Multi-state modelling of intermittently-observed data

A new Bayesian model and software `msmbayes`

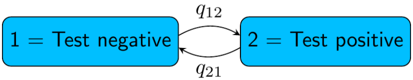
Christopher Jackson

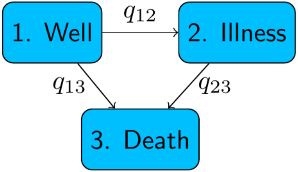
Royal Statistical Society Conference, Brighton, Sep 2024

MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

Multi-state models





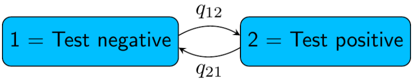
... or any other state and transition structure

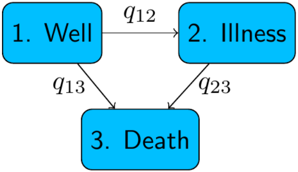
Parameters: continuous-time models with transition intensities / rates / hazards $q_{rs} = \exp(\beta_{rs}\mathbf{x})$

Estimate:, e.g.,

- expected time spent in a state (e.g. duration of an infection)
- probabilities of transition between states...

Multi-state models





... or any other state and transition structure

Parameters: continuous-time models with transition intensities / rates / hazards $q_{rs} = \exp(\beta_{rs}\mathbf{x})$

Estimate:, e.g.,

- expected time spent in a state (e.g. duration of an infection)
- probabilities of transition between states...

Data

Multi-state models get applied to a wide range of data structures

Intermittent observation: In our applications, we only know the state at a finite set of times — e.g. when person is tested for infection

Person	Time	Infection
1	0	Yes
1	2	No
1	5	No
2	1	No
2	8	Yes
...	...	...

Don't know transition times between states:

- e.g. when someone got the infection, when it cleared

Some infections may be completely unobserved for people in the data

Model estimation and challenges

Standard framework based on maximum likelihood estimation
(Kalbfleisch and Lawless, JASA 1985)

`msm` package for R (CRAN, Jackson 2011 J. Stat. Soft.) is widely used.

Strong assumptions Markov assumption: exponentially-distributed staying time in state.

- Can relax by adding latent states ("phase-type" models), however...

...Estimation can be challenging

- May be lots of parameters: transition intensities and covariate effects
- With intermittent observation, hard to tell which parameters are informed by data.
- Estimation algorithm doesn't converge if parameters not identifiable

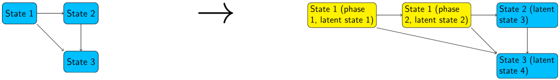
Model estimation and challenges

Standard framework based on maximum likelihood estimation (Kalbfleisch and Lawless, JASA 1985)

`msm` package for R (CRAN, Jackson 2011 J. Stat. Soft.) is widely used.

Strong assumptions Markov assumption: exponentially-distributed staying time in state.

- Can relax by adding latent states ("phase-type" models), however...



...Estimation can be challenging

- May be lots of parameters: transition intensities and covariate effects
- With intermittent observation, hard to tell which parameters are informed by data.
- Estimation algorithm doesn't converge if parameters not identifiable

Model estimation and challenges

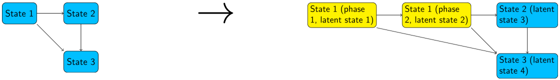
Standard framework based on maximum likelihood estimation

(Kalbfleisch and Lawless, JASA 1985)

`msm` package for R (CRAN, Jackson 2011 J. Stat. Soft.) is widely used.

Strong assumptions Markov assumption: exponentially-distributed staying time in state.

- Can relax by adding latent states (“phase-type” models), however...



...Estimation can be challenging

- May be lots of parameters: transition intensities and covariate effects
- With intermittent observation, hard to tell which parameters are informed by data.
- Estimation algorithm doesn’t converge if parameters not identifiable

Solution: Bayesian estimation

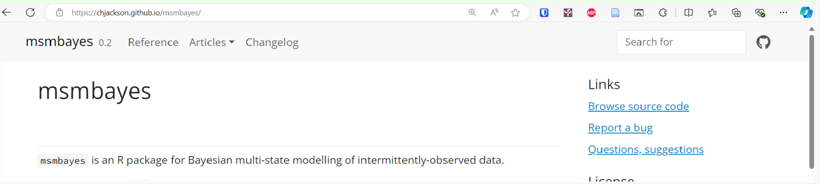
Using at least weakly informative priors

- in most scientific analyses there is background information about the thing being studied!

