J. KSIAM **Vol.3**, No.2, 99-106, 1999

# SPLINE HAZARD RATE ESTIMATION USING CENSORED DATA

Myung Hwan Na

**Abstract**

In this paper, the spline hazard rate model to the randomly censored data is introduced. The unknown hazard rate function is expressed as a linear combination of B-splines which is constrained to be linear(or constant) in tails. We determine the coefficients of the linear combination by maximizing the likelihood function. The number of knots are determined by Bayesian Information Criterion. Examples using simulated data are used to illustrate the performance of this method under presenting the random censoring.

## 1 Introduction

Reliability engineers, biostatisticians, and actuaries are all interested in lifetimes. In particular, they are interested in five lifetime distribution representations: the hazard rate function $h(t)$, the cumulative hazard rate function $H(t)$, the reliability function $R(t)$, the probability density function $f(t)$, and the mean residual life function $m(t)$. Perhaps, the hazard rate function is the most popular of the five representations for life–time modelling. The hazard rate function is defined as

$$h(t) = \frac{f(t)}{R(t)}, \quad t \geq 0.$$

Thus, the hazard rate function is the ratio of the probability density function to the reliability function. Throughout this paper we assume that the hazard rate function satisfy two conditions:

$$\int_0^\infty h(t)dt = \infty,\ h(t) > 0 \text{ for all } t > 0. \tag{1}$$

A smooth estimation of the hazard rate function is a very important topic in both theoretical and applied statistics; Anderson and Senthilselvan(1980) use quadratic spline with discontinuity in the slope at the times of death. O'sullivan(1988) use smoothing splines for the log-hazard function. Senthilselvan(1987) use hyperbolic spline function which is continuous with its first derivative discontinuous only at a finite number of points. Kooperberg, et al.(1995) use cubic splines and two additional log terms for

Key Words: Hazard Rate, Spline, Censoring, Simulation

log-hazard function. The discussion section of Abrahamowicz, et al.(1992) contains a good review of many of the papers on the use of splines to estimate density in the presence of censored data.

In this paper we introduce the spline hazard rate model to the random censored data. The unknown hazard rate function is expressed as a fuction from a space of cubic splines constrained to be linear(or constant) in tails. The coefficients of the linear combination are determined by maximizing the likelihood function. The number of knots are determined by Bayesian Information Criterion(BIC). Examples using simulated data are used to illustrate the performance of this method under presenting the random censoring.

Section 2 is devoted to an introduction to spline model for the hazard rate function. A maximum likelihood estimation procedure is discussed in section 3. Section 4 contains knot deletion procedure. Section 5 contains examples using simulated data.

## 2 SPLINE HAZARD RATE MODEL

Let K denote a nonnegative integer. When $K \geq 1$, let $\xi_1, \cdots, \xi_K$ be a (simple) knot sequence in $[0, \infty)$ where $0 < \xi_1 < \cdots < \xi_K < \infty$. Let $S_0$ denote the collection of twice continuously differentiable functions $s$ on $[0, \infty)$ such that the restriction of $s$ to each of the intervals $[0, \xi_1], [\xi_1, \xi_2], \cdots, [\xi_K, \infty)$ is a cubic polynomial, i.e., $s$ is a polynomial of order 4 (or less) on each of the intervals. Then $S_0$ is the (K+4)-dimensional vector space of cubic splines corresponding to the knot positions $\xi_1, \cdots, \xi_K$. Set $S$ denote the subspace of $S_0$ consisting of the natural cubic splines with knots at $\xi_1, \cdots, \xi_K$, i.e., the functions in $S$ that are linear (or constant) on $[0, \xi_1]$ and $[\xi_K, \infty)$. This linear vector space is $K$-dimensional and has a basis $B_1, \cdots, B_K$ of S. When $K = 0$, there are no basis functions depending on $t$. For exhaustive treatment of splines, the reader should consult Greville(1969), de Boor(1978), and Schumaker(1981). For statistical applications we refer to Smith(1979), and Wegman and Wright(1983).

