SAS/IML MODULES FOR KERNEL ESTIMATION OF THE HAZARD FUNCTION FROM CENSORED DATA

Olaf Gefeller
Abteilung Medizinische Statistik, Georg-August-Universität Göttingen

Abstract

The kernel method employing nearest neighbour distances as bandwidths for the nonparametric estimation of the hazard function from censored data is delineated in this paper. One particular problem in this context consists of the definition of nearest neighbour distances in the case of censored data. Different approaches are reviewed and a new method incorporating the full information of the censored observations into the calculation of the nearest neighbour distances is proposed. The implementation of the resulting nearest neighbour kernel estimators has taken place using the SAS/IML* matrix language. Due to its flexible programming features and its matrix orientation the programs could be implemented efficiently. In the paper the SAS/IML* modules for the estimation of the hazard function from censored data are provided in detail.

Keywords: censored data, hazard function, kernel method, nearest neighbour distance

1. Introduction

Hazard functions are nowadays in common use as an important tool in various fields of applications, e.g. in the analysis of survival data from medical studies or in issues of reliability testing in industry. Their nonparametric estimation from censored data via kernel methods has received considerable attention in the statistical literature as well as in practical applications (Padgett, 1988, Izenman, 1991). The popularity of the kernel method is due to its easy-to-understand explicit definition, its theoretical tractability, and its good practical performance. The crucial point in the practical realization of the method lies in the particular choice of the smoothing parameter, which is usually termed ‘bandwidth’. Nearest neighbour kernel estimators employing nearest neighbour distances as bandwidths have been proposed in this context (Tanner, 1983, Tanner & Wong, 1984, Liu & Van Ryzin, 1985, Cheng, 1987) to reveal the feature of ‘data-adaption’, i.e., in regions with many observations the bandwidth should be small, whereas when data are sparse the bandwidth should become large. A particular problem arises in the presence of censored data, because the natural definition of nearest neighbour distances has to be transferred to the censored setting.

2. Notation and Statistical Model

Let $T_1, \ldots, T_n$ be i.i.d. nonnegative random variables ("lifetimes") with identical distribution function $F$ and density function $f$. Let $C_1, \ldots, C_n$ be i.i.d. nonnegative random variables ("censoring times") with identical distribution function $G$ and density function $g$. Assume further that lifetimes $T_i$ and censoring times $C_i$ are independent for all $i$. Under this setting of the random censorship model, one observes the bivariate sample $(X_1, \delta_1), \ldots, (X_n, \delta_n)$, where
with \( I\{\cdot\} \) denoting the indicator function. In this case of randomly right censored data nonparametric estimates of the survivorship function \( S(x) := 1 - F(x) \) and the cumulative hazard function \( H(x) := -\log S(x) \) are well known. The Kaplan–Meier estimator \( \hat{S} \) of the survivorship function (Kaplan & Meier, 1958) and the Nelson–estimator \( \hat{H} \) of the cumulative hazard function (Nelson, 1969) belong to the standard repertoire of survival analysis (Kalbfleisch & Prentice, 1980). No standard solution exists for the nonparametric estimation problem of the hazard function \( h(x) := f(x)/S(x) \). A huge variety of methods has been discussed in the literature. We restrict our attention to the kernel method using nearest neighbour distances as bandwidths. Then, the estimator of \( h \) can be written as

\[
\hat{h}(x) = \sum_{i=1}^{n} \frac{\delta_{(i)}}{n-i+1} \cdot \frac{1}{R(k,x)} \cdot K\left( \frac{x-X_{(i)}}{R(k,x)} \right),
\]

where \( \delta_{(i)} \) denotes the censoring indicator corresponding to the \( i \)-th order statistic \( X_{(i)} \), \( K \) the kernel function satisfying some regularity conditions (Parzen, 1962), and \( R(k,x) \) the distance of \( x \) to its \( k \)-th nearest neighbour among the observations \( X_1, \ldots, X_n \).

### 3. Definitions of Nearest Neighbour Distances in the Case of Censored Data

The problem of transferring the definition of \( R(k,x) \) from the uncensored to the censored setting has been tackled in different ways:

1. Most authors ignore all censored data and define \( R(k,x) \) as the distance of \( x \) to its \( k \)-th nearest neighbour among the uncensored observations \( X_i, \delta_i = 1, i = 1, \ldots, n \). As Schäfer (1985) points out these distances have the disadvantage of being "biased" by the censoring distribution \( G \) in the sense that they adapt to the conditional density of \( T_i \) under the condition \( T_i \leq C_i \) of being uncensored, rather than to the density function \( f \) or the hazard function \( h \) to be estimated.

