

The use of right-censored data in nonparametric predictive inference

F.P.A. Coolen & K.J. Yan

Dept. of Mathematical Sciences, University of Durham
Durham, DH1 3LE, England
Frank.Coolen@durham.ac.uk

Abstract

We illustrate how nonparametric predictive inference can deal with data including right-censoring times. We present lower and upper bounds for the survival function of a future individual, based on data including event times and right-censoring times. Our results are illustrated via an example, and compared with the Berliner-Hill and Kaplan-Meier methods.

1 Introduction

Berliner and Hill (1988) present a method for nonparametric survival analysis, using the assumption $A_{(n)}$ introduced by Hill (1968), to derive predictive probabilities for a future observation on the basis of data including right-censored observations. $A_{(n)}$ defines probabilities for a future random quantity, based on observations $x_1 < \dots < x_n$, as $P(X_{n+1} \in (x_j, x_{j+1})) = 1/(n+1)$, for $j = 0, \dots, n$, where $x_0 = 0$ and $x_{n+1} = \infty$, assuming the random quantities are non-negative ('lifetimes'). This is a post-data assumption, related to exchangeability. It implies that the rank of the next observation has equal probability to be any value from 1 to $n+1$. $A_{(n)}$ has been justified in the literature, and is appropriate for situations with very little information in addition to the data, or if one does not wish to use additional information. A brief overview with references is given by Coolen and van der Laan (2001), who give an example of nonparametric predictive inference based on $A_{(n)}$. $A_{(n)}$ does not provide precise probabilities for most events of interest, but it can be used to derive bounds for probabilities, which have strong consistency properties in the theory of interval probability (Augustin and Coolen 2001, Weichselberger 2001).

Let the data consist of u event times, $0 < t_1 < \dots < t_u$, and $v = n - u$ right-censoring times, $0 < c_1 < \dots < c_v$. Let $t_0 = 0$ and $t_{u+1} = \infty$, and let the right-censoring times in (t_i, t_{i+1}) be $c_1^i < \dots < c_{v_i}^i$. We assume that there are no ties among the data, the method is easily adapted for ties (Coolen and Yan 2002). Let n_t be the number of individuals with observation time greater than t . We call this the number of individuals 'at risk' at time t , at an observation time the corresponding individual is not included in n_t . We denote the number of individuals at risk just prior to t by \tilde{n}_t , so $\tilde{n}_t = n_t + 1$ if t is an observation time in the data, else $\tilde{n}_t = n_t$.

Berliner and Hill (1988) discuss that, when solely using $A_{(n)}$, one cannot make use of exact censoring information, but can get good approximate results by replacing each right-censoring time c_r by the information that the censored individual survived the largest observed event time smaller than c_r . Their inferences are in terms of probabilities for $X_{n+1} \in (t_i, t_{i+1})$, as predictive survival function they suggest the probability mass for X_{n+1} in (t_i, t_{i+1}) to be uniformly distributed. We call this the 'uniform Berliner-Hill' survival function.

Coolen and Yan (2002) generalize $A_{(n)}$ to allow the exact censoring information to be taken into account, they call this 'right-censoring $A_{(n)}$ ' ('rc- $A_{(n)}$ '). They derive optimal bounds for the survival function for X_{n+1} corresponding to rc- $A_{(n)}$. These results are summarized here, and an example is given to compare this approach with the Berliner-Hill method and the product-limit estimator by Kaplan and Meier (1958), which is fundamentally different as it provides an estimate of the population survival function instead of a predictive survival function. Further details are given in Coolen and Yan (2002). Yan (2002) applies rc- $A_{(n)}$ -based inference to compare two groups of lifetime data and for life-tables.

2 Right-censoring $A_{(n)}$

To describe rc- $A_{(n)}$, we need to introduce notation for partial specification of probability distributions, which we call M -function (Coolen and Yan 2002).

Definition (M -function)

A partial specification of a probability distribution for a real-valued random quantity T can be provided via probability masses assigned to intervals, without any further restriction on the spread of the probability mass within each interval. A probability mass assigned, in such a way, to an interval (a, b) , is denoted by $M_T(a, b)$, and referred to as M -function value for T on (a, b) .

Clearly, all M -function values for T on all intervals should sum up to one, each M -function value should be in $[0, 1]$, and $A_{(n)}$ implies $M_{X_{n+1}}(x_j, x_{j+1}) = 1/(n+1)$ for $j = 0, \dots, n$.

