Inference for the parameters of HMMs

Assume an HMM model where $X_i \in \{0,1\}$, $Y_i \in \{1,2,3\}$, and we have observed both states in some stretch of data:

X	0	0	0	0	1	1	1	1	1	0
Y	1	2	1	1	2	3	2	3	3	1

- Counting transitions, we get

	0	1
0	3	1
1	1	4

and

	1	2	3
0	4	1	0
1	0	2	3

.

- In practice, we can use pseudocounts just as in the Markov chain case. In the example above, using all pseudocounts equal to 1, we get

$$P = \begin{bmatrix} 4/6 & 2/6 \\ 2/7 & 5/7 \end{bmatrix}, Q = \begin{bmatrix} 5/8 & 2/8 & 1/8 \\ 1/8 & 3/8 & 4/8 \end{bmatrix}$$

  where P is the transition matrix of the Markov chain, and Q is the stochastic matrix of transition probabilities from X_i to Y_i.
- As for Markov chains, these results can be obtained by using priors for P and Q that are products of Dirichlet distributions.

More on inference of parameters for HMMs

- The Bayesian paradigm may be used to make predictions for later observations: In the example above, with $X_0, \ldots X_9$, $Y_0, \ldots Y_9$ observed, the probability vector with the three possible values of Y_{10} can be computed with the matrix product $\mathrm{E}\left(P_{x_9}\right)\mathrm{E}\left(Q\right)$.
- The priors can be adapted to incorporate actual prior information.
- For example, prior knowledge about the transitions from states of X_i to states of Y_i might lead you to model $Y_i \sim \text{Poisson}(\lambda_{X_i})$, so for each value of X_i the Y_i are Poisson distributed with parameter λ_{X_i}. Fixing a prior also on the λ_{X_i} parameters, we may then find the posteriors for these in similar ways as we have done before.

More inference questions for HMMs (for information)

- We focused above on the case where (some) parameters of the HMM are not fully known.
- If the HMM parameters are given and the Y_i are observed, the goal may instead be to learn about the values of the X_i (these methods are not part of the course):
  - Find the sequence $X_0, \dots, X_k$ with the maximum probability given the observed $Y_0, \dots, Y_k$ and the given model: The *Viterbi algorithm*.
  - Find the marginal distribution for each X_i given the observed $Y_0, \dots, Y_k$ and the model: The Forward-Backward algorithm.
  - Find the *joint distribution* of $X_0, \dots, X_k$ given the observed $Y_0, \dots, Y_k$ and the model. In practice: Find a sequence $X_0, \dots, X_k$ that is a *sample* from this joint distribution. This may also be done with a Forward-Backward algorithm.

MVE550 2021 Lecture 6
Dobrow Chapter 4
Introduction to branching processes
Probability generating functions

Petter Mostad

Chalmers University

November 18, 2022

Introduction

- Many real phenomena can be described as developing with a tree-like structure, for example
  - Growth of cells.
  - Spread of viruses or other pathogens in a population.
  - Nuclear chain reactions.
  - Spread of funny cat videos on the internet.
  - Spread of a surname over generations.
- The process with which one node gives rise to “children” can be described as random: We will assume the probabilistic properties of this process is the same for all nodes.
- We will assume all nodes are organized into *generations*.
- We are only concerned with the size of each generation.
- How large are the generations? How much does the size vary? Will the process become *extinct*?

Branching processes

A branching process is discrete Markov chain $Z_0, Z_1, \ldots, Z_n, \ldots$ where

- the state space is the non-negative integers
- $Z_0 = 1$
- 0 is an absorbing state
- Z_n is the sum $X_1 + X_2 + \cdots + X_{Z_{n-1}}$, where the X_j are independent random non-negative integers all with the same *offspring distribution*. In other words

$$Z_n = \sum_{i=1}^{Z_{n-1}} X_i.$$

- Connecting each of the Z_n individuals in generation n with their offspring in generation $n+1$ we get a tree illustrating the branching process.
- The offspring distribution is described by the probability vector $a = (a_0, a_1, \ldots,)$ where $a_j = \Pr(X_i = j)$.
- To focus on the interesting cases we assume $a_0 > 0$ and $a_0 + a_1 < 1$.

