & 1 \\
\end{array}
\end{array}
\]

Note that the row entries in this array do not simply sum to 1, but the entries for mutually exclusive and exhaustive cases should: A, H and B, or A and C, etc.
Calculations with our HMM

For our $F_2$ intercross there are certain calculations we would like to do which the HMM formalism makes straightforward. In fact, they are all instances of calculations generally of interest with HMM, and if the number of steps is not too large, and the state spaces not too big, there are neat algorithms for carrying them out which are worth knowing.

**Major problems**: Given the combined observations $O$ of the genotypes $Y = (Y_1, Y_2, \ldots, Y_n)$ from many different $F_2$ offspring, transition matrices $P(r_i)$, and emission probabilities $q(\varepsilon_t)$, $t=1,2,\ldots,n$,

a) calculate the log likelihood $l( r \mid O) = \log p(O \mid r)$ for different parameter values and different marker orders;

b) calculate MLE for the recombination fractions $r$ by maximizing $l( r \mid O)$; and for individual offspring with genotype data $Y$,

c) probabilistically reconstruct the individual unobserved ordered genotypes $X = (X_1, X_2, \ldots, X_n)$ by calculating $\arg\max_X p(X \mid Y, r)$.

We’ll see the algorithms in due course. We might also wish to simulate from $p(X \mid Y, r)$.
Another problem: reconstructing haplotypes

The problem here is to reconstruct the children's haplotypes as in the figure, from marker data on both the children and the parents.
Some basic notions from pedigree analysis

Founders: individuals without parents in the pedigree.
Non-founders: individuals with one or more parent in the pedigree

Calculations on pedigrees typically involve evaluating probabilities of data under specific hypotheses, perhaps going on to form likelihood ratios, or inferring unobserved genetic states, e.g. of genotypes or haplotypes.

To get started we need (in addition to the pedigree) phenotypic data, e.g. on markers or disease states, and a full genetic model. This means population genotype or allele frequencies, transmission probabilities, (probabilities of offspring haplotypes given parental haplotypes), and penetrances (probabilities of phenotypes given genotypes). Generally genotype frequencies in populations are assumed to satisfy the Hardy-Weinberg equilibrium formulae, and to be in linkage equilibrium across loci, and no interference is assumed in recombination. All of these are simplifications which have been found to be generally satisfactory.
Inheritance vectors

We plan to view our (small) pedigree as one large HMM. The states of the Markov chain along each chromosome are what we term inheritance vectors, which have a component for every meiosis in the pedigree, and entries 0 or 1. So if we have \( n \) non-founders, our inheritance vectors are binary vectors of length \( 2n \). Each non-founder has 2 components of the inheritance vector assigned to them: one for their maternal and one for their paternal meiosis. At any locus on a chromosome, the entries in the inheritance vector for a non-founder are 0 if the parental variant passed on at that locus was grandmaternal, and 1 otherwise. For individuals whose parent are founders, this definition is incomplete, but we will make a choice below. Before doing so, let me note that slightly different definitions of inheritance vector are in the literature, and in Lange’s excellent book, they are replaced by what he terms descent graphs. Clearly the idea is the same: given an inheritance vector, and allelic assignments for the founders, we can follow the descent of the alleles down the pedigree using our vector, and it is easy to envisage its graphical representation.

Consider a two-parent two-child nuclear family and suppose that the parents are \( 1/2 \) and \( 3/4 \), while the first child is \( 1/3 \) and the second is \( 2/4 \). Then the inheritance vector is of length 4, \( v = (v_{gm}, v_{gp}, v_{bm}, v_{bp}) \), where \( gm \) represents the girl’s maternal meiosis, \( gp \) her paternal meiosis, and so on.
Exercises and references

Exercise 1. Figure out whether the sequence of observed states (i.e. A, H, or B) in an $F_2$ intercross is a Markov chain. (I asserted that collapsing the state space of a Markov chain will not, in general, lead to another Markov chain. But sometime it does. Does it in this case?)

Exercise 2. Suppose that 1, 2 and 3 are markers along a backcross mouse’s chromosome in that order, and that $X_t$ denotes its genotype at marker $t$. Prove that (contrary to my assertion in class) in general $(X_1, X_2, X_3)$ will not be Markov in the incorrect order, e.g. that in general,

$$\text{pr}(X_2 = H | X_1 = A, X_3 = H) \neq \text{pr}(X_2 = H | X_1 = A),$$

which would be true if $(X_3, X_1, X_2)$ were Markov in that order (3-1-2).

References. As already mentioned, Lange’s book is a good reference for this topic applied to genetics. Closely related material can be found in my old class notes on my web site, see my Stat 260 for 1998, esp. weeks 3, 7 and 8, and weeks 3, 6 and 7 of 2000.