

Applications of Smoothed Monotone Regression Splines and Smoothed Bootstrapping in Survival Analysis

Donald E. Ramirez^{1,2} and Philip W. Smith³

¹ University of Virginia, Department of Mathematics, Kerchof Hall 209, Charlottesville, VA 22903 USA

³ WindWard Technologies, Inc., 12039 Mulholland, Meadows Place, TX 77477 USA

Abstract. An estimate of a survival curve $S(x)$ for censored data is given by the non-parametric Kaplan-Meier method, which provides an estimate of the empirical survival curve and estimators of the standard errors. We use the software package Confit, developed by WTI, which is designed to produce a smoothed approximating spline, subject to imposed constraints on the function or its derivatives over an interval. The algorithm solves a constrained least-squares problem parameterized by an appropriate spline subspace (using a B-spline representation). The constraints impose some additional constraints on these coefficients that are converted into a quadratic programming problem. We will discuss the algorithm used to solve the quadratic programming problem, and give applications to illustrate our method on several data sets.

1 Introduction

In modelling data, the statistician is searching for a functional form $\phi(x)$ that the data satisfies. In certain cases, we wish to impose additional constraints on either the functional form $\phi(x)$ or its derivative $\phi'(x)$. For example, to model the cumulative distribution function $F(x)$ or the survival $S(x) = 1 - F(x)$, monotonicity is essential. For a review of some of the important statistical applications of splines see Smith (1979).

A set of basic splines which are useful in computational statistics is the set of B-splines. For a nondecreasing knot sequence $\{t_0, t_1, \dots, t_N\}$, $B_i^0(x)$ is the indicator function for $[t_i, t_{i+1})$ and the higher order splines ($k \geq 1$) are defined recursively by

$$B_i^k(x) = \left(\frac{x - t_i}{t_{i+k} - t_i} \right) B_i^{k-1}(x) + \left(\frac{t_{i+k+1} - x}{t_{i+k+1} - t_{i+1}} \right) B_{i+1}^{k-1}(x), \quad t_i \leq x \leq t_{i+k+1}. \quad (1)$$

The B-splines are nonnegative and are normalized by $\sum_{i=0}^N B_i^k(x) = 1$. (See for example, DeBoor (1978, p. 110).) We will denote by $B^{N-1}(x, \{t_0, \dots, t_N\})$ the $N - 1$ degree B-spline for $\{t_0, \dots, t_N\}$.

²The first author was partially supported by the Center for Advanced Studies at the University of Virginia

A closely related set of splines is given by $M_i^k(x) = kB_i^k(x)/(t_{i+m} - t_i)$ (Curry and Schoenberg, 1966). Since the M -splines are nonnegative functions, their integrals provide a set of monotone functions. Ramsay (1988) used for his monotone regression splines sums of these integrated splines with positive coefficients. Ramsay described a nonlinear algorithm using the gradient projection constrained optimization algorithm for determining the optimal coefficients. Gaylord and Ramirez (1991), building on the work of Ramsay, introduced an adaptive linear algorithm using weighted regression to force regression splines to be monotone.

Smoothing splines, pioneered by Wahba (1990), are found by minimizing a combination of the least-squares errors $\sum_{i=0}^N (y_i - \phi(x_i))^2$ and the L^2 norm of the generalized curvature $\int_{t_0}^{t_N} (\phi^{(m)}(x))^2 dx$. Kelly and Rice (1990) constrained the B-spline coefficients to be monotone to enforce monotonicity of the smoothing spline. Turlach (1997) has proposed calculating smoothing splines with constraints. His approach leads to a quadratic programming problem, with the infinite number of constraints replaced by a suitable finite number, which can be solved using the algorithm of Goldfarb and Idnani (1983).

In this paper, we will briefly describe the software package Confit which finds an approximating spline where the user is able to prescribe constraints on the function and/or the derivatives. We then present some applications of monotone splines that are efficiently solved with Confit.

