Comparison with known results

1 Known results

The `icmstate` package can be used to non-parametrically estimate the transition intensities in interval-censored Markov multi-state models without loops. Although estimation for general multi-state models is quite novel, some results already exist for specific interval-censored multi-state models. In this vignette, we compare the estimates obtained using the `icmstate` package with the known results.

2 Frydman (1995) non-parametric estimator

Frydman [1995] has derived the non-parametric estimator for a specific illness-death model (See Figure 1). The estimator can only be used if death times are observed exactly, and for each person it is known whether they have transitioned through illness on the way to death or not.

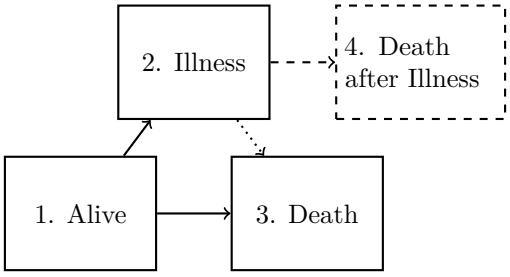


Figure 1: Graphical representation of the (extended) illness-death model. Extended: solid and dashed lines. Standard: solid and dotted lines.

The non-parametric estimator described in Frydman [1995] is available through the `msm_frydman()` function. It allows for the non-parametric estimation of a mix of cumulative distribution functions and cumulative intensity functions.

2.1 Comparison with the icmstate package

