Understanding product integration.

A talk about teaching survival analysis.

Jan Beyersmann, Arthur Allignol, Martin Schumacher. Freiburg, Germany
DFG Research Unit FOR 534
jan@fdm.uni-freiburg.de

- It is product integration that switches from hazards to probabilities.
- Product integration is not unusually difficult, but notoriously neglected.
- This talk: Use R for approaching product integration.
- One R function for approximating the true survival function and for computing Kaplan-Meier.
- Generalizes to more complex models; e.g. useful for numerical approximation and simulation with time-dependent covariates.

Survival analysis is hazard-based.

- Survival time T, censoring time C: $T \wedge C$, $\mathbf{1}(T \leq C)$

- The hazard is 'undisturbed' by censoring: cumulative hazard $A(t)$, hazard $A(\mathrm{d}t) = P(T \in \mathrm{d}t \,|\, T \geq t) = P(T \wedge C \in \mathrm{d}t, T \leq C \,|\, T \wedge C \geq t)$

- $A(\mathrm{d}t)$ estimated by increments of the Nelson-Aalen estimator:

$$\widehat{A}(\mathrm{d}t) = \frac{\#\text{ observed alive} \rightarrow \text{dead transitions at } t}{\#\text{ observed to be alive just prior } t}$$

- Kaplan-Meier is a deterministic function of the Nelson-Aalen estimator $\int \widehat{A}(\mathrm{d}t)$, and we have

$$\prod_{t_i \leq t} \left(1 - \widehat{A}(\mathrm{d}t_i)\right) \xrightarrow{P} \exp\left(-\int_0^t A(\mathrm{d}u)\right) = P(T > t)$$

- The convergence statement is not very intuitive.

Product integration π

- Recall $A(\mathrm{d}u) = P(T < u + \mathrm{d}u \,|\, T \geq u)$.

  $\Rightarrow 1 - A(\mathrm{d}u) = P(T \geq u + \mathrm{d}u \,|\, T \geq u)$

- Survival function $P(T > t) = P(T \geq t + \mathrm{d}t)$ should be an infinite product over $[0, t]$ of $1 - A(\mathrm{d}u)$-terms:

$$\begin{aligned} S(t) &= \pi_0^t \, (1 - A(\mathrm{d}u)) \\ &\approx \prod_{k=1}^{K} (1 - \Delta A(t_k)) \approx \prod_{k=1}^{K} P(T > t_k \,|\, T > t_{k-1}), \end{aligned}$$

  for a partition (t_k) of $[0, t]$

- $P(T > t) = \exp\left(-\int_0^t A(\mathrm{d}u)\right)$: solution of a product integral.

- Kaplan-Meier is a product integral of the empirical hazards.

- Roadmap:
  - Check this via R.
  - Use exactly the same code for true survival function and Kaplan-Meier.

A simple R function for product integration

- Pass partition of $[0,t]$ and cumulative hazard to `prodint`

```
prodint <- function(time.points,A){
  prod(1-diff(apply(X=matrix(times), MARGIN=1, FUN=A)))
}
```

- E.g. exponential distribution with cumulative hazard $A(t) = 0.9 \cdot t$

  `A.exp <- function(time.point){return(0.9*time.point)}`

  on the time interval $[0,1]$:

```
> times <- seq(0,1,0.001)
> prodint(times,A.exp);exp(-0.9*max(times))
[1] 0.4064049
[1] 0.4065697
```

- The vector of time points does not have to be equally spaced:

```
> prodint(runif(n=1000, min=0, max=1), A.exp)
[1] 0.4063475
```

- Conclusion: $\prod_{k=1}^{K}(1 - \Delta A(t_k))$ approaches $S(t)$ and we write $\prod_0^t (1 - \mathrm{d}A(u))$ for the limit.

- Can be tailored to return a survival **function**.

From Nelson-Aalen to Kaplan-Meier via product integration

- Recall: empirical hazard

$$\widehat{A}(\mathrm{d}t) = \frac{\#\text{ observed alive} \to \text{dead transitions at } t}{\#\text{ observed to be alive just prior } t}$$

- Nelson-Aalen estimator $\int \widehat{A}(\mathrm{d}t)$ of the cumulative hazard.

- Kaplan-Meier is the product integral of one minus Nelson-Aalen:

$$\widehat{S}(t) = \prod_0^t \Big(1 - \widehat{A}(\mathrm{d}u)\Big) = \prod_{t_k \le t} \Big(1 - \widehat{A}(\mathrm{d}t_k)\Big)$$

- Continuous mapping theorem:

$$\widehat{S}(t) = \prod_0^t \Big(1 - \widehat{A}(\mathrm{d}u)\Big) \xrightarrow{P} \prod_0^t (1 - A(\mathrm{d}u)) = S(t)$$

- Kaplan-Meier can be computed by `prodint` applied to $\int \widehat{A}(\mathrm{d}t)$.