Definition (rc- $A_{(n)}$)

The assumption ‘right-censoring $A_{(n)}$ ’ (rc- $A_{(n)}$) is that the probability distribution for a nonnegative random quantity X_{n+1} , based on u event times and v right-censoring times, as described above, is partially specified by ($i = 0, \dots, u$, $k = 1, \dots, l_i$, with $t_0 = 0$ and $t_{u+1} = \infty$)

$$M_{X_{n+1}}(t_i, t_{i+1}) = \frac{1}{n+1} \prod_{\{r: c_r < t_i\}} \frac{\tilde{n}_{c_r} + 1}{\tilde{n}_{c_r}},$$

$$M_{X_{n+1}}(c_k^i, t_{i+1}) = \frac{1}{(n+1)\tilde{n}_{c_k^i}} \prod_{\{r: c_r < c_k^i\}} \frac{\tilde{n}_{c_r} + 1}{\tilde{n}_{c_r}}.$$

The product terms are defined as one if the product is taken over an empty set. The M -function values for X_{n+1} on other intervals are zero. This implicitly assumes non-informative censoring, as a post-data assumption related to exchangeability, at any time t , of all individuals known to be at risk at t , see Coolen and Yan (2002), who also justify rc- $A_{(n)}$. If there are no censorings then rc- $A_{(n)}$ is identical to $A_{(n)}$.

The rc- $A_{(n)}$ -based partially specified probability distribution for X_{n+1} enables derivation of bounds of probabilities for events concerning X_{n+1} , following the same procedure as described for $A_{(n)}$ by Augustin and Coolen (2001). The maximum lower bound for $P(X_{n+1} \in B)$ is derived by summing all M -function values for X_{n+1} on intervals that are completely within B . We denote this lower bound, which is a lower probability (Weichselberger 2001), by $\underline{P}(X_{n+1} \in B)$. The minimum upper bound is derived by summing all M -function values for X_{n+1} on intervals that have non-empty intersection with B . This upper probability is denoted by $\overline{P}(X_{n+1} \in B)$. When only assuming rc- $A_{(n)}$, it can only be deduced that the probability for this event is between these lower and upper probabilities.

3 Lower and upper survival functions

We now consider the survival function $S(t) = P(X_{n+1} \in (t, \infty))$. The rc- $A_{(n)}$ -based lower (\underline{S}) and upper (\overline{S}) survival functions for X_{n+1} , are derived as follows (Coolen and Yan 2002). At observed event times,

$$\underline{S}_{X_{n+1}}(t_i) = \overline{S}_{X_{n+1}}(t_i) = \sum_{j=i}^u P(X_{n+1} \in (t_j, t_{j+1})), \quad \text{for } i = 0, 1, \dots, u,$$

where, for $i = 0, \dots, u$,

$$P(X_{n+1} \in (t_i, t_{i+1})) = M_{X_{n+1}}(t_i, t_{i+1}) + \sum_{k=1}^{l_i} M_{X_{n+1}}(c_k^i, t_{i+1}) = \frac{1}{n+1} \prod_{\{r: c_r < t_{i+1}\}} \frac{\tilde{n}_{c_r} + 1}{\tilde{n}_{c_r}}.$$

Computation of these probabilities is simplified by (for $i = 1, \dots, u$)

$$P(X_{n+1} \in (t_i, t_{i+1})) = P(X_{n+1} \in (t_{i-1}, t_i)) \times \left[\frac{\tilde{n}_{c_i^i} + 1}{\tilde{n}_{c_i^i}} \right].$$

So, at observed event times, rc- $A_{(n)}$ -based lower and upper survival functions for X_{n+1} are equal. At other times these functions are (for $i = 0, 1, \dots, u$),

$$\begin{aligned} \overline{S}_{X_{n+1}}(t) &= \overline{S}_{X_{n+1}}(t_i), \text{ for all } t \in [t_i, t_{i+1}), \\ \underline{S}_{X_{n+1}}(t) &= \sum_{j=i+1}^u P(X_{n+1} \in (t_j, t_{j+1})) + \sum_{\{k: c_k^i \geq t\}} M_{X_{n+1}}(c_k^i, t_{i+1}), \text{ for all } t \in (t_i, t_{i+1}]. \end{aligned}$$

The upper survival function is constant between event times, and decreases at t_i by $P(X_{n+1} \in (t_{i-1}, t_i))$. On $[0, t_1)$ it is equal to one, while on $[t_u, \infty)$ it is a positive constant, which is a consequence of the fact that no further assumptions are added to the data. The lower survival function decreases at each observation, so also at censoring times, and is zero beyond the largest observation, both if this is an event or censoring time. Calculation can be slightly simplified by (for $i = 0, 1, \dots, u$)

$$\underline{S}_{X_{n+1}}(t) = \overline{S}_{X_{n+1}}(t_{i+1}) + \sum_{\{k: c_k^i \geq t\}} M_{X_{n+1}}(c_k^i, t_{i+1}), \text{ for all } t \in (t_i, t_{i+1}).$$

The effect of a censoring at c_r is increased difference between upper and lower survival functions beyond c_r . In interval probability theory, this difference is often inversely related to the amount of information on which such bounds are based, so censoring can be regarded as loss of information.

The rc- $A_{(n)}$ -based $P(X_{n+1} \in (t_i, t_{i+1}))$ are identical to those in the Berliner-Hill method (Coolen and Yan 2002), the only difference between these methods is within such intervals. The uniform Berliner-Hill survival function can be less than the lower survival function for some t , e.g. if there are multiple censorings close to t_{i+1} in (t_i, t_{i+1}) .

