PENALIZED LIKELIHOOD HAZARD ESTIMATION: ALGORITHM AND EXAMPLES

by

Chong Gu
Purdue University

Technical Report #92-25

Department of Statistics
Purdue University

June 1992

◇

Penalized Likelihood Hazard Estimation: Algorithm and Examples

CHONG GU*

Purdue University

Abstract

Based on an earlier theoretical analysis, the practical implementation of penalized likelihood hazard estimation using censored life time data is studied. An algorithm with an automatic smoothing parameter is implemented in a portable code and is examined via simulations. The algorithm is an adaptation of the performance-oriented iteration developed earlier for density estimation and the performance is measured by a proxy of symmetrized Kullback-Leibler. A key ingredient of the algorithm is a cross-validation scheme based on the martingale structure of censored data. Various practical aspects of the methodology are discussed via examples.

Key words and phrases. Censored data, cross-validation, hazard, smoothing parameter, symmetrized Kullback-Leibler.

1 Introduction

Censored life time data are common in life testing, medical follow up and other studies. Let T_i be the life time of an item and C_i be the censoring time beyond which the item is dropped from the study. One observes (X_i, δ_i), $i = 1, \cdots, n$, where $X_i = \min(T_i, C_i)$ and $\delta_i = I_{[T_i \le C_i]}$. Assume that T_i follow a common survival function $S(t) = \text{Prob}(T > t)$. Of interest is the estimation of the hazard function $\lambda(t) = -d \log S(t)/dt$. Assume $\lambda(t) > 0$ on $\{t : \tilde{S}(t) = \text{Prob}(X \ge t) > 0\}$ and let $\eta(t) = \log \lambda(t)$. We shall use $e^{\eta(t)}$ to indicate the hazard in the remaining of the article and reserve λ for the smoothing parameter, to be discussed shortly.

*Research supported by NSF under Grant DMS-9101730.

Data and model are two sources of information in a statistical analysis. Data carry noise but are "unbiased", while models, or constraints, help to reduce noise but are responsible for "biases". In a parametric analysis with strict constraints, η is assumed to be a member of a parametric family, say $P_\theta = \{\eta(t,\theta) : \theta \in \Theta\}$, where $\eta(t,\theta)$ is known up to a finite dimensional parameter θ, and the estimation is usually via the maximum likelihood (ML) in P_θ. In a constraint-free nonparametric analysis, ML over "arbitrary" functions results in a delta sum estimator of η corresponding to the Kaplan-Meier estimator of the survival function. See, e.g., Kalbfleisch and Prentice (1980) and Fleming and Harrington (1991).

Strict constraints and no constraint represent two extremes on the spectrum of bias-variance tradeoff. To strike a compromise between the two extremes, smooth function models with soft constraints are needed. Defining a quadratic roughness functional $J(\eta)$ with a finite dimensional null space $J_\perp$, a convenient way of specifying smooth function models with soft constraints is via $M_\rho = \{\eta : J(\eta) \le \rho\}$ for some $\rho \ge 0$. An example of $J(\eta)$ is $\int \ddot{\eta}^2$ which has the linear polynomials as $J_\perp$. When $\rho = 0$, the model M_0 reduces to a parametric model with $P_\theta = J_\perp$. When $\rho = \infty$, the model M_∞ is usually "arbitrary" so delta sum results. For $\rho \in (0,\infty)$, the ML estimator over M_ρ usually falls on the sphere $\{\eta : J(\eta) = \rho\}$, and Lagrange method converts the constrained ML problem into a penalized likelihood problem

$$\min \; -\frac{1}{n}\sum_{i=1}^{n}\{\delta_i\eta(X_i) - \int_0^{X_i} e^\eta\} + \frac{\lambda}{2}J(\eta), \tag{1.1}$$

where the first term is the minus log likelihood and the Lagrange multiplier λ is called a smoothing parameter. $\lambda = \infty$ corresponds to $\rho = 0$ and $\lambda = 0$ to $\rho = \infty$. The minimization of (1.1) is implicitly over $\mathcal{H} = \{\eta : J(\eta) < \infty\}$. $J(\eta)$ forms a natural square (semi) norm in $\mathcal{H}$ and, supplemented by a square norm in $J_\perp$, makes $\mathcal{H}$ a Hilbert space. It shall be assumed that evaluation is continuous in $\mathcal{H}$ so the likelihood part of (1.1) is continuous. See O'Sullivan (1988) and Gu (1991b).

Penalized likelihood method was introduced by Good and Gaskins (1971) in the context of nonparametric probability density estimation. Its use in hazard estimation was proposed by Anderson and Senthilselvan (1980), Bartoszynski, Brown, McBride and Thompson (1981), and O'Sullivan (1988). The asymptotic convergence rates of the solution $\hat{\eta}$ of (1.1) were calculated by Cox and O'Sullivan (1990) and Gu (1991b). An algorithm for computing a B-spline approximation of $\hat{\eta}$ was proposed and studied by O'Sullivan (1988). A computable data-adaptive approximation $\hat{\eta}_n$ of $\hat{\eta}$ was proposed by Gu (1991b) and shown to share the same asymptotic convergence rates as $\hat{\eta}$.