Using the `npmsm()` function it is not possible to directly derive the NPMLE for the standard illness-death model described above. We can however derive an estimate by cleverly adjusting the fitted multi-state model. If we specify a transition matrix corresponding to the illness death model, a transition from healthy to death can indicate that the subject has either passed away after experiencing illness or directly. If we consider the extended illness-death model instead (See Figure 1), then we could fit the desired model. In this case, death after illness and death are separate states, allowing us to distinguish between these two pathways.

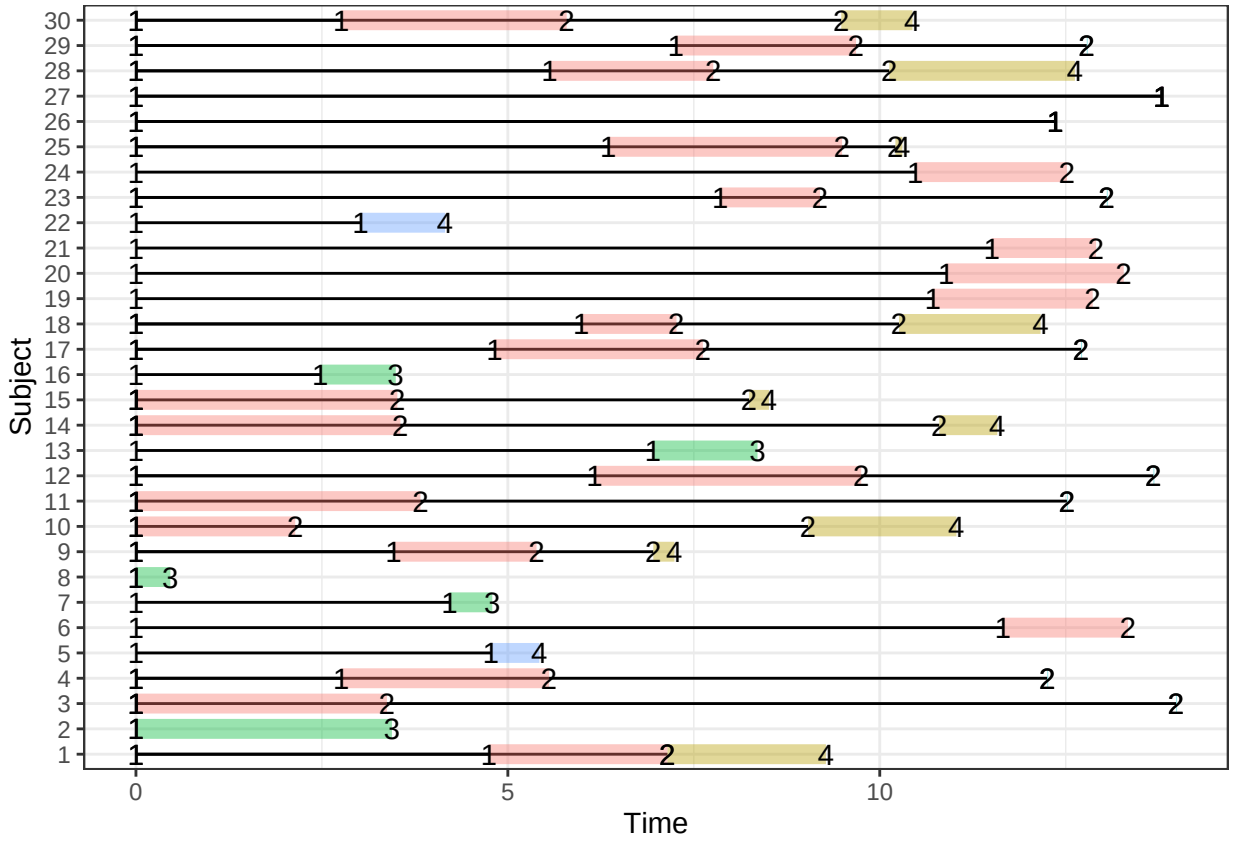
To compare the two estimators, we generate data that adheres to these requirements by simulating from an extended illness-death model:

```r
library(icmstate)
library(mstate)
#> Loading required package: survival
library(msm)
set.seed(1)
tmat_EID <- mstate::transMat(x = list( c(2, 3), c(4), c(), c() ))
qmatrix <- rbind(
    c(-0.15, 0.1, 0.05, 0),
    c(0, -0.1, 0, 0.1),
    c(0, 0, 0, 0),
    c(0, 0, 0, 0)
  )
n <- 30

#time = observation time, subject = subject identifier
simdat <- data.frame(time = c(replicate(n, c(0, seq(2, 12, by=2) + runif(6, 0, 2)))),
                     subject = rep(1:n, each = 7))
#Simulate interval-censored data. See help(simmulti.msm)
dat <- simmulti.msm(data = simdat, qmatrix = qmatrix, start = 1,
                    death = c(3,4))[, 1:3]
names(dat)[1] <- "id"
```