Advantages: Stabilises computation $\rightarrow$

- Meaningful posterior that reflects level of knowledge about parameters
- Identifies where the data are uninformative
  - if posterior is similar to prior

msmbayes R package

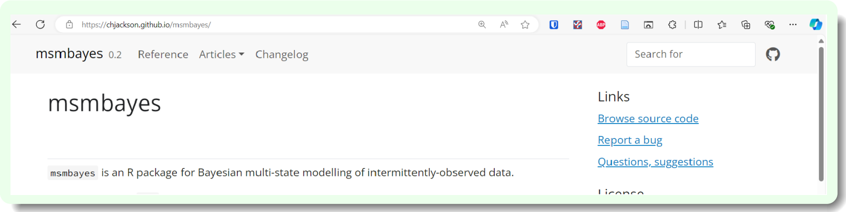


Internally uses Stan for MCMC (or faster approximations)

Familiar interface, like common R modelling packages

```
Q <- rbind(c(0, 1),
           c(1, 0)) # 2-state transition structure
priors <- list(
  msmprior("time(1,2)", median=10,  upper=30),
  msmprior("time(2,1)", median=0.5, upper=1)
)
msmbayes(data = infsim, state="state", time="months",
         subject="subject", qmatrix=q, priors=priors,...)
```

msmbayes R package



Internally uses Stan for MCMC (or faster approximations)

Familiar interface, like common R modelling packages

```
Q <- rbind(c(0, 1),
           c(1, 0)) # 2-state transition structure
priors <- list(
  msmprior("time(1,2)", median=10,  upper=30),
  msmprior("time(2,1)", median=0.5, upper=1)
)
msmbayes(data = infsim, state="state", time="months",
         subject="subject", qmatrix=q, priors=priors,...)
```

Intuitive interface to specify priors

Prior estimate and credible limits on quantities with clear interpretation

In our application, prior guess e.g.

- 10 months (up to 30 months) for mean time until next infection $1/q_{12}$
- 2 weeks (up to 1 month) for mean length of infection $1/q_{21}$

```
priors <- list(
  msmprior("time(1,2)", median=10,  upper=30),
  msmprior("time(2,1)", median=0.5, upper=1)
)
```

(log-normal priors automatically deduced)

Intuitive interface to specify priors

Prior estimate and credible limits on quantities with clear interpretation

In our application, prior guess e.g.

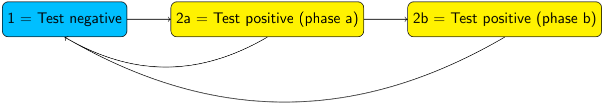
- 10 months (up to 30 months) for mean time until next infection $1/q_{12}$
- 2 weeks (up to 1 month) for mean length of infection $1/q_{21}$

```
priors <- list(
  msmprior("time(1,2)", median=10,  upper=30),
  msmprior("time(2,1)", median=0.5, upper=1)
)
```

(log-normal priors automatically deduced)

Application: estimating infection duration (simulated data)

Two latent "test positive" states → non-exponential duration distribution



MLE fails due to non-identifiability of one parameter

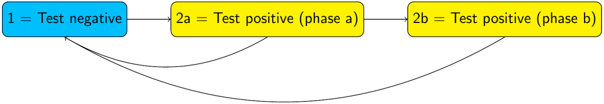
Priors and posteriors for mean times to transition

Transition rate from phase b - test negative is not identifiable (posterior close to prior)

However we still get a useful posterior for the mean infection duration (a function of these rates), which reflects this uncertainty: CI (2,30) days

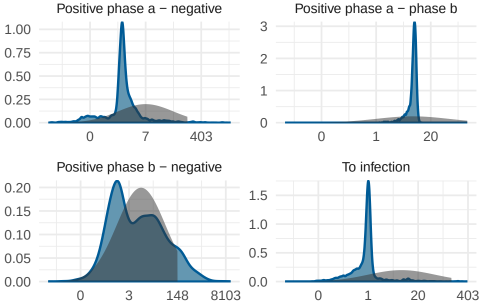
Application: estimating infection duration (simulated data)

Two latent "test positive" states → non-exponential duration distribution



MLE fails due to non-identifiability of one parameter

Priors and posteriors for mean times to transition



Transition rate from phase b - test negative is not identifiable (posterior close to prior)

However we still get a useful posterior for the mean infection duration (a function of these rates), which reflects this uncertainty: CI (2,30) days

Summary and ongoing work

Bayesian approaches improve estimation in multi-state models for intermittently observed data

Remaining challenges for "phase-type" models with latent states

- appropriate number of latent phases
- priors for "nuisance" latent transition rates

Application to cohort studies of respiratory infections in the UK

- SIREN study of healthcare workers
- COVID-19 Infection Survey

Scalability of computation: approximate Bayesian inference

Summary and ongoing work

Bayesian approaches improve estimation in multi-state models for intermittently observed data

Remaining challenges for "phase-type" models with latent states

- appropriate number of latent phases
- priors for "nuisance" latent transition rates

Application to cohort studies of respiratory infections in the UK

- SIREN study of healthcare workers
- COVID-19 Infection Survey

Scalability of computation: approximate Bayesian inference