Let $\Theta$ denote the collection of all column-vector $\theta = (\theta_1, \cdots, \theta_K)^t \in R^K$ such that $\sum_{j=1}^K \theta_j B_j(t) > 0$ for all $t > 0$. Given $\theta \in \Theta$, consider the model

$$h(t|\theta) = \sum_{j=1}^{K} \theta_j B_j(t), \quad t > 0$$

for the hazard rate function. For this spline model, the corresponding cumulative hazard rate function, reliability function, and probability density function are respectively given by

$$\begin{aligned} H(t|\theta) &= \sum_{j=1}^{K} \theta_j C_j(t), \\ R(t|\theta) &= \exp\left(-\sum_{j=1}^{K} \theta_j C_j(t)\right), \end{aligned}$$

$$f(t|\theta) = \left(\sum_{j=1}^{K}\theta_j B_j(t)\right)\exp\left(-\sum_{j=1}^{K}\theta_j C_j(t)\right)$$

where $C_j(t) = \int_0^t B_j(u)du$.

In particular, when $K = 0$, this spline model includes exactly exponential distribution. When $K = 1$, this is exact hazard rate function of the Rayleigh distribution. The MLE $\hat{\theta}$ is obtained by maximizing the likelihood function. We refer to $\hat{h}(\cdot) = h(\cdot|\hat{\theta})$ as the spline hazard rate estimate.

## 3 MAXIMUM LIKELIHOOD ESTIMATION

Let $T_1, T_2, \cdots, T_n$ be independent identically distributed(i.i.d.) with a life distribution function(d.f.) $F$ and let $C_1, C_2, \cdots, C_n$ be i.i.d. with d.f. $G$. $C_i$ is the censoring time associated with $T_i$. In random censoring case we can only observe $(Y_1, \delta_1), \cdots, (Y_n, \delta_n)$ where $Y_i = \min(T_i, C_i)$, $\delta_i = I(T_i \le C_i)$, $1 \le i \le n$. It is assumed that $T_i$ and $C_i$ are independent. The random variable $Y_i$ is said to be uncensored or censored according as $\delta_i = 1$ or $\delta_i = 0$. Note that the partial likelihood corresponding to the data $(y_i, \delta_i)$ equals $[f(y_i)]^{\delta_i}[1 - F(y_i)]^{1-\delta_i}$ (see Miller, 1981), so the log-likelihood for the data $(y_i, \delta_i)$ equals

$$\psi(y_i, \delta_i) = \delta_i \log h(y_i) - H(y_i).$$

Thus the log-likelihood function corresponding to the spline model is determined by

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n}\psi(y_i, \delta_i) \\ &= \sum_{i=1}^{n}\delta_i \log(h(y_i)) - \sum_{i=1}^{n}H(y_i). \end{aligned}$$

Moreover,

$$\frac{\partial}{\partial\theta_j}l(\theta) = \sum_{i=1}^{n}\frac{\delta_i B_j(y_i)}{h(y_i)} - \sum_{i=1}^{n}C_j(y_i), \qquad 1 \le j \le K$$

and

$$\frac{\partial^2}{\partial\theta_j\partial\theta_k}l(\theta) = -\sum_{i=1}^{n}\frac{\delta_i B_j(y_i)B_k(y_i)}{(h(y_i))^2}, \qquad 1 \le j, k \le K.$$

It follows from the last result that $l(\theta)$ is a concave function. Thus the MLE is unique if it exists.

Let $S(\theta)$ denote the score function of $l(\theta)$, that is, $K$-dimensional column vector with entries $\partial l(\theta)/\partial\theta_j$, and let $H(\theta)$ denote the Hessian of $l(\theta)$, the $K \times K$ matrix with entries $\partial^2 l(\theta)/\partial\theta_j\partial\theta_k$. The maximum likelihood equation for $\hat{\theta}$ is $S(\hat{\theta}) = 0$. We use

Newton-Raphson method with step-halving for computing $\hat{\theta}$, to start with an initial guess $\hat{\theta}^{(0)}$ and iteratively determine $\hat{\theta}^{(m+1)}$ by the formula

$$\hat{\theta}^{(m+1)} = \hat{\theta}^{(m)} + \frac{1}{2^M} I^{-1}(\hat{\theta}^{(m)}) S(\hat{\theta}^{(m)})$$

where $I(\theta) = -H(\theta)$ and $M$ is the smallest nonnegative integer such that

$$l\{(\hat{\theta}^{(m)} + \frac{1}{2^M} I^{-1}(\hat{\theta}^{(m)}) S(\hat{\theta}^{(m)})\} \geq l\{(\hat{\theta}^{(m)} + \frac{1}{2^{M+1}} I^{-1}(\hat{\theta}^{(m)}) S(\hat{\theta}^{(m)})\}.$$

We stop the iterations when $l(\hat{\theta}^{(m+1)}) - l(\hat{\theta}^{(m)}) \leq 10^{-6}$.