Let us visualise the generated data:

```r
visualise_msm(dat, tmat = tmat_EID)
```



We observe some direct $1 \to 4$ transitions. To use Frydman's result, we must have an interval for the $1 \to 2$ transition. To avoid copmlicated data processing, we simply remove subjects 5 and 22.

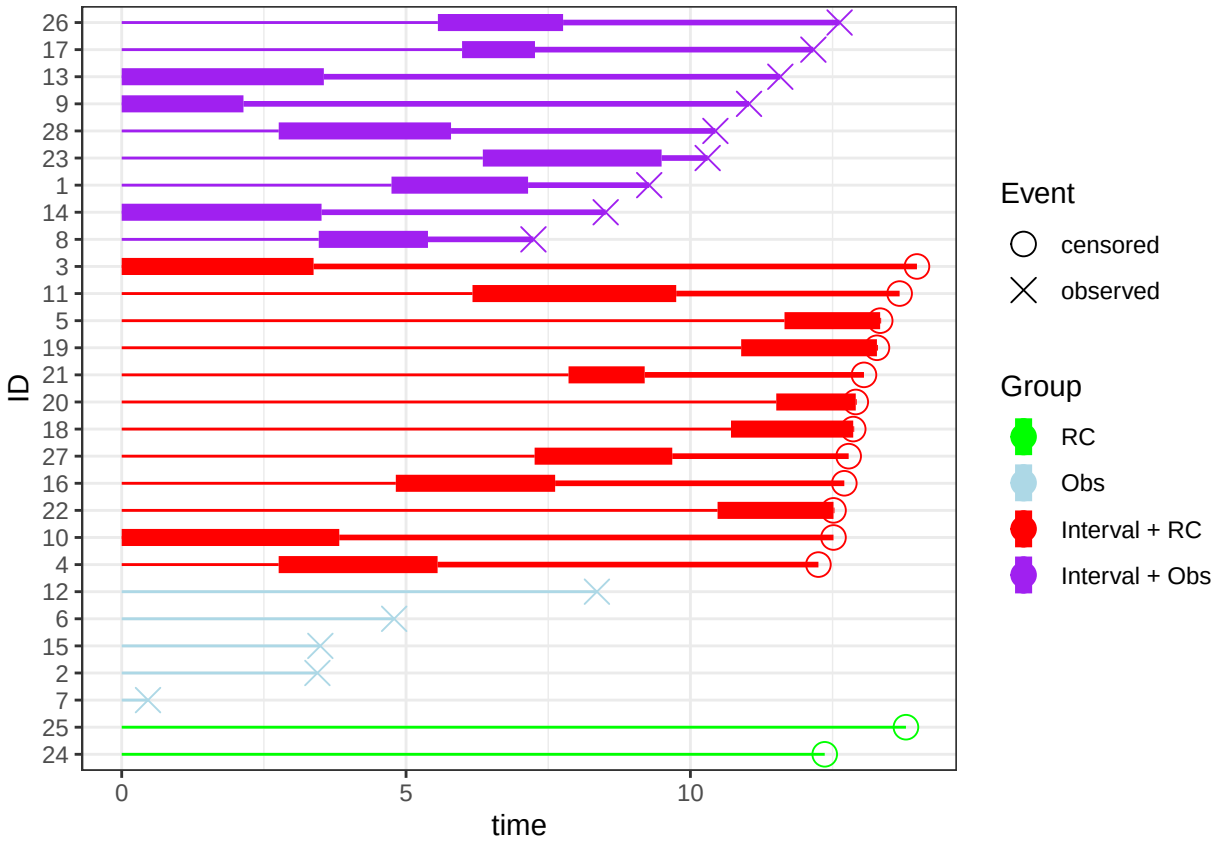
```
dat <- dat[!dat[, "id"] %in% c(5, 22),]
```

Unfortunately, the data format used by the `msm_frydman()` function differs quite a lot from the `npmsm()` function, so we will need to do some data processing anyways. To see what the data needs to look like, see `help(msm_frydman)`. In general, the data should be represented by the tuple (δ, Δ) with δ indicating whether a transition to the illness (state 2) has occurred, and Δ indicating whether a transition to a death state (state 3 or 4) has occurred. We write a function which transforms to the correct data format:

```
#Create Frydman data
msmtoFrydman <- function(gd){
  #Create Frydman data
  gd <- remove_redundant_observations(gd, tmat = tmat_EID)
  gd_frydman <- NULL
  for(j in unique(gd$id)){
    tempdat <- subset(gd, id == j)
    tempstates <- unique(tempdat$state)
    if(length(tempstates) == 1){ #If we only observe the subject in 1 state, right censored in 1
      gdi_frydman <- data.frame(delta = 0, Delta = 0,
                                L = NA,
                                R = NA,
                                time = tempdat$time[length(tempdat$time)])
    } else if(length(tempstates) == 2){ #If we only observe the subject in 2 states, either 1->2 or 1->
      if(all(tempstates %in% c(1,2))){
        gdi_frydman <- data.frame(delta = 1, Delta = 0,
                                  L = tempdat$time[which.min(tempdat$state == 1)-1],
                                  R = tempdat$time[which.min(tempdat$state == 1)],
                                  time = tempdat$time[length(tempdat$time)])
      } else if(all(tempstates %in% c(1,3))){
        gdi_frydman <- data.frame(delta = 0, Delta = 1,
                                  L = NA,
                                  R = NA,
                                  time = tempdat$time[which.min(tempdat$state == 1)])
      }
    } else if(length(tempstates) == 3){ #If we observe 3 states, then 1->2->3 must have occured
      gdi_frydman <- data.frame(delta = 1, Delta = 1,
                                L = tempdat$time[which.min(tempdat$state == 1)-1],
                                R = tempdat$time[which.min(tempdat$state == 1)],
                                time = tempdat$time[length(tempdat$time)])
    }
    gd_frydman <- rbind(gd_frydman, gdi_frydman)
  }
  return(gd_frydman)
}
```

We can then visualise the Frydman data as well using the `visualise_data()` function:

```
dat_frydman <- msmtoFrydman(dat)
visualise_data(dat_frydman)
```