Expected generation size

- Note that the state 0 is absorbing: This absorbtion is called *extinction*.
- As $a_0 > 0$, all nonzero states are transient.
- Define $\mu = \mathrm{E}(X_i) = \sum_{j=0}^{\infty} j a_j$ (the expected number of children).
- Then we may compute that

$$\mathrm{E}(Z_n) = \mathrm{E}\left(\sum_{i=1}^{Z_{n-1}} X_i\right) = \mathrm{E}\left(\mathrm{E}\left(\sum_{i=1}^{Z_{n-1}} X_i \mid Z_{n-1}\right)\right) = \cdots = \mathrm{E}(Z_{n-1})\,\mu.$$

- We get directly that

$$\mathrm{E}(Z_n) = \mu^n\,\mathrm{E}(Z_0) = \mu^n$$

- We subdivide Branching processes into three types:
  - *Subcritical* if $\mu < 1$. Then $\lim_{n\to\infty} \mathrm{E}(Z_n) = 0$.
  - *Critical* if $\mu = 1$. Then $\lim_{n\to\infty} \mathrm{E}(Z_n) = 1$.
  - *Supercritical* if $\mu > 1$. Then $\lim_{n\to\infty} \mathrm{E}(Z_n) = \infty$.
- We can prove that if $\lim_{n\to\infty} \mathrm{E}(Z_n) = 0$ then the probability of extinction is 1.

Variance of the generation size

- Continue with $\mu = \mathrm{E}\left(X_i\right)$ denoting the expected number of children and let $\sigma^2 = \mathrm{Var}\left(X_i\right)$ denote the variance of the number of children.
- Using the law of total variance, we get

$$
\begin{aligned}
\mathrm{Var}\left(Z_n\right) &= \mathrm{Var}\left(\mathrm{E}\left(Z_n \mid Z_{n-1}\right)\right) + \mathrm{E}\left(\mathrm{Var}\left(Z_n \mid Z_{n-1}\right)\right) \\
&= \mathrm{Var}\left(\mathrm{E}\left(\sum_{i=1}^{Z_{n-1}} X_i \mid Z_{n-1}\right)\right) + \mathrm{E}\left(\mathrm{Var}\left(\sum_{i=1}^{Z_{n-1}} X_i \mid Z_{n-1}\right)\right) \\
&= \mathrm{Var}\left(\mu Z_{n-1}\right) + \mathrm{E}\left(\sigma^2 Z_{n-1}\right) \\
&= \mu^2 \,\mathrm{Var}\left(Z_{n-1}\right) + \sigma^2 \mu^{n-1}
\end{aligned}
$$

- From this we prove by induction, for $n \geq 1$,

$$
\mathrm{Var}\left(Z_n\right) = \sigma^2 \mu^{n-1} \sum_{k=0}^{n-1} \mu^k = \begin{cases} n\sigma^2 & \text{if } \mu = 1 \\ \sigma^2 \mu^{n-1}(\mu^n - 1)/(\mu - 1) & \text{if } \mu \neq 1 \end{cases}
$$

Probability generating functions

- For *any* discrete random variable X taking values in $\{0, 1, 2, \ldots, \}$ define the probability generating function $G(s)$, or $G_X(s)$, as

$$G(s) = \mathrm{E}\left(s^X\right) = \sum_{k=0}^{\infty} s^k \Pr(X = k).$$

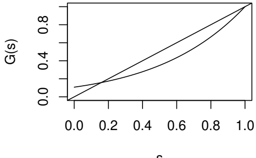
- The series converges absolutely for $|s| \leq 1$. We assume s is a real number in $[0, 1]$.
- We get a 1-1 correspondence between probability vectors on $\{0, 1, 2, \ldots, \}$ and functions represented by a series where the non-negative coefficients sum to 1.
- Specifically, if $G_X(s) = G_Y(s)$ for all s for random variables X and Y then X and Y have the same distribution.
- The correspondence of X with $G_X(s)$ provides an important and surprisingly useful computational tool.