The purpose of this article is to examine an automatic algorithm for calculating $\hat{\eta}_n$ and to study the practical performance of the methodology with the help of the algorithm. The algorithm is identical in structure to an algorithm of Gu (1991a) in the context of density estimation. A key ingredient in the algorithm is a simple cross-validation scheme based on the martingale structure of censored data. The remaining of the article is organized as follows. Section 2 reviews some background theoretical results, sets up the numerical problem, and discusses the algorithm. Section 3 discusses the cross-validation scheme. Section 4 presents simulation results to demonstrate various aspects of the algorithm and the methodology. Section 5 illustrates an application in data analysis.

2 Formulation and Algorithm

2.1 Theoretical background

Let η_0 be the true hazard and assume $\eta_0 \in \mathcal{H}$. Define $\mathrm{SKL}(\hat{\eta}, \eta_0) = \int_0^\infty (e^{\hat{\eta}} - e^{\eta_0})(\hat{\eta} - \eta_0)\tilde{S}$ and $V(\eta) = \int \eta^2 e^{\eta_0}\tilde{S}$, where $\tilde{S}(t) = \mathrm{Prob}(X \geq t)$ is the survival function of X. $\mathrm{SKL}(\hat{\eta}, \eta_0)$ is an appropriately weighted symmetrized Kullback-Leibler between $\hat{\eta}$ and η_0 and $V(\hat{\eta} - \eta_0)$ is its quadratic approximation. Under certain conditions, it can be shown that

$$\mathrm{SKL}(\hat{\eta}, \eta_0) \sim V(\hat{\eta} - \eta_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda), \tag{2.1}$$

where the r codes the smoothness implied by $J(\eta)$. For $J(\eta) = \int \ddot{\eta}^2$, $r = 4$. See Gu (1991b).

A Hilbert space in which evaluation is continuous is known as a reproducing kernel Hilbert space possessing a reproducing kernel, a positive-definite bivariate function R with the reproducing property that $\langle R(t, \cdot), \eta \rangle = \eta(t)$, where $\langle \cdot, \cdot \rangle$ is the inner product in the space; see, e.g., Wahba (1990, Chapter 1). Given a square norm in $J_\perp$, $\mathcal{H}$ has a tensor sum decomposition such that J is a square norm in $\mathcal{H} \ominus J_\perp$. Let R_J be the reproducing kernel in the space $\mathcal{H} \ominus J_\perp$ with J as the inner product, $\mathcal{H}_n = J_\perp \oplus \{R_J(X_i, \cdot), \delta_i = 1\}$, and $\hat{\eta}_n$ be the minimizer of (1.1) in $\mathcal{H}_n$. With a mild further condition in extra to what are needed for (2.1), it was shown in Gu (1991b) that $\mathrm{SKL}(\hat{\eta}_n, \eta_0) \sim V(\hat{\eta}_n - \eta_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$. Note that $\hat{\eta}_n$ is computable but $\hat{\eta}$ is not, while they share the same asymptotic convergence rates. This article is about the computation of $\hat{\eta}_n$.

2.2 Numerical preliminaries

Let $N = \sum_{i=1}^{n} \delta_i$ and let T_i, $i = 1, \cdots, N$, denote the observed failure times. Write $\{\phi_\nu\}_{\nu=1}^{M}$ as a basis of $J_\perp$ and $\xi_i = R_J(T_i, \cdot)$. By definition, a function in $\mathcal{H}_n$ has an expression

$$\eta = \sum_{i=1}^{N} c_i \xi_i + \sum_{\nu=1}^{M} d_\nu \phi_\nu = \boldsymbol{\xi}^T \boldsymbol{c} + \boldsymbol{\phi}^T \boldsymbol{d}, \tag{2.2}$$

where $\boldsymbol{\xi}$ and $\boldsymbol{\phi}$ are vectors of functions and $\boldsymbol{c}$ and $\boldsymbol{d}$ are vectors of coefficients. Substituting (2.2) into (1.1), noting that

$$J(g) = \langle \sum_{i=1}^{N} c_i \xi_i \,, \sum_{j=1}^{N} c_j \xi_j \rangle = \sum_{i=1}^{N} \sum_{j=1}^{N} c_i c_j R_J(T_i, T_j) \tag{2.3}$$

where $\langle R_J(t, \cdot), R_J(s, \cdot) \rangle = R_J(t, s)$ is used, the problem becomes to minimize

$$A_\lambda(\boldsymbol{c}, \boldsymbol{d}) = -\frac{1}{n} \mathbf{1}^T (Q\boldsymbol{c} + S\boldsymbol{d}) + \int \bar{Y} \exp(\boldsymbol{\xi}^T \boldsymbol{c} + \boldsymbol{\phi}^T \boldsymbol{d}) + \frac{\lambda}{2} \boldsymbol{c}^T Q \boldsymbol{c}, \tag{2.4}$$

with respect to $\boldsymbol{c}$ and $\boldsymbol{d}$, where Q is $N \times N$ with (i, j)th entry $\xi_i(T_j) = R_J(T_i, T_j)$, S is $N \times M$ with (i, ν)th entry $\phi_\nu(T_i)$, and $\bar{Y} = (1/n) \sum_{i=1}^{n} Y_i = (1/n) \sum_{i=1}^{n} I_{[X_i \geq t]}$ is the empirical survival function of X.