And finally we can fit both models (we need to specify a tolerance for both, as they are both EM algorithms):

```
mod_npmsm <- npmsm(gd = dat, tmat = tmat_EID, maxit = 300, exact = c(3,4),
                   tol = 1e-6)
mod_frydman <- msm_frydman(data = dat_frydman, tol = 1e-6)
```

The first thing we can do is compare the support intervals (i.e. the intervals where the intensities can be non-zero) for the $1 \to 2$ transition. For this we can use the `support_npmsm()` function for the `npmsm` fit, which numerically determines on which intervals the transition intensities are estimated to be non-zero. The `msm_frydman` fit automatically returns the support intervals, as these can be determined from the theory in the article.

```
supp_npmsm <- support_npmsm(mod_npmsm, cutoff = 1e-9)
supp_frydman <- mod_frydman$supportMSM$Q_mat
print("Frydman Support")
#> [1] "Frydman Support"
supp_frydman
#>           L         R
#> 1  0.000000  2.141358
#> 2  2.760988  3.374046
#> 3  3.464627  3.514174
#> 4  4.820168  5.385463
#> 5  5.559970  5.790891
#> 6  6.350254  7.145707
#> 7  7.260828  7.266987
#> 8  8.510179  9.196185
#> 9  9.276002  9.493397
```

```
#> 10 12.361733 12.516332
#> 11 12.623554 12.862947
#> 12 13.790188       Inf
print("icmstate Support")
#> [1] "icmstate Support"
supp_npmsm$`State 1 -> State 2`$support
#>              L         R         dA
#> [1,]  0.000000  2.141358 0.10607498
#> [2,]  2.760988  3.374046 0.05852194
#> [3,]  3.464627  3.514174 0.15038906
#> [4,]  4.820168  5.385463 0.07107790
#> [5,]  6.350254  7.145707 0.17411522
#> [6,]  7.260828  7.266987 0.22807737
#> [7,]  8.510179  9.196185 0.24185272
#> [8,] 12.361733 12.516332 1.13222010
```

The support sets from Frydman [1995] do not necessarily all need to contain non-zero intensities, as some of them can still be zero. We therefore would like to compare the two estimators on their estimated survival functions. However, both models estimate different quantities. Where the `npmsm()` fit estimates transition intensities and can recover transition probabilities, the `msm_frydman()` fit estimates quite specific quantities. Let S be the (unobserved) entry time into state 2, and T the entry time into state 3 and finally $V = T - S$. Then the `msm_frydman()` function estimates the following quantities:

- $F_{12}(s) = \mathbb{P}(S \leq S, \delta = 1)$
- $F_{13}(s) = \mathbb{P}(S \leq s, \delta = 0)$
- $\Lambda_{23}(u) = \mathbb{P}(T = u | T \geq u, \delta = 1)$

From these quantities we can then also recover

- $F(s) = \mathbb{P}(S \leq s) = F_{12}(s) + F_{13}(s)$
- $G_s(v) = \mathbb{P}(V > v | S = s, \delta = 1) = \prod_{x<u\leq s+v}(1 - \Lambda_{23}(\{u\}))$

It can be shown that the estimators from Frydman [1995] can be compared with the transition intensities/probabilities in the following way:

MSM	Frydman (95)	Support
$P_{11}(0, s)$	$1 - F(s) = 1 - F_{12}(s) - F_{13}(s)$	$\mathcal{S}_{12} \cup \mathcal{S}_{13}$
$P_{13}(0, s)$	$F_{13}(s)$	$\mathcal{S}_{13}$
$P_{13}(0, s) + P_{11}(0, s)$	$1 - F_{12}(s)$	$\mathcal{S}_{12}$
$P_{24}(s, s + v) = \prod_{s<\tau_k\leq s+v}(1 - \alpha_{24}^k)$	$G_s(v) = \prod_{s<u\leq s+v}(1 - \Lambda_{23}(\{u\}))$	$\mathcal{S}_{23}$