2 Confit

Confit is a package designed to produce a smoothed approximating spline subject to a finite number of constraints on the function or its derivatives. It fits a spline function $y = S(x)$ to the data $\{(x_i, y_i) : i = 0, \dots, N\}$ with weights $\{w_i : i = 0, \dots, N\}$, by minimizing the error sum of squares, subject to specified shape constraints. Specifically, it solves the problem: minimize $\sum_{i=0}^N w_i (S(x_i) - y_i)^2$ subject to $\alpha_j \leq S^{(p_j)}(x) \leq \beta_j$ on $a_j \leq x \leq b_j$, for $j = 1, \dots, m$. The algorithm involves a least-squares problem which is solved for the coefficients of an appropriate B-spline. When constraints are imposed on the approximating spline, the problem is converted into a quadratic programming problem that Confit solves with amazing speed. For any given constraint on an interval, Confit checks for compliance on a grid of 500 equally spaced points. The actual computation involves building a sequence of optimization problems by adding the most violated point one at a time (for each constraint) until compliance is observed. To create a finer mesh, the interval can be subdivided. Confit uses an Excel compatible spreadsheet for its data entry and output. A 90 day free trial version of Confit is available at the web site <http://web.wti.net/~wti>.

3 Survival curves for censored exponential data

The data used here is simulated data from an exponential distribution with mean $\mu = 24$ and sample size $N = 25$. The (five) values > 70 are considered as being right-censored values. An estimate y of the survival curve $S(x)$ is given by the Kaplan-Meier method. This method also provides an estimate of the standard error for these estimates. We choose the inverse of the square of the standard error for the weights w . This follows the procedure for a logistic response function (see Neter *et al.*, p. 363).

x	0	1	2	5	6	9	10	12	14	17
y	.96	.92	.88	.84	.72	.68	.64	.60	.56	.52
w	651	339	237	186	124	115	109	104	101	100
x	18	20	21	31	35	38	41	49	63	70
y	.48	.44	.40	.36	.32	.28	.24	.20	.16	.12
w	100	101	104	109	115	124	137	156	186	237

Table 3.1. Censored Exponent Data with $m = 24$ and $N = 25$.

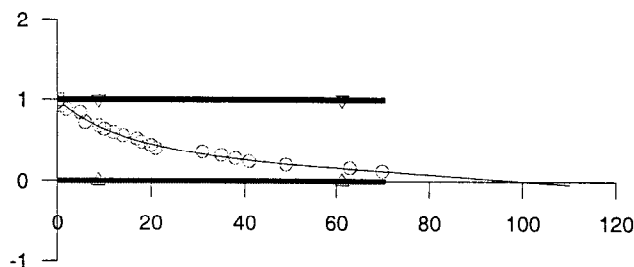


Fig. 3.1. B-Spline approximation for censored exponential data from Confit.

To model this data, we introduce the four constraints: (1) $S(0) = 1$, (2) $0 \leq S(x) \leq 1$ on $[0, 70]$, (3) $S'(x) \leq 0$ on $[0, 70]$, and (4) $S''(x) = 0$ on $[70, 120]$. The first and second constraints are the cdf constraints; the third constraint is the monotonicity constraint. The fourth constraint extends the spline linearly downward to the x -axis.

Using the defaults for Confit, the B-spline approximation of the data produces a fourth degree spline approximation with eight B-splines with knot sequence $\{0, 0, 0, 0, 0, 27.5, 55, 82.5, 110, 110, 110, 110, 110\}$ and coefficients $\{1, .654, .410, .218, .137, .041, -.011, -.037\}$. Thus $S(x) = B^4(x, \{0, 0, 0, 0, 0, 27.5\}) + .654B^4(x, \{0, 0, 0, 0, 27.5, 55\}) + \dots - .037B^4(x, \{82.5, 110, 110, 110, 110, 110\})$.