Let $\mu_\eta(h) = \int h \bar{Y} e^\eta$, $V_\eta(f, h) = \mu_\eta(fh)$, and $V_\eta(h) = V_\eta(h, h)$. Note that the empirical survival function $\bar{Y}$ is used in stead of $\tilde{S}$ in the definition of V_η here. Write $\tilde{\eta} = \boldsymbol{\xi}^T \tilde{\boldsymbol{c}} + \boldsymbol{\phi}^T \tilde{\boldsymbol{d}}$ as the current iterate of η. For fixed λ, the one-step Newton update for minimizing (2.4) can be shown to satisfy (cf. Gu, 1991a, §2)

$$\begin{pmatrix} V_{\xi,\xi} + \lambda Q & V_{\xi,\phi} \\ V_{\phi,\xi} & V_{\phi,\phi} \end{pmatrix} \begin{pmatrix} \boldsymbol{c} \\ \boldsymbol{d} \end{pmatrix} = \begin{pmatrix} Q\mathbf{1}/n - \mu_\xi + V_{\xi,\eta} \\ S^T \mathbf{1}/n - \mu_\phi + V_{\phi,\eta} \end{pmatrix}, \tag{2.5}$$

where $V_{\xi,\xi}$ is $N \times N$ with (i, j)th entry $V_{\tilde{\eta}}(\xi_i, \xi_j)$, $V_{\xi,\phi}$ $N \times M$ with (i, ν)th entry $V_{\tilde{\eta}}(\xi_i, \phi_\nu)$, $V_{\phi,\phi}$ $M \times M$ with (ν, μ)th entry $V_{\tilde{\eta}}(\phi_\nu, \phi_\mu)$, $V_{\xi,\eta}$ $N \times 1$ with ith entry $V_{\tilde{\eta}}(\xi_i, \tilde{\eta})$, $V_{\phi,\eta}$ $M \times 1$ with νth entry $V_{\tilde{\eta}}(\phi_\nu, \tilde{\eta})$, μ_ξ $N \times 1$ with ith entry $\mu_{\tilde{\eta}}(\xi_i)$, and μ_ϕ $M \times 1$ with νth entry $\mu_{\tilde{\eta}}(\phi_\nu)$.

2.3 Algorithm

A choice of λ selects from among $\{M_\rho, \rho \in (0, \infty)\}$ a smooth model, and a proper bias-variance tradeoff via smoothing parameter selection determines the performance of the estimator. Among natural performance criteria are $\mathrm{SKL}(\hat{\eta}_n, \eta_0)$ and $V(\hat{\eta}_n - \eta_0)$, where for practicality we shall estimate

the survival function $\tilde{S}$ appearing in SKL and V by the empirical survival function $\bar{Y}$. From $\tilde{\eta}$, the one-step Newton update of (2.5) provides a group of estimates with a varying λ, and we shall try to select a well performing update via a proper choice of λ. Based on $\tilde{\eta}$, $L_{\tilde{\eta}}(\eta, \eta_0) = V_{\tilde{\eta}}(\eta)/2 - V_{\tilde{\eta}}(\eta, \tilde{\eta}) + \mu_{\tilde{\eta}}(\eta) - \mu_{\eta_0}(\eta)$ can be shown to be a proxy of $\mathrm{SKL}(\eta, \eta_0)$ or $V(\eta - \eta_0)$ (cf. Gu, 1991a, §3), where all terms are computable except $\mu_{\eta_0}(\eta)$. When an estimate of $\mu_{\eta_0}(\eta)$ is available, a performance-oriented iteration can be conducted to jointly update (λ, η) by choosing λ to minimize $\hat{L}_{\tilde{\eta}}(\eta, \eta_0)$ among the one-step Newton updates, where $\hat{L}_{\tilde{\eta}}(\eta, \eta_0)$ is $L_{\tilde{\eta}}(\eta, \eta_0)$ with an estimated $\mu_{\eta_0}(\eta)$; see Gu (1991a, §3). Hiding the differences under the definitions of quantities appearing in (2.5), and using a certain estimate of $\mu_{\eta_0}(\eta)$ which we shall discuss in §3, the computing formulas and the algorithm in Gu (1991a, §§3-4) hold verbatim for hazard estimation. We will not duplicate further details here, but only note that the algorithm takes as inputs the following information.

1. *Data*: The observed failure times T_i and the empirical survival function $\bar{Y}$ of X.

2. *A class of models*: The reproducing kernel $R_J(t, s)$ and the null space basis $\phi_\nu(t)$.

3. *Base measure*: A quadrature formula (mesh points and weights) for calculating integrals.

If converges, the algorithm returns $\boldsymbol{c}$ and $\boldsymbol{d}$ associated with an automatic λ as the estimate.

3 Cross-Validation

We now discuss the estimation of $\mu_{\eta_0}(\eta) = \int_0^\infty \eta \bar{Y} e^{\eta_0}$ for η a one-step Newton update from $\tilde{\eta}$. We shall adopt the notation definitions of Gu (1991a, §§3-4) up to the quantities appearing in (2.5). To quote (3.1) of Gu (1991a),

$$\eta = \boldsymbol{\xi}^T H^{-1}(Q\mathbf{1}/n) + \boldsymbol{\xi}^T H^{-1}\boldsymbol{v}_\xi + (\boldsymbol{\phi} - V_{\phi,\xi}H^{-1}\boldsymbol{\xi})^T E^{-1}\boldsymbol{u}_{\phi|\xi} = h_1 + h_2 + h_3.$$

We shall illustrate that the formulas (3.4) and (3.5) in Gu (1991a) estimate $\mu_{\eta_0}(h_2+h_3)$ and $\mu_{\eta_0}(h_1)$, respectively.

