Bioinformatics Research Center
Aarhus University

Advanced Statistical Methods in Bioinformatics
Asger Hobolth
December 4, 2018

Handout 15

Exam: Basic information

The exam is a 48 hours take-home exam. The exam is ready for download on the blackboard webpage for the course in the Exam directory from 9.00 am on January 2, 2019. A printed version of your report must be handed in to me (building 1110, office 324) or Ellen Noer (building 1110, office 319), before 12.00 am on January 4, 2019.

Exam: Syllabus

Ewens and Grant (2005):

1. Finite Markov chains in Discrete Time: Section 4.5-4.10 page 161-170.
2. Hidden Markov Models: Chapter 12 page 409-429.
3. Continuous-Time Markov Chains: Section 11.7 page 403-408.
4. Evolutionary Continuous-Time Markov Models: Section 14.3 page 484-496.

All exercises and other material from the 15 handouts.

All material from the lectures.

Mandatory Projects I-III

Papers:

1. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J., and Stratton, M.R. (2013). Deciphering signatures of mutational processes operative in human cancer. *Cell Reports*, 3, 246–259.
2. Fu and Koutras (1994). Distribution theory of runs: A Markov chain approach. *Journal of the American Statistical Association*, **427**, 1050-1058.
3. Hobolth, A. and Jensen, J.L. (2005). Statistical inference in evolutionary models of DNA sequences via the EM algorithm. *Statistical Applications in Genetics and Molecular Biology*, **18**.
4. Hobolth, Guo, Kousholt and Jensen (2018). A unifying framework and comparison of algorithms for non-negative matrix factorization. *Manuscript.*

Lectures in Week 49

Parameter estimation for discretely observed continuous time Markov chains. As an example we discussed the EM algorithm for the symmetric two-state model, and in particular the R code below.

```
##------------------------------------------------------------
## Analytical estimation for symmetric bi-allelic model
##------------------------------------------------------------
## Data
n00 <- 900 ; n01 <- 100 ; n <- n00+n01
## Estimation of alpha (assume T=1 for convenience)
```

```
phat<- n01/n
alp<- -0.5*log(1-2*phat)
##---------------------------------------------------------
## E-step:
##---------------------------------------------------------
## Mean number of jumps from state 0 to state 1
## and from state 1 to state 0
## E[N(0,1)|X(0)=a,X(T)=b] or E[N(1,0)|X(0)=a,X(T)=b]
## for the two possible cases of endpoints (a,b).
## The two means are the same for endpoints (0,0)
## E[N(0,1)|X(0)=0,X(T)=0]=E[N(1,0)|X(0)=0,X(T)=0].
## Furthermore, we have
## E[N(0,1)|X(0)=0,X(T)=1]=1+E[N(1,0)|X(0)=0,X(T)=1],
## so we need to consider two cases:
## Case 1: E[N(0,1)|X(0)=0,X(T)=0]
## Case 2: E[N(0,1)|X(0)=0,X(T)=1]
##--------------------------------------------------------
## We check that the equations for the cases make sense
##--------------------------------------------------------
tm <- seq(0.01,2,len=100)
alp <- 1
## By choosing alp=1 the waiting time to the next jump is
## exponential with rate 1
## Case 1: Endpoints (0,1) end expected (0,1) jumps
mn1 <- 0.5*alp*tm*(1-exp(-2*alp*tm))/(1+exp(-2*alp*tm))
plot(tm,mn1,type="l",col="purple",lwd=2,ylim=c(0,2),
     ylab="Mean number of jumps")
abline(a=0,b=0.5,col="black")
## Case 2: Endpoints (0,1) and expected (0,1) jumps
mn2 <- 0.5*( 1 + alp*tm*(1+exp(-2*alp*tm))/(1-exp(-2*alp*tm)) )
points(tm,mn2,type="l",col="blue",lwd=2)
legend("topleft",c("Case 1: Mean of N(0,1) with endpoints (0,0)",
                   "Case 2: Mean of N(0,1) with endpoints (0,1)"),
       col=c("purple","blue"),lty=1,lwd=2,bty="n")
##-----------------------------------------------------------------------
## EM algorithm for the symmetric bi-allelic Jukes-Cantor model
##------------------------------------------------------------------------
## Starting value
alp <- 2
## Iterations
for (iter in 1:50){
  mn1 <- 0.5*alp*(1-exp(-2*alp))/(1+exp(-2*alp))
  mn2 <- 0.5*( 1+alp*(1+exp(-2*alp))/(1-exp(-2*alp)))
  alp.new<- ( n00*2*mn1 + n01*(2*mn2-1) )/n
  alp<-alp.new
  cat("Iteration:",iter,"; alpha:",alp,"\n")
}
```