The support column indicates that we can only compare the quantities on the right endpoints of the corresponding support sets. The subscript indicates the transition of the support set. As an example, if the support set $\mathcal{S}_{12} = \{(1, 2], (2.5, 3]\}$ then the comparison of the quantities $P_{13}(0, s) + P_{11}(0, s)$ and $1 - F_{12}(s)$ can only be made on the right endpoints 2 and 3. We determine the relevant right-endpoints:

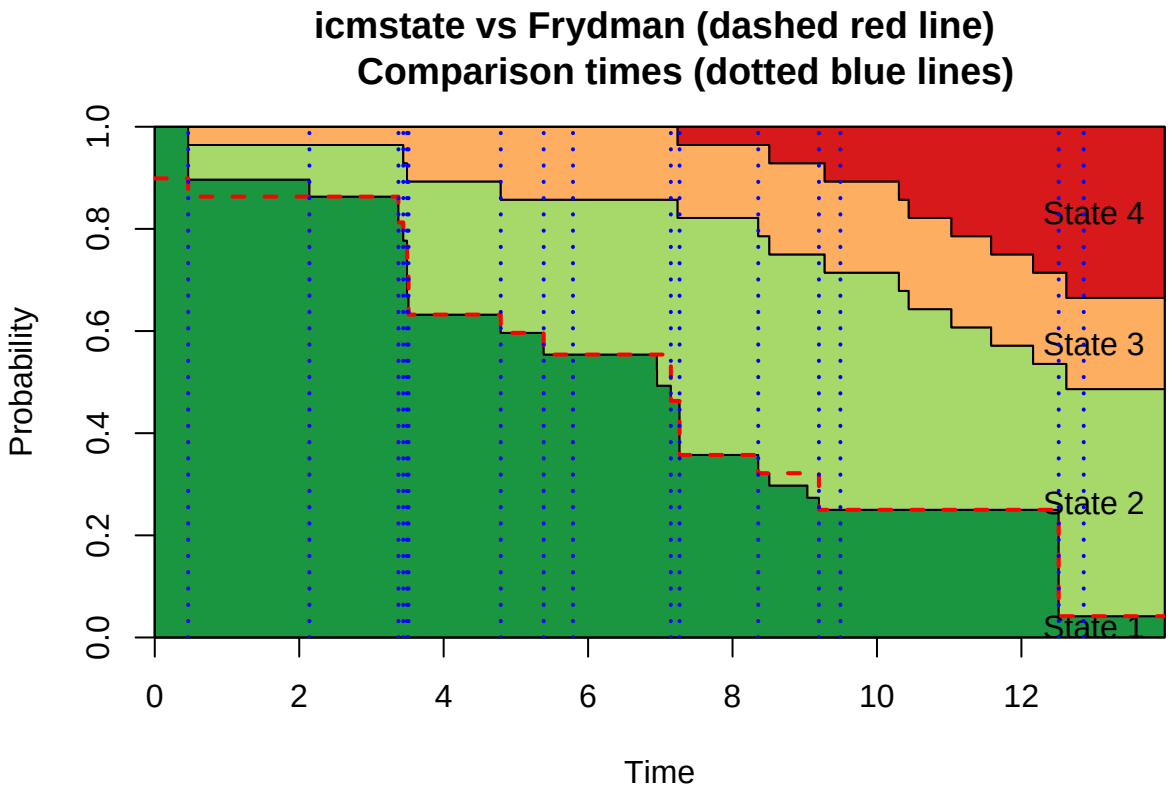
```
#Right-endpoints of the 1->2 transition
RE12 <- supp_frydman[, 2]
#Right-endpoints of the 1->3 transition
RE13 <- mod_frydman$data_idx$e_k_star
#Right-endpoints of the 2->3 transition
RE23 <- mod_frydman$data_idx$t_n_star
```

Let us perform the comparison.