Let $N(t) = I_{[X \le t, \delta=1]}$ and $A(t) = \int_0^t Y(u)e^{\eta_0(u)}du$ where $Y(t) = I_{[X \ge t]}$. Under independent censorship, $M(t) = N(t) - A(t)$ is a martingale. Given a predictable function h on $[0, \infty)$, the Stieltjes integral $\int_0^t h(u)dM(u)$ is also a martingale under mild conditions. See, e.g., Fleming and Harrington (1991, §2.7). A deterministic (meaning independent of $M(t)$) continuous function is

predictable, and in practice ϕ_ν and ξ_i are usually chosen to be continuous. We shall use the martingale moment property to estimate $\mu_{\eta_0}(\eta)$. Specifically, since $E(\int_0^\infty h dM) = 0$, one may use $\int_0^\infty h d\bar{M}$ to "estimate" 0 where $\bar{M} = (1/n)\sum_{i=1}^n M_i$, which leads to estimating $\int_0^\infty h\bar{Y}e^{\eta_0}$ by $\int_0^\infty h d\bar{N} = (1/n)\sum_{i=1}^n \delta_i h(X_i)$ where $\bar{N} = (1/n)\sum_{i=1}^n N_i$. Applying this to estimate $\mu_{\eta_0}(h_2 + h_3)$ yields (3.4) of Gu (1991a).

We need a simple cross-validation for estimating $\mu_{\eta_0}(h_1)$. Write $h_1 = (1/n)\sum_{i=1}^n \delta_i \boldsymbol{\xi}^T H^{-1}\boldsymbol{\xi}(X_i) = (1/n)\sum_{i=1}^n \tilde{h}_i$. Note that $\tilde{h}_i$ is dependent on M_i so $\int_0^t \tilde{h}_i(u)dM_i(u)$ is not a martingale. To estimate $\int_0^\infty \tilde{h}_i\bar{Y}e^{\eta_0}$ for $\delta_i = 1$, we shall first approximate $\bar{Y}$ by $\bar{Y}_{(i)}$ where $\bar{Y}_{(i)} = \sum_{j\neq i} Y_j/(n-1)$, and then use the martingale moment estimate $\sum_{j\neq i}\delta_j\tilde{h}_i(X_j)/(n-1)$ for $\int_0^\infty \tilde{h}_i\bar{Y}_{(i)}e^{\eta_0}$. Equation (3.5) of Gu (1991a) results after collecting terms, and formula (3.6) and Algorithm 4.1 of Gu (1991a) follow.

4 Simulations

Simulation results are presented in this section to demonstrate various aspects of the automatic algorithm and the methodology in general. A real data application will be presented in the next section.

Failure times T_i were generated from a Weibull distribution with a hazard $e^{\eta_0(t)} = 24t^2$. Independent censoring times C_i were generated from a truncated exponential distribution with a survival function $\text{Prob}(C \geq t) = e^{-t/3}$, $t \leq 1$, and $\text{Prob}(C \geq 1+) = 0$. The models were specified via $M_\rho = \{\eta : J(\eta) = \int_0^1 \ddot{\eta}^2 < \rho\}$, where the J has a null space $J_\perp = \{1, t\}$. Taking $(\int_0^1 \eta)^2 + (\int_0^1 \dot{\eta})^2$ as a square norm in $J_\perp$, $\mathcal{H}\ominus J_\perp = \{\eta : \int_0^1 \eta = \int_0^1 \dot{\eta} = 0, \int_0^1 \ddot{\eta}^2 < \infty\}$ and $R_J(t,s) = k_2(t)k_2(s) - k_4(|t-s|)$, where $k_2 = (k_1^2 - 1/12)/2$, $k_4 = (k_1^4 - k_1^2/2 + 7/240)/24$, and $k_1 = (\cdot - .5)$; see, e.g., Craven and Wahba (1979). Integrals defining $\text{SKL}(\hat{\eta}_n, \eta_0)$, $V(\hat{\eta}_n - \eta_0)$, and quantities in (2.5) were calculated by averaging the integrands over 300 equally spaced mesh points on $(0, 1)$.

For each of three sample sizes $n = 100$, $n = 150$, and $n = 200$, one hundred replicates of data were generated as described above. The number of failures N averaged to 86.7, 129.7, and 172.8, respectively, for the three sets of replicates. These groups of replicates shall be labeled by their (average) N/n ratios as 87/100, 130/150, and 173/200. The automatic algorithm converged on all but one $n = 200$ data set. $\text{SKL}(\hat{\eta}_n, \eta_0)$ and $V(\hat{\eta}_n - \eta_0)$ were evaluated at all the converged automatic fits, where the empirical survival function $\bar{Y}$ was used instead of $\tilde{S}$ so the definitions of SKL and V vary slightly from data to data. Fixed-λ solutions of (2.4) were also calculated on a