What does $G_X(s)$ look like?

- $G_X(1) = 1$ and $G_X(0) = \Pr(X = 0)$.
- We get

$$\begin{aligned}
G'(s) &= \sum_{k=1}^{\infty} ks^{k-1} \Pr(X = k) = \mathrm{E}\left(Xs^{X-1}\right) \\
G''(s) &= \sum_{k=2}^{\infty} k(k-1)s^{k-2} \Pr(X = k) = \mathrm{E}\left(X(X-1)s^{X-2}\right) \\
G'''(s) &= \sum_{k=3}^{\infty} k(k-1)(k-2)s^{k-3} \Pr(X = k) = \mathrm{E}\left(X(X-1)(X-2)s^{X-3}\right)
\end{aligned}$$

- So the derivatives are non-negative, and $G'(s)$ and $G''(s)$ are positive for $s \in (0, 1)$.
- Below: $G_X(s)$ when $X \sim \text{Binomial}(10, 0.2)$. (Diagonal added)



Some properties of probability generating functions

- To go from X to $G_X(s)$: Compute the infinite (or finite) sum.
- To go from $G_X(s)$ to X: Use that we have

$$P(X = j) = \frac{G^{(j)}(0)}{j!}.$$

- If X and Y are independent,

$$G_{X+Y}(s) = \mathrm{E}\left(s^{X+Y}\right) = \mathrm{E}\left(s^X s^Y\right) = \mathrm{E}\left(s^X\right)\mathrm{E}\left(s^Y\right) = G_X(s)G_Y(s)$$

- $\mathrm{E}(X) = G'(1)$
- $\mathrm{E}(X(X-1)) = G''(1)$.
- As a consequence, $\mathrm{Var}(X) = G''(1) + G'(1) - G'(1)^2$.

Probability generating functions for Branching processes

Assume we have a Branching process $Z_0, Z_1, \ldots,$ with independent random variables X counting the offspring at each node.

- Write $G_n(s) = G_{Z_n}(s) = \mathrm{E}\left(s^{Z_n}\right)$ and $G(s) = G_{X_k}(s) = \mathrm{E}\left(s^{X_k}\right)$.
- We get

$$
\begin{aligned}
G_n(s) &= \mathrm{E}\left(s^{\sum_{k=1}^{Z_{n-1}} X_k}\right) = \mathrm{E}\left(\mathrm{E}\left(s^{\sum_{k=1}^{Z_{n-1}} X_k} \mid Z_{n-1}\right)\right) \\
&= \mathrm{E}\left(\mathrm{E}\left(\prod_{k=1}^{Z_{n-1}} s^{X_k} \mid Z_{n-1}\right)\right) = \mathrm{E}\left(G(s)^{Z_{n-1}}\right) = G_{n-1}(G(s)).
\end{aligned}
$$

- As $G_0(s) = \mathrm{E}\left(s^{Z_0}\right) = s$, it follows that $G_n(s) = G(G(G(\ldots G(s)\ldots)))$, with n iterations of the G function.
- This result can be applied numerically to compute $G_n(s)$, but it is even more important theoretically.

Extinction probability theorem

THEOREM

- Let G be the probability generating function for the offspring distribution for a branching process. The probability of eventual extinction is the smallest positive root of the equation $s = G(s)$.
- Thus in (subcritical and) critical cases the extinction probability is 1.

- Proof: Let e_n be the probability that the process is extinct in generation n. Then

  $$e_n = \Pr(Z_n = 0) = G_n(0) = G(G_{n-1}(0)) = G(\Pr(Z_{n-1} = 0)) = G(e_{n-1})$$

  We get for the probability of extinction

  $$e = \lim_{n\to\infty} e_n = \lim_{n\to\infty} G(e_{n-1}) = G(\lim_{n\to\infty} e_{n-1}) = G(e)$$

  so e is a root of G. Starting with any positive root x, we get $e_0 = 0 < x$ and applying the increasing function G repeatedly on both sides yields $e_n < x$, taking the limit yields $e \leq x$.