2. As a consequence of his criticism Schäfer proposes an alternative definition of \( R(k,x) \) in the case of censored data:

\[
R_2(k,x) := \inf\{ r > 0 \mid \hat{H}(x+r) - \hat{H}(x-r) \geq \frac{k-1}{n} \}.
\]

This definition of \( R(k,x) \) attempts to incorporate the information of the censored observations by using the Nelson estimator \( \hat{H} \). But this proposal suffers from serious conceptual drawbacks, which can be recognized immediately. If no censored data were observed in the sample, \( R_2(k,x) \) is not identical to usual definition of nearest neighbour distances in the uncensored setting. In addition, one inherent property of \( \hat{H} \) has an awkward effect on the definition of nearest neighbour distances: the heights of the steps in \( \hat{H} \) increase automatically by definition as \( x \to X_{(n)} \). Consequently, in the right tail of the lifetime distribution this effect dominates the value \((k-1)/n\) used in the definition of \( R_2(\cdot) \). Consider e.g.
the situation of an uncensored largest observation \( X_{(n)} \), i.e. \( \delta_{(n)} = 1 \). Then, for all \( t \geq (X_{(n-1)} + X_{(n)})/2 \) it follows: \( R_2(k,t) = \left| X_{(n)} - t \right| \) for all \( k \in \{1, \ldots, n\} \). In this extreme example \( R_2(\cdot) \) is completely independent of \( k \), which cannot be a desirable property of any definition of nearest neighbour distances. Even in less extreme situations the increasing heights of the steps in \( \widehat{H} \) influence the actual values of \( R_2(\cdot) \).

(3) To avoid these problems we propose a modification of the following type:

\[
R_3(k, x) := \sup \{ r > 0 \mid \widehat{S}(x - r) - \widehat{S}(x + r - 0) \leq \frac{k-1}{n} \},
\]

where \( \widehat{S}(x + r - 0) \) denotes the limit from the left of the Kaplan-Meier estimate at the point \( x + r \). By using \( \widehat{S} \) instead of \( \widehat{H} \) the beforementioned drawbacks of \( R_2(\cdot) \) are resolved. In the uncensored setting \( R_3(\cdot) \) is identical to the natural definition of the nearest neighbour distances (\( \widehat{S} \) reduces to 1 minus the usual empirical distribution function.). The heights of the steps in \( \widehat{S} \) do not show any systematic trend, they adapt to the configuration of the censored observations. Our definition \( R_3(\cdot) \) incorporates the information of the censored data in a plausible way. The open interval \( I := (x - R_3(k, x), x + R_3(k, x)) \) constitutes the largest possible interval which covers an empirical mass of at most \( (k-1)/n \) given to \( I \) by \( \widehat{S} \). Due to the configuration of censored observations the actual empirical mass of \( I \) can be smaller than \( (k-1)/n \). This is compensated by a concentration of empirical mass at the boundaries of \( I \). Consequently, the closed version of \( I \) covers an empirical mass of at least \( k/n \). This procedure resembles the situation of tied observations in the uncensored setting. A recently conducted simulation study (Gefeller & Dette, 1991) has demonstrated that employing \( R_3(\cdot) \) as the definition of nearest neighbour distances in the calculation of kernel estimators of the hazard function leads to a substantial reduction in the mean integrated squared error of the corresponding estimator.

4. Computational Realization in SAS/IML

The computational realization of these nearest neighbour kernel estimators of the hazard function took place within the SAS/IML* software, which provides a flexible and easy-to-use matrix language enabling us to employ fast algorithms in computing all estimators. The specific SAS/IML* modules to calculate the three nearest neighbour distances and the module to compute the resulting nearest neighbour kernel estimator (using the 'biweight' kernel function) are presented below.