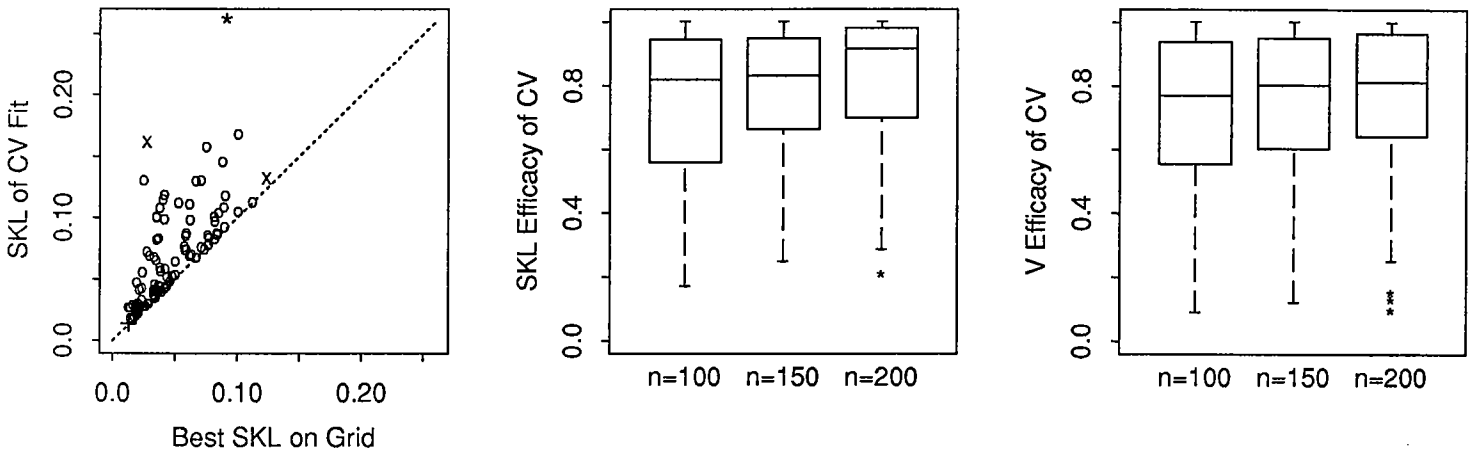


Figure 4.1: The efficacy of cross-validation in the automatic algorithm.

grid $\log_{10}\lambda = (-6)(.1)(-2.5)$ and SKL and V evaluated.

The relative effectiveness of the cross-validation smoothing parameter selection is illustrated in Figure 4.1. The left frame of Figure 4.1 plots the SKL of the automatic fit against the smallest SKL obtained on the grid for the $n = 100$ replicates, where a point on the dotted line indicates a perfect performance of the automatic algorithm; the best automatic fit and the poorest automatic fit are plotted as a plus and a star and two other points are plotted as crosses. The center and right frames of Figure 4.1 present the relative efficacy of the cross-validation in SKL score and in V score for the $n = 100$, $n = 150$, and $n = 200$ replicates, where the efficacy is defined as the smallest score on the grid divided by the score evaluated at the automatic fit. Slightly improved relative efficacy associated with larger sample sizes is visible.

To check on the effect of censoring on the relative performance of cross-validation, I applied heavier censoring on the $n = 150$ and $n = 200$ replicates by changing the censoring survival function to $e^{-4t/3}$ and e^{-2t} for $t \leq 1$, respectively. This caused the number of failures N average to 84.5 and 87.4, which are comparable to the 86.7 of the $n = 100$ replicates with a censoring survival function $e^{-t/3}$. Censoring with a survival function e^{-t}, $t \leq 1$, was also applied to the $n = 200$ replicates which made N average to 129.7. These replicates are labeled as 85/150, 87/200, and 130/200, respectively. The SKL efficacy and V efficacy of cross-validation on these replicates are plotted in Figure 4.2 together with those of the other three groups of replicates. It appears that the sample

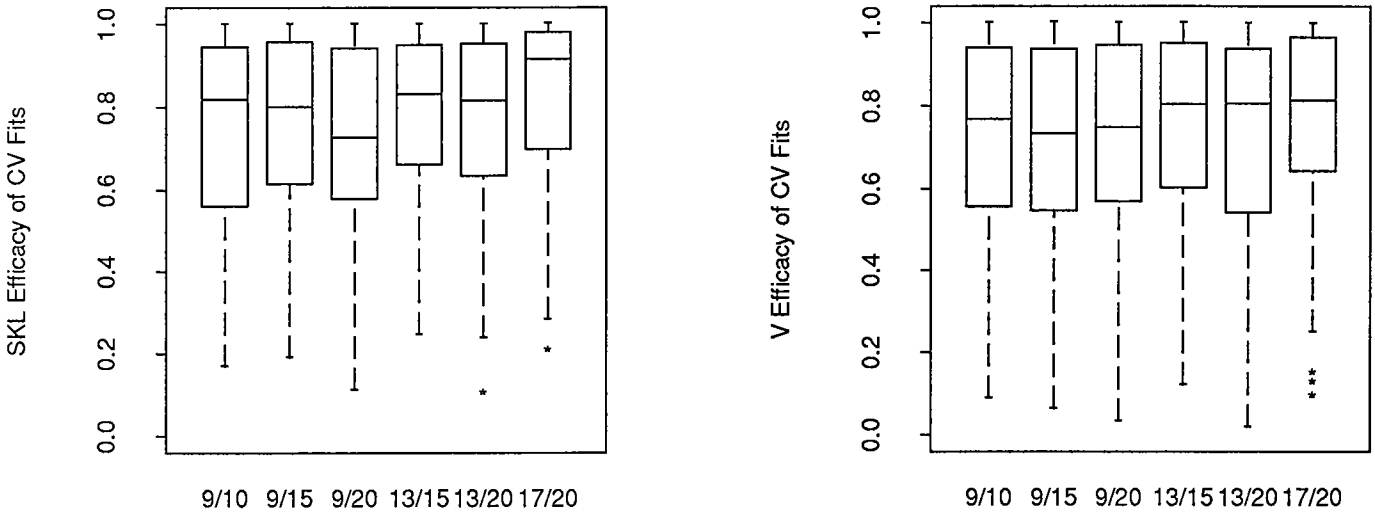


Figure 4.2: A comparison of cross-validation efficacy at different N/n ratios. The last digit in the N/n label is rounded.