We can estimate quantiles for the data from the smoothed spline. The estimates for the 50%, 95%, and 99% quantiles are 17.1 (16.6), 88.0 (71.9), and

97.8 (110.5), respectively, where the actual values are shown in parentheses following. The plot of the monotone spline approximation is shown in Figure 3.1.

4 Survival curves for Weibull data

The data here are $N = 25$ values from a Weibull distribution with $\alpha = 1.5$ and $\beta = 1$. The cumulative hazard function $H(x) = -\log(1 - F(x)) = (x/\beta)^\alpha$ provides another application for smoothed monotone regression splines. The hazard function $h(x) = H'(x) = (\alpha/\beta)(x/\beta)^{\alpha-1}$ is assumed to be nonnegative and monotone increasing ($\alpha > 1$). The ordered data $\{x_{(1)}, \dots, x_{(N)}\}$ is paired with the empirical distribution by

$$F(x_{(i)}) = (i - 1/3)/(N + 1/3). \quad (2)$$

This follows the convention of Hoaglin *et al.* (1983, p. 44) for quantiles. Confit is used to fit a monotone spline to the data $\{(x_{(i)}, y_i) : i = 1, \dots, N\}$ with $y_i = -\log(1 - F(x_{(i)})) = H(x_{(i)})$. The constraints required are (1) $H(x) \geq 0$, (2) $h(x) = H'(x) \geq 0$, and (3) $h'(x) \geq 0$ for monotonicity of the hazard function. Using the defaults for Confit, the B-spline approximation of the data produces a fourth degree spline approximation with nine B-splines with 14 knots.

The survival curve estimate is recovered from $S(x) = \exp(-H(x)) = \exp(-(x/\beta)^\alpha)$. To extend the spline approximation of $S(x)$ downward to the x -axis with an exponential curve; the corresponding values for α and β used are

$$\alpha = -x_{(N)}h(x_{(N)})/\log(S(x_{(N)})) \quad (3)$$

$$\beta = (\alpha x_{(N)}^{\alpha-1} h(x_{(N)}))^{1/\alpha} \quad (4)$$

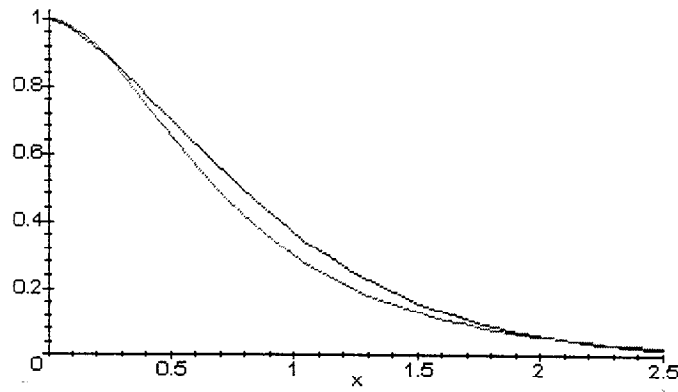


Fig. 4.1. B-Spline approximation for Weibull data with underlying curve.

We can estimate quantiles for the data from $S(x)$. The estimates for the 50%, 95%, and 99% quantiles are 0.679 (0.783), 2.09 (2.08), and 3.07 (2.77), respectively, where as before the actual values are shown in parentheses following. The plot of the monotone approximation is shown in Figure 4.1 with $S(x)$ the lower curve from $[0.24, 2.04]$.

5 Density estimation

The mouse autopsy data set ($N = 40$) from Hoel (1972, Table 1: Other Causes) is a most challenging data set for density estimation. We wish to estimate the empirical distribution function $F(x)$ using a smoothed monotone spline. The pairs of values for the spline approximation are $\{(x_{(i)}, F(x_{(i)})) : i = 1, \dots, 40\}$ using Equation 2. The x -values range from 40 to 763. Density estimation on this data set often yields negative estimates for the pdf in the interval $[70, 100]$. The constraints we used for our B-spline approximation were (1) $0 \leq F(x) \leq 1$ on $[40, 763]$, (2) $F'(x) = f(x) \geq 0$ on $[70, 100]$, (3) $F''(x) = f'(x) = 0$ on $[763, 770]$, and (4) $F''(x) = 0$ on $[25, 40]$. Constraints (1) and (2) are the cdf constraints, while constraints (3) and (4) extend $F(x)$ linearly. Using the defaults for Confit, the B-spline approximation of the cdf $F(x)$ for the mouse data produces a fourth degree spline approximation with 21 B-splines and 26 knots. The plot of the density estimator $f(x) = F'(x)$ is shown in Figure 5.1.