2.1.1 Survival in state 1

We compare $P_{11}(0, s)$ with $1 - F(s) = 1 - F_{12}(s) - F_{13}(s)$. Transition probabilities can be recovered using the `transprob()` function in the package. The cdf of the $1 \to 2$ and $1 \to 3$ transition are contained in the `msm_frydman()` fit through the `$cdf` list. Note that the output of the cdf for the $1 \to 2$ transition contains both the upper and lower value of the cdf, as we do not know where exactly the cdf makes the jump in the support intervals. Additionally, we can only compare the values with each other on the right-endpoints of the support intervals.

```
#Transition probabilities from state 1 from time 0
P11 <- transprob(mod_npmsm, predt = 0)
#Extract times of interest
times1 <- P11[[1]][,1]
#1-F(s) for Frydman estimator:
#We take min(F_{12}(x)) as the cdf has only jumped at the right-endpoints
Frydman1minF <- sapply(times1, function(x) 1- (mod_frydman$cdf$F13(x) +
                                                 min(mod_frydman$cdf$F12(x))))
#Comparison plot
plot(P11, main = "icmstate vs Frydman (dashed red line)
     Comparison times (dotted blue lines)")
lines(times1, Frydman1minF, col = "red", type = "s", lwd = 2, lty = 2)
abline(v = unique(c(RE12, RE13)), col = "blue", lwd = 2, lty = 3)
```

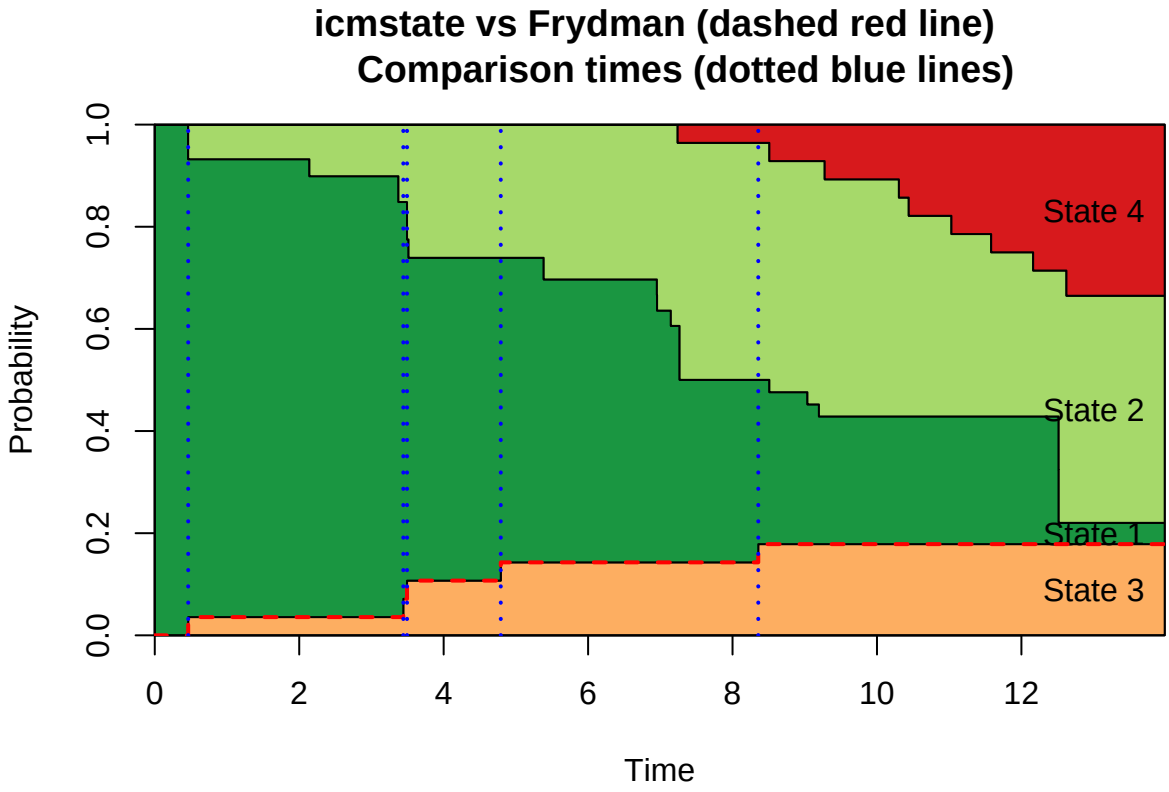


We can see that the quantities correspond on the right-endpoints.