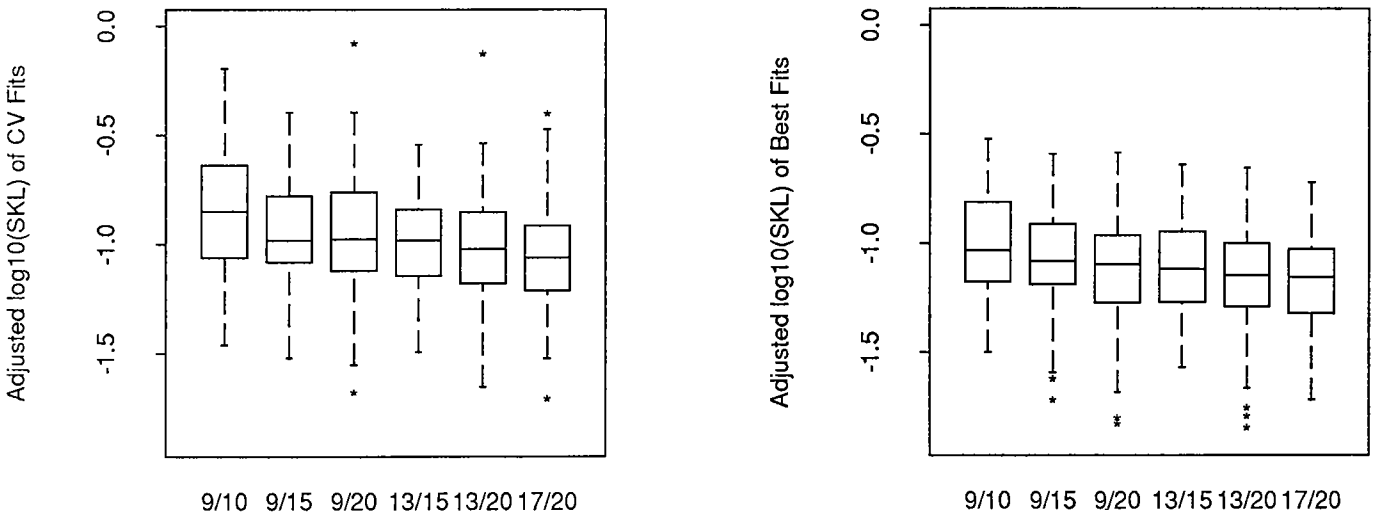


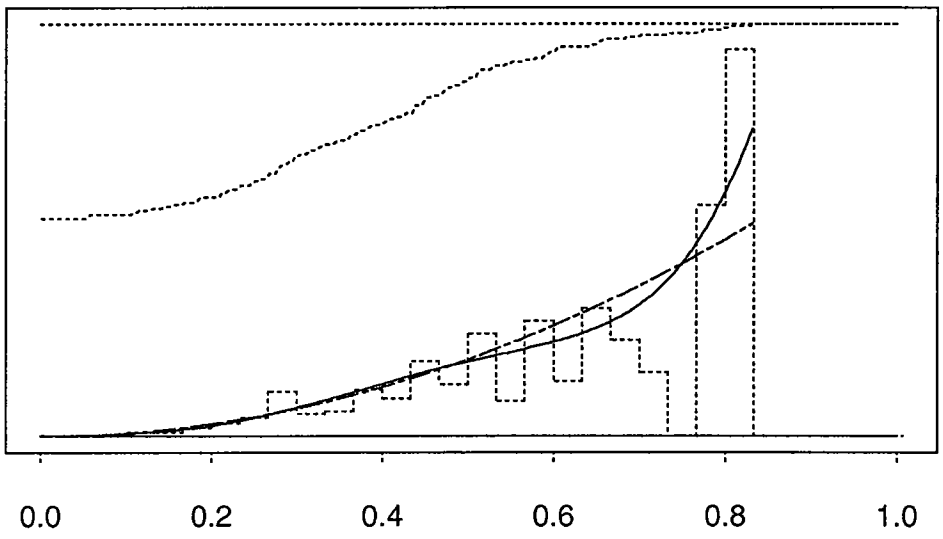
Figure 4.3: A comparison of absolute performance of the estimates at different N/n ratios. The last digit in the N/n label is rounded.

size n doesn't have a clear impact on the relative efficacy of cross-validation, but the number of failures N, and to a less extent the censoring ratio $1 - N/n$, might have slight bearing on it. The mechanism behind it is not understood yet.