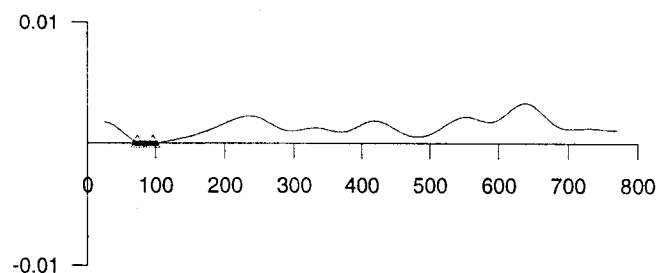


Fig. 5.1. B-spline approximation for the density for the mouse data from Confit.

6 Smoothed bootstrapping

The seminal work of Efron (1982) popularized the bootstrap that has now become a common tool for use in exploring the assessment of errors in statistical estimation problems. Consider the problem of estimating the maximum value from a set of five values from the censored exponential data above.

Since the data is censored, the range is not defined when a censored value is used in the resampled data. To avoid this difficulty, one can use the monotone regression spline approximation $S(x)$ for the empirical distribution function $F(x)$, which provides a semiparametric model of the underlying distribution. For a uniformly distributed sequence $\{u_1, \dots, u_N\}$ from $[0, 1]$ the resampled data consist of $\{S^{-1}(u_1), \dots, S^{-1}(u_N)\}$. This procedure has been used in Gaylord and Ramirez (1991). Resampling with 100 sets of five values, we found that 90% of the range values fell in the interval $[18, 92]$.

References

- Curry, H. B. and Schoenberg, I. J. (1966). On Polya frequency functions. IV. The fundamental spline functions and their limits. *J. Analyse Math.*, **17**, 71-107.
- DeBoor, C. (1978). *A Practical Guide to Splines*. New York: Springer-Verlag.
- Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*, Philadelphia, SIAM.
- Gaylord, C. and Ramirez, D. (1991). Monotone regression splines for smoothed bootstrapping, *Computational Statistics Quarterly*, **6**, 85-97.
- Goldfarb, D. and Idnani, A. (1983). A numerically stable dual method for solving strictly convex quadratic programs, *Mathematical Programming*, **27**, 1-33.
- Hoaglin, D. Mosteller, F. and Tukey, J. (1983). *Understanding Robust and Exploratory Data Analysis*. New York, John Wiley.
- Hoel, D. (1972). A representation of mortality data by competing risks, *Biometrics*, **28**, 475-488.
- Kelly, C. and Rice, J. (1990). Monotone smoothing with applications to dose-response curves and the assessment for synergism, *Biometrics*, **46**, 1071-1085.
- Kleinbaum, D. (1996). *Survival Analysis*. New York, Springer-Verlag.
- Neter, J., Wasserman, W. and Kutner, M. (1983). *Applied Linear Regression Models*. Homewood, Illinois, Richard D. Irwin, Inc.
- Ramsay, J. (1988). Monotone regression splines in action, *Statistical Science*, **4**, 425-461.
- Smith, P. L. (1979). Splines as a useful and convenient statistical tool, *Amer. Statistician*, **33**, 57-62.
- Turlach, B. (1997). Constrained smoothing splines revisited. *Statistics Research Report No. SRR 008-97*, Centre for Mathematics and its Applications, Australian National University.
- Wahba, G. (1990). *Spline Models for Observational Data*, Philadelphia, SIAM.