2.1.2 Transition to death

We perform a similar comparison between $P_{13}(0, 2)$ and $F_{13}(s)$.

```
FrydmanF13 <- sapply(times1, function(x) mod_frydman$cdf$F13(x))
#Comparison plot
plot(P11, main = "icmstate vs Frydman (dashed red line)
     Comparison times (dotted blue lines)", ord = c(3, 1, 2, 4))
lines(times1, FrydmanF13, col = "red", type = "s", lwd = 2, lty = 2)
abline(v = unique(RE13), col = "blue", lwd = 2, lty = 3)
```

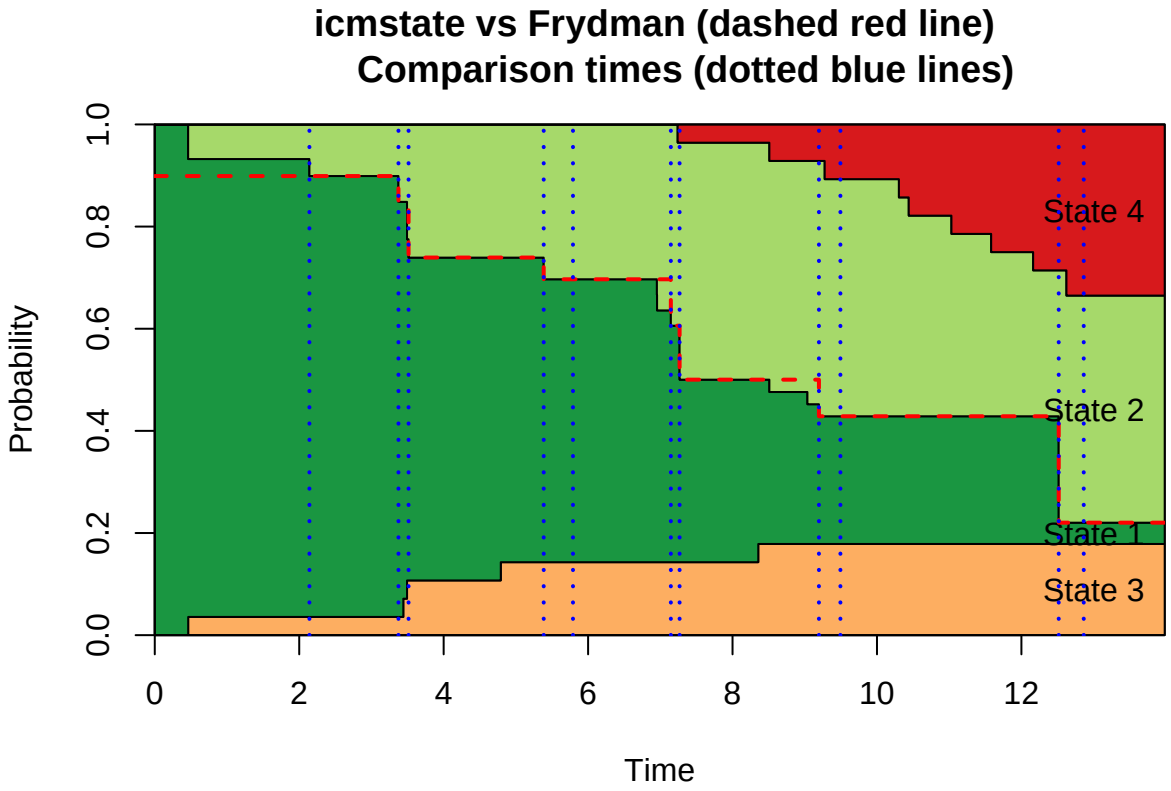


Again, the two estimators align.

2.1.3 Not catching illness

We compare $P_{13}(0,s) + P_{11}(0,s)$ with $1 - F_{12}(s)$.

```
FrydmanF12 <- sapply(times1, function(x) min(mod_frydman$cdf$F12(x)))
#Comparison plot
plot(P11, main = "icmstate vs Frydman (dashed red line)
     Comparison times (dotted blue lines)", ord = c(3, 1, 2, 4))
lines(times1, 1-FrydmanF12, col = "red", type = "s", lwd = 2, lty = 2)
abline(v = unique(RE12), col = "blue", lwd = 2, lty = 3)
```