`prodint` computes Kaplan-Meier.

- 100 event times $\sim \exp 0.9$: `event.times <- rexp(100,0.9)`

- 100 censoring times `cens.times` $\sim u[0,5]$: `runif(100,0,5)`

- Observed times `obs.times <- pmin(event.times, cens.times)`

  About 24% of the observations censored.

- Compute Nelson-Aalen with `mvna` or

```
fit.surv <- survfit(Surv(obs.times,c(event.times<=cens.times)))
A <- function(time.point){
  sum(fit.surv$n.event[fit.surv$time <= time.point]/
      fit.surv$n.risk[fit.surv$time <= time.point])
}
```

  and estimate the survival function at, e.g., time 1

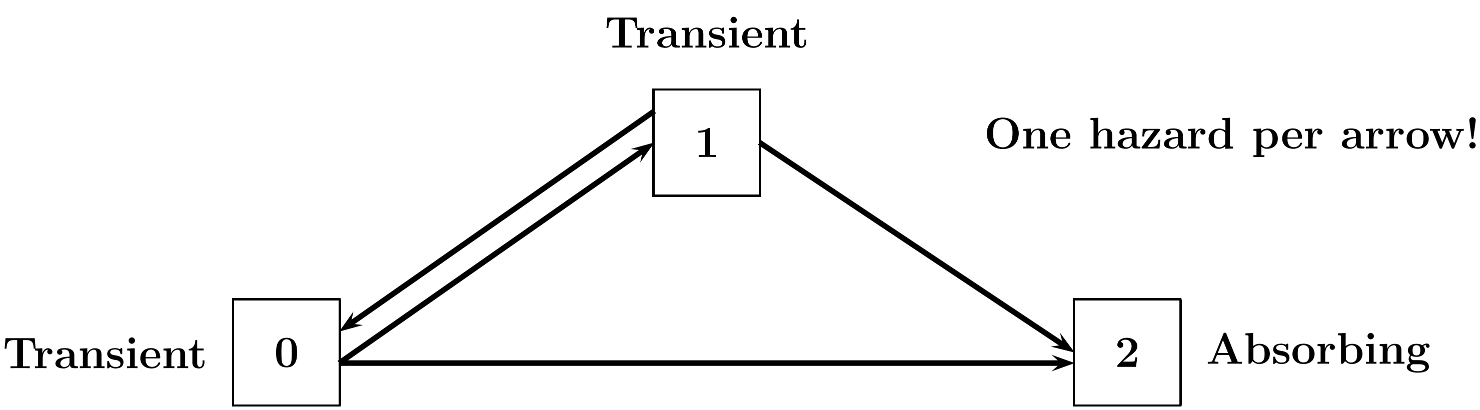
```
> prodint(obs.times[obs.times<=1],A)
[1] 0.4370994
```

- Value of `fit.surv$surv` for time 1 is 0.4370994.

Why is product integration useful?

- Survival analysis is hazard-based.
- It is product integration that recovers both the underlying and the empirical distribution function.
- Properties of Nelson-Aalen estimator are easiest to study.
- Properties of product integration (continuity, Hadamard-differentiability) allow to transfer results to Kaplan-Meier: consistency, asymptotic distribution.
- Generalizes to quite complex models where Kaplan-Meier and the exp($-$cumulative hazard)-formula fail, but are often erroneously applied.

Matrix-valued product integration for multivariate hazards.



- Closed formulae for transition probabilities usually not available.
- Can be approximated using product integration.
- Can be estimated by applying product integration to multivariate Nelson-Aalen: Aalen-Johansen.
- R: packages `mvna`, `etm`, matrix-valued function `prodint`
- E.g. useful for time-dependent covariates: estimation, simulation.
- Standard assumptions: time-inhomogeneous Markov or random censoring.

A brief summary and some references

- Move from hazards to probabilities thru product integration both in the modelling and the empirical world.
- We can and should do this teaching survival analysis.
- Works in more complex models (incl. competing risks), avoiding hypothetical quantities.

- R. Gill and S. Johansen. A survey of product-integration with a view towards application in survival analysis. *Annals of Statistics*, 18(4):1501–1555, 1990.
- O. Aalen and S. Johansen, An empirical transition matrix for non-homogeneous Markov chains based on censored observations, *Scand J Stat* vol. 5 pp. 141–150, 1978.
- P. Andersen, Ø. Borgan, R. Gill, and N. Keiding.*Statistical models based on counting processes.* Springer, 1993.
- J. Beyersmann, T. Gerds, and M. Schumacher. Letter to the editor: comment on 'Illustrating the impact of a time-varying covariate with an extended Kaplan-Meier estimator' by Steven Snapinn, Qi Jiang, and Boris Iglewicz in the November 2005 issue of The American Statistician. *The American Statistician*, 60(30):295–296, 2006.
- Arthur's talk on `mvna`.