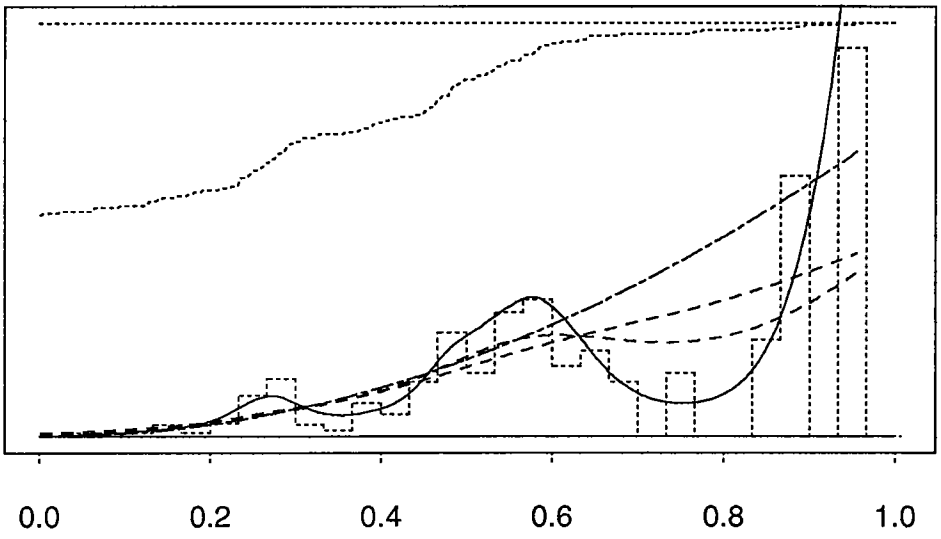
We now look at the absolute performance of the method at different N/n ratios. Since the empirical survival function $\bar{Y}$ of X, which serves as a weight function in the definition of SKL and V scores, varies randomly from case to case within the same group and varies systematically between the groups, the scores are not quite directly comparable. No attempt is made here to correct for the random variation, but intuitively the systematic variation may be adjusted to an extent by dividing the scores by the average total weight $\int_0^1 \tilde{S}$, where $\tilde{S}$ is the theoretical survival function of X under the corresponding censorship. After such an adjustment, $\log_{10}$ SKL of the automatic fits and of the best fits on grid in the six groups of replicates are plotted in Figure 4.3. It appears that the sample size n and the censoring ratio $1 - N/n$ have some influence on the precision of the estimates, as expected.

To perceive the practical performance of the method, the automatic fits corresponding to the plus and the star in the left frame of Figure 4.1 are plotted in the top and bottom frames of Figure 4.4 as solid lines with the true hazard superimposed as dot-dash lines. The lines extend only to a point beyond which the empirical survival function $\bar{Y}$ is zero. The data are superimposed as dotted lines in forms of the empirical survival function $\bar{Y}$ of X_i (upside down from top of the frames, in inflated scale) and the empirical hazard of discretized data (bar plots). For the "poorest" fit, the minimum-SKL fit and minimum-V fit on the grid are also superimposed as dashed lines. It can be seen that the "poorest" automatic fit fits the data reasonablely well, but the data appear to be rather atypical given the test hazard. I also looked at two other poor fits, those marked as crosses in the left frame of Figure 4.1, in Figure 4.5. The plots are constructed in the same manner as in Figure 4.4. The "improvable" automatic fit demonstrates a plausible fit to the data were the "truth" not given. The one with a nearly perfect relative efficacy virtually sits in the null space $J_\perp$, and in fact, the fit calculated from this data set is insensitive to the choice of λ over a broad range.

In summary, our limited experiments indicate that the relative efficacy of cross-validation improves slightly as the number of failures N increases, the absolute performance improves as the sample size n increases, and in general the method fits data well.

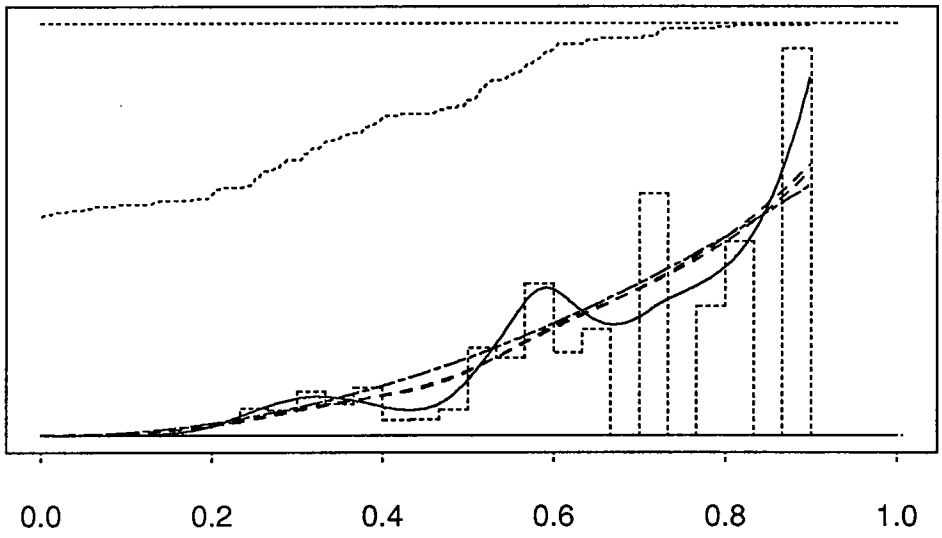


A Good Automatic Fit (N=84)