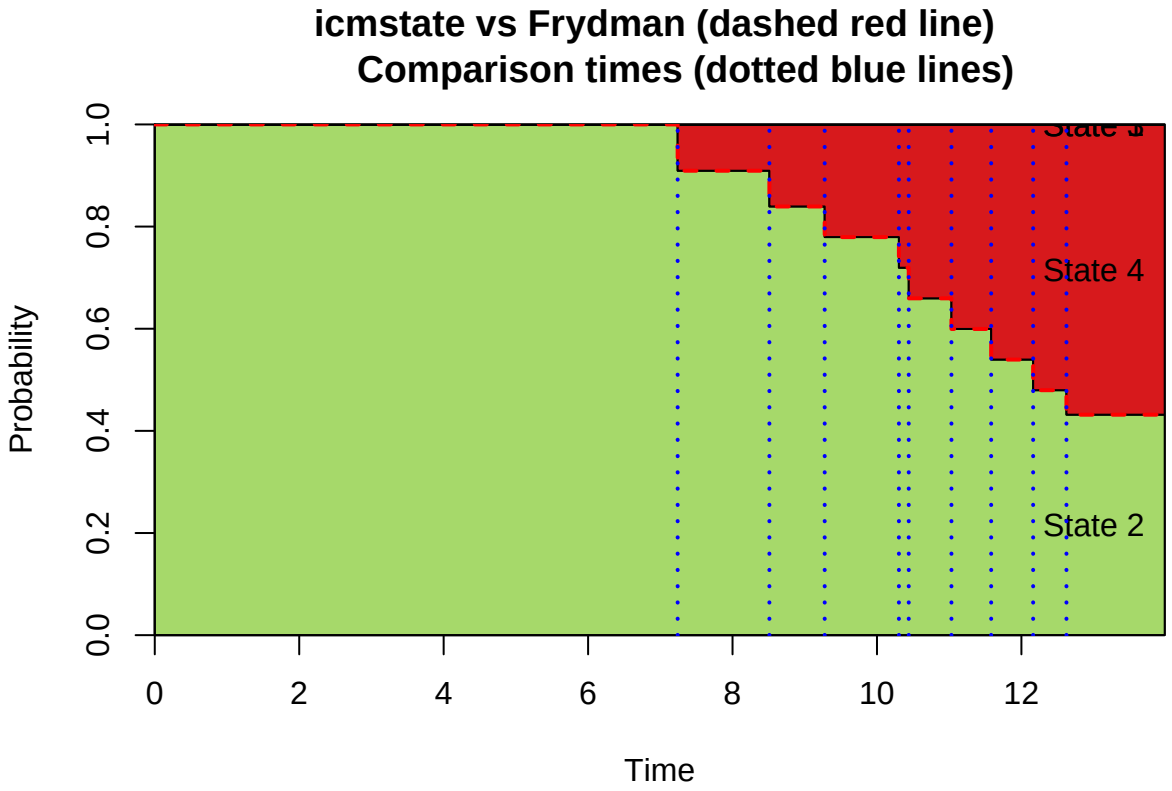


The estimate coincide!

2.1.4 Dying after illness

We compare $P_{24}(0,s) = \prod_{0<\tau_k\leq s}(1-\alpha_{24}^k)$ with $G_s(v) = \prod_{0<u\leq s}(1-\Lambda_{23}(\{u\}))$.

```
#Calculate dA23
FrydmandA23 <- c(0, diff(sapply(times1, function(x) mod_frydman$cdf$Lambda23(x))))
#We calculate the product integral for the Frydman estimator
FrydmanG <- cumprod(1-FrydmandA23)
#Comparison plot
plot(P11, main = "icmstate vs Frydman (dashed red line)
     Comparison times (dotted blue lines)", from = 2, ord = c(2, 4, 1, 3))
lines(times1, FrydmanG, col = "red", type = "s", lwd = 2, lty = 2)
abline(v = unique(RE23), col = "blue", lwd = 2, lty = 3)
```



And they coincide again!

References

H. Frydman. Nonparametric estimation of a Markov 'illness-death' process from interval- censored observations, with application to diabetes survival data. *Biometrika*, 82(4):773, December 1995. ISSN 0006-3444. doi: 10.2307/2337344.