Hill (1992) compared the Berliner-Hill method with the Kaplan-Meier method, which, although its explicit inferential aim is quite different, namely estimation of the underlying population survival function instead of prediction for one future individual, turn out to be pretty similar. The Berliner-Hill method gives more mass to the upper tail of the distribution than the Kaplan-Meier method. Comparison of our rc- $A_{(n)}$ -based method with Kaplan-Meier leads to the identical conclusion, as we get the same probabilities between observed event times as the Berliner-Hill method.

4 Example

To illustrate rc- $A_{(n)}$ based lower and upper survival functions, we use survival data (in days, t^* denotes right-censoring) which are part of an example discussed in Coolen and Yan (2002), who provide details on context and the original source. The 16 observations are

$$90, 142, 150, 269, 291, 468^*, 680, 837, 890^*, 1037, 1090^*, 1113^*, 1153, 1297, 1429, 1577^*$$

Figure 1 gives the lower and upper survival functions for X_{17} based on rc- $A_{(16)}$, together with the uniform Berliner-Hill survival function and the Kaplan-Meier estimate. In this figure, it has been assumed that 1700 is a known upper bound for these observations and random quantity.

Figure 1 illustrates some of the issues addressed above. For example, the Kaplan-Meier estimate puts quite a lot of mass beyond 1429. We should remark that, although we have plotted it as constant after that largest observed event time, several alternatives have been suggested on how to define it if the largest observation is a censoring time, as is the case here at 1577, e.g. some would leave it undefined beyond 1577. At observed event times, our upper and lower survival functions are equal, and indeed also coincide

with the uniform Berliner-Hill survival function. Our method, via the lower survival function, is the only one that clearly indicates where censorings take place. The uniform Berliner-Hill survival function beyond the largest event time 1429 is influenced by our choice (just for the presentation) to set 1700 as an upper bound for the random quantity X_{17} . Without such an upper bound, it is not clear how the Berliner-Hill survival function should be defined on this interval. And, this conveniently chosen upper bound allows us to illustrate that, as mentioned above, the uniform BH survival function can actually become smaller than our lower bound, which happens here in a very small interval just prior to the censoring time 1577.

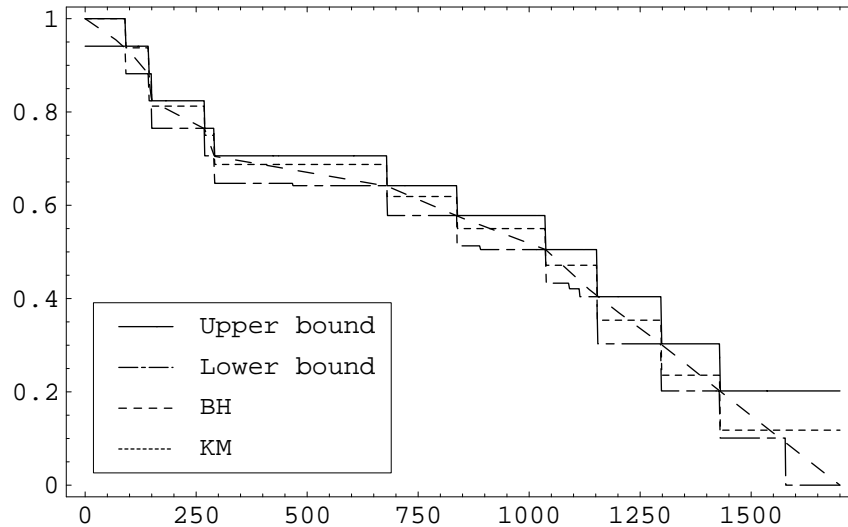


Figure 1: Survival functions; upper and lower, uniform Berliner-Hill, and Kaplan-Meier.

References

- Augustin, T. and Coolen, F.P.A. (2001). Nonparametric predictive inference and interval probability. (In submission.)
- Berliner, L.M. and Hill, B.M. (1988). Bayesian nonparametric survival analysis (with discussion). *Journal of the American Statistical Association* 83, 772-784.
- Coolen, F.P.A. and van der Laan, P. (2001). Imprecise predictive selection based on low structure assumptions. *Journal of Statistical Planning and Inference* 98, 259-277.
- Coolen, F.P.A. and Yan, K.J. (2002). Nonparametric predictive inference with right-censored data. (In submission.)
- Hill, B.M. (1968). Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association* 63, 677-691.
- Hill, B.M. (1992). Bayesian nonparametric survival analysis a comparison of the Kaplan-Meier and Berliner-Hill estimators. In J.P. Klein and P.K. Goel (Eds.), *Survival Analysis: State of the Art*, NATO ASI series, pp. 25-46. Kluwer Academic Publishers.
- Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53, 457-481.
- Weichselberger, K. (2001). *Elementare Grundbegriffe einer Allgemeineren Wahrscheinlichkeitsrechnung I. Intervalwahrscheinlichkeit as Umfassendes Konzept*. Heidelberg: Physika.
- Yan, K.J. (2002). *Nonparametric predictive inference with right-censored data*. PhD-thesis, University of Durham.