Explanation of main variables:

- **n**: number of observations
- **k**: value specifying the order of the nearest neighbour distance (\( k \)-th nearest neighbour distance)
- **data**: a \( 2 \times n \) matrix; first row contains all ordered lifetimes (censored and uncensored), second row contains the corresponding censoring indicator
- **nndist**: vector containing the \( k \)-th nearest neighbour distance at each point of the grid
- **begin, end, step**: variables declaring position and spacing of the grid
uncens: vector containing all uncensored lifetimes in ascending order

uncenobs: number of uncensored observations

nelson: a $2 \times (n+1)$ matrix containing all values of the Nelson estimator of the cumulative hazard function; first row contains estimated values, second row contains arguments of the estimator

kaplan: a $2 \times (n+1)$ matrix containing all values of the Kaplan-Meier estimator of the survivorship function; first row contains estimated values, second row contains arguments of the estimator

jumpind: vector containing the column position numbers of all uncensored observations in the matrix data

hazard vector containing the estimated values of the hazard function at each point of the grid

Listing of the SAS/IML* modules:

(1) start nddef1(data,k,nndist,begin,end,step);
    uncensor=data[1,loc(data[2,])]; j=0;
    do x = begin to end by step;
        j=j+1;
        distance = abs(uncensor - x);
        nndist[j] = distance[loc(rank(distance)=k)];
    end;
    finish;

(2) start nddef2(n,data,k,nndist,begin,end,step,uncenobs);
    maxind=min(uncenobs+1,k+3);
    intmass=j(2,maxind,0);
    nelson=j(2,n+1,0);
    nelson[1,1]=0;
    nelson[2,1]=0;
    do i=1 to n;
        nelson[1,i+1]=nelson[1,i]+data[2,i]/(n-i+1);
        nelson[2,i+1]=data[1,i];
    end;
    uncensor=data[1,loc(data[2,])];
    jumpind = loc(data[2,]);
    j=0;
do x = begin to end by step;
  j=j+1;
  distance = abs(uncensor - x);
  left=data[1,jumpind[loc(rank(distance)=1)]];
  right=left;
  intmass[1,1]=distance[loc(rank(distance)=1)];
  intmass[2,1]=0;
  do m=2 to maxind;
    if data[1,jumpind[loc(rank(distance)=m)] < left
      then do;
        left=data[1,jumpind[loc(rank(distance)=m)]];
        intmass[1,m]=distance[loc(rank(distance)=m)];
        intmass[2,m]=nelson[1,jumpind[loc(uncensor=right)]] - nelson[1,jumpind[loc(uncensor=left)]];  
      end;
    else do;
      right=data[1,jumpind[loc(rank(distance)=m)]];
      intmass[1,m]=distance[loc(rank(distance)=m)];
      if left > 0
        then intmass[2,m]=nelson[1,jumpind[loc(uncensor=right)-1]] - nelson[1,jumpind[loc(uncensor=left)-1]]
          else intmass[2,m]=nelson[1,jumpind[loc(uncensor=right)-1]];
      end;
    end;
  end;
  kindex=maxind;
  do i=1 to maxind while (kindex=maxind);
    if intmass[2,i] < (k-1)/n
      then kindex=i-1;
    end;
  ndist[j]=intmass[1,kindex];
  end;
finish;

(3) start nngefn(data,k,ndist,begin,end,step,uncenobs);
  maxind=min(uncenobs+1,k+3);
  intmass=j(2,maxind,0);
5. Discussion

In this paper we have considered computational aspects of a special subclass of kernel estimators, those in which the bandwidths are determined by nearest neighbour distances. Due to their computational simplicity and promising asymptotic properties, kernel-type estimators enjoy great popularity in all applications of the nonparametric estimation of the hazard function nowadays. The choice of the bandwidth-parameter is crucial for the behaviour of this class of estimators. For the particular subclass of nearest neighbour kernel estimators, to our knowledge, the problem of defining nearest neighbour distances in the case of censored data has not been adequately taken account of in the literature. Intuitively, it should be clear that ignoring the information of the censored observations in calculating nearest neighbour distances, as it is the existing advice in almost all papers on this topic, cannot be the optimal solution to this problem. Therefore, we have developed an approach to comprise the information of all data in the calculation of nearest neighbour distances inspired by
Schiäfer's idea (Schiäfer, 1985), but not suffering from the serious drawbacks of his realization. In the situation of no censoring our method yields by definition identical results as the usual calculation of nearest neighbour distances. As shown in a simulation study comparing the effects of the different approaches on the properties of the resulting hazard function estimators (Gefeller & Dette, 1991), the new definition of nearest neighbour distances in the case of censored data leads to an estimator with a mean integrated squared error smaller than the corresponding mean integrated squared error of the other estimators in all situations covered in the study. On the basis of these results it seems justified to recommend the usage of the new method in practical applications dealing with the estimation of hazard functions from censored data whenever nearest neighbour distances are involved in the calculations. Computationally, the method can be easily implemented using the SAS/IML* software. The particular modules have been presented in this paper.

References


SAS/IML is a registered trademark of SAS Institute Inc., Cary, NC, USA.