# 4 KNOT DELETION PROCEDURE

In this section we determine the rules for selecting the number and location of knots. In order to determine the number and location, we can directly apply the stepwise knot deletion method of Smith(1982). First place enough initial knots appropriately and then delete unnecessary knots. According to Stone(1991), for twice continuously differentiable $h(t)$, an optimal rate of convergence $n^{-2/5}$ can be achieved if the number of knots is increased proportionally to $n^{1/5}$. So we use the integer $K$ closest to $4n^{1/5}$ as a number of initial knots. We will describe an initial knot placement rule: place two knots at the first and the last order statistics and the remaining knots as closely as possible to the equi-spaced percentiles. For example, if the number of initial knots is five, they are placed at the 0, 25, 50, 75, 100 percentiles.

First, consider the problem that the estimate of hazard rate function take negative values. The estimates of the hazard rate may take negative values for large value of $K$ or on intervals $[0, \xi_1]$ and $[\xi_K, \infty)$. So we used following method to satisfy conditions (1) and (2).

$(i)$ If $\hat{\theta}_1(\xi_1 + \xi_2 + \xi_3) + \hat{\theta}_2$ less than 0, we set $\theta_1 = 0$, i.e. the function is constant on $[0, \xi_1]$.

$(ii)$ If $\hat{\theta}_K$ less than 0, we set $\theta_K = 0$, i.e. the function is constant on $[\xi_K, \infty)$.

$(iii)$ If the minimum value of $\hat{h}(x_i)$, $i = 1, \cdots, n$, is negative, we delete the closest knot to $x_{j^*}$, where $x_{j^*}$ is argument of minimum value of $\hat{h}(x_i)$.

Now consider the problem of deleting unnecessary knots. Following Smith(1982), the absence of a knot $\xi$ of a spline $\sum_{j=1}^{K} \theta_j B_j(t)$ means that

$$\sum_{j=1}^{K} \theta_j \delta_j(\xi) = 0$$

where $\delta_j(\xi) = B_j^{(3)}(\xi-) - B_j^{(3)}(\xi+)$, $B_j^{(3)}(\xi-)$ and $B_j^{(3)}(\xi+)$ are, respectively, the left- and right-hand limit of $\partial^3 B_j(t)/\partial t^3$ at $\xi$.

At any step we compute

$$\eta_k = \frac{|\hat{\psi}_k|}{\mathrm{SE}(\hat{\psi}_k)} \quad k = 1, \cdots, K$$

where $\hat{\psi}_k = \sum_{j=1}^{K} \hat{\theta}_j \delta_j(\xi_k)$ and $\mathrm{SE}(\hat{\psi}_k) = \{\delta^t(\xi_k)(I(\hat{\theta}))^{-1}\delta(\xi_k)\}^{1/2}$. And we delete the knot having the smallest value of $\eta_k$. In this manner, we arrive at a sequence of models indexed by $J$, which ranges from 0 to $K$. Let $I_L = 1$ when the estimated function is constant on $[0, \xi_1]$, $I_L = 0$ otherwise. Let $I_R = 1$ when the estimated function is constant on $[t_K, \infty)$, $I_R = 0$ otherwise. Let $\hat{l}_J$ denote the log-likelihood function for the Jth model evaluated at the MLE for that model. Let $\mathrm{BIC} = -2\hat{l}_J + \log(n)(K - J - I_L - I_R)$ be the Bayesian Information Criterion(Schwarz, 1978) for the $J$th model. We choose the model corresponding to that value $\hat{J}$ of $J$ that minimizes BIC. This model has $K - \hat{J}$ knots and $K - \hat{J} - I_L - I_R$ free parameters.

# 5 EXAMPLES

The spline hazard rate estimation procedure described in Section 2 is applied to Weibull, Gamma and Dhillon distributions. The density functions are respectively given by:

$$\begin{aligned}
f(t) &= \frac{\alpha}{\beta}(\frac{t}{\beta})^{\alpha-1}\exp(-(\frac{t}{\beta})^{\alpha}), \\
f(t) &= \frac{1}{\Gamma(\alpha)}\frac{t^{\alpha-1}}{\beta^{\alpha}}\exp(-\frac{t}{\beta}), \\
f(t) &= \alpha\sigma(\sigma t)^{\alpha-1}\exp((\sigma t)^{\alpha})\exp(1-\exp(\sigma t)^{\alpha}).
\end{aligned}$$

The simulation is performed on the subroutine IMSL of the package FORTRAN.

In Figure 1, we show the true hazard rate function(solid) corresponding to Weibull distribution with parameter $\alpha = 0.8$ and $\beta = 1$. The dotted line corresponds to the estimated hazard rate function based on a sample of size 200. Figure 2 is similar to Figure 1, but the underlying distribution for Figure 2 is Gamma distribution. The data for Figure 2 is from Gamma distribution with parameter $\alpha = 2$ and $\beta = 1$ based on a sample of size 200. In the figure we show the true hazard rate function corresponding to this Gamma distribution together with the estimate for the hazard rate function based on the spline model. In Figure 3, we show the result of similar calculation based on a sample of size 200 from Dhillon distribution with parameter $\alpha = 1$ and $\beta = 1$, i.e., extreme value distribution. In the figure we show the true hazard rate function corresponding to this Dhillon distribution together with the estimate for the hazard rate function based on the spline model.

From these examples, we have found that the spline hazard rate estimate yields a reasonable estimate for the hazard rate function.

**Figure 1.** Spline hazard rate estimate for Weibull distribution with $\alpha = 0.8$ and $\beta = 1$ based on sample of size 200.

**Figure 2.** Spline hazard rate estimate for Gamma distribution with $\alpha = 2$ and $\beta = 1$ based on sample of size 200.

**Figure 3.** Spline hazard rate estimate for Dhilon distribution with $\alpha = 1$ and $\sigma = 1$ based on sample of size 200.

## REFERENCES

1. Abrahamowicz, M., Ciampi, A. and Ramsay, J. O. (1992): "Nonparametric Density Estimation for Censored Survival Data : Regression-Spline Approach", *The Canadian Journal of Statistics*, Vol. 20, 171-185.

2. Anderson, J. A. and Senthilselvan, A. (1980): "Smooth Estimates for the Hazard Function", *Journal of the Royal Statistical Society*, Ser. B, Vol. 42, 322-327.

3. de Boor, C. (1978), *A Practical Guide to Splines*, Springer-Verlag, New York.

4. Greville, T. N. E. (1969), *Theory and Application of Spline Function*, Academic Press, New York.

5. Kooperberg, C., Stone C. J. and Truong, Y. K. (1995): "Hazard Regression", *Journal of American Statistic Association*, Vol. 90, 78-94.

6. Miller, R. (1981) *Survival Analysis*, John Wiley & Sons, New York.

7. O'sullivan F. (1988): "Fast Computation of Fully Automated Log-Density and Log-Hazard Estimates", *SIAM Journal of Scientific and Statistical Computing*, Vol. 9, 363-379.

8. Schumaker, L. L. (1981): *Spline functions; Basic Theory*, Wiley, New York.

9. Schwarz, G. (1978): "Estimating the dimension of model", *Annals of Statistics*, Vol. 6, 461-464.

10. Senthilelvan, A. (1987): "Penalized Likelihood Estimation of Hazard and Intensity Functions", *Journal of the Royal Statistical Society*, Ser. B, Vol. 49, 170-174.

11. Smith, P. L. (1979): "Splines as a Useful and Convenient Statistical Tools", *The American Statistician*, Vol. 33, pp. 57-62.

12. Smith, P. L. (1982): "Curve fitting and modeling with splines using statistical variable selection methods" NASA, Langley Research Center, Hampla, VA, NASA Report 166034.

13. Stone C. J. (1991): Generalized Multivariate Regression Splines, Technical Report No. 318, Dept. statist. Univ. California, Berkeley.

14. Wegman, E. J. and Wright, I. W. (1983), "Splines in Statistics", *Journal of American Statistic Association*, Vol. 78, 351-366.

Department of Statistics, Seoul National
University, Seoul 151-742, Korea
e-mail: nmh@stats.snu.ac.kr