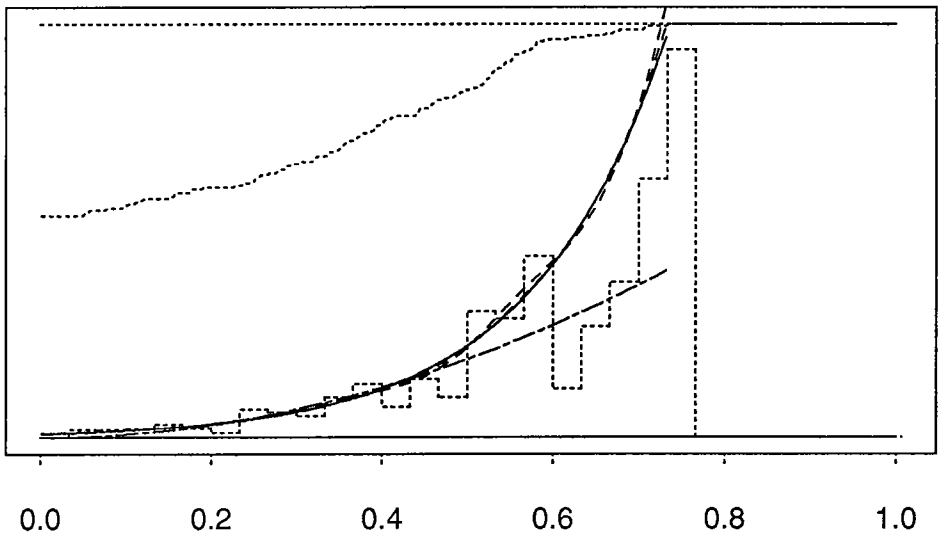


The 'Poorest' Automatic Fit (N=88)

Figure 4.4: A good fit and the "poorest" fit in the $n = 100$ simulations. The solid line is the automatic fit and the dot-dash line the truth. The dashed lines are SKL-optimal and V-optimal fits. The dotted lines represent data.



An 'Improvable' Poor Fit (N=81)



A 'Best Possible' Poor Fit (N=90)

Figure 4.5: Two other poor fits in the $n = 100$ simulations. The solid line is the automatic fit and the dot-dash line the truth. The dashed lines are SKL-optimal and V-optimal fits. The dotted lines represent data.

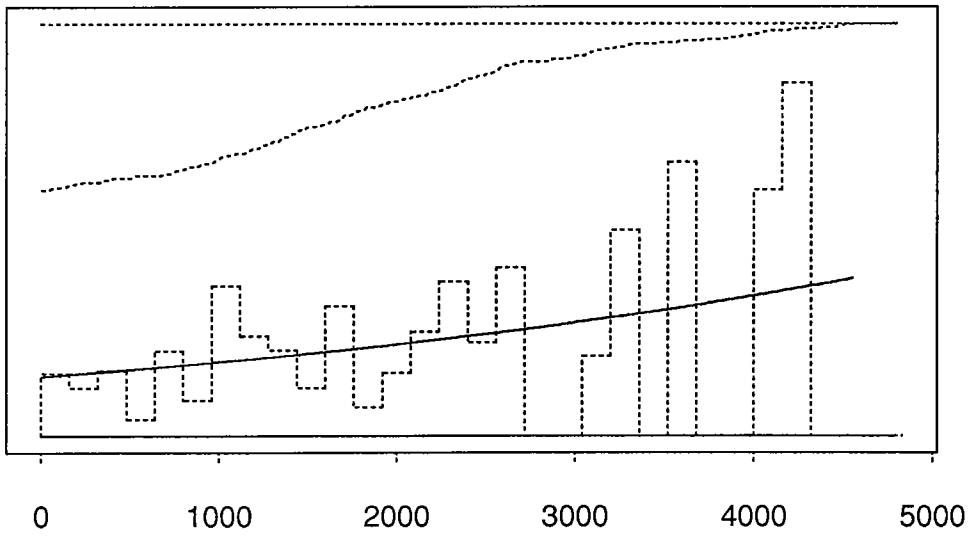
5 Application

Between January, 1974 and May, 1984, the Mayo Clinic conducted a double-blinded randomized trial in primary biliary cirrhosis of the liver (PBC), comparing the drug D-penicillamine (DPCA) with a placebo. The data are tabulated in Fleming and Harrington (1990) with a concise description. Of the 312 patients participated in the trial, 158 were treated with DPCA with 65 recorded deaths, and 154 were treated with the placebo with 60 recorded deaths. The trial lasted nearly 4800 days. I mapped $[0, 4800]$ onto $[0, 1]$ and applied the automatic algorithm to the data using the same $J_\perp$ and R_J as specified in §4. Separate estimation for the two groups of patients yields the fits presented in Figure 5.1 with the data superimposed in the same manner as in Figures 4.4 and 4.5. The automatic fit for the DPCA hazard virtually sits in $J_\perp$. The dashed line in the placebo frame indicates the corresponding parametric fit in $J_\perp$. The hazard pattern demonstrated by the placebo automatic fit asks for some explanation. Combining all the 312 patients together, the automatic fit again sits in the null space $J_\perp$ which is plotted in Figure 5.2 with the combined data, the DPCA fit, and the placebo fits superimposed. It appears that DPCA might have slightly reduced the hazard in the first three years or so during the treatment, but from about three and half years onward until about eight and half years, the DPCA treated patients had higher hazard compared with those in the placebo group. This, of course, could be due to the delayed failures of weak patients under the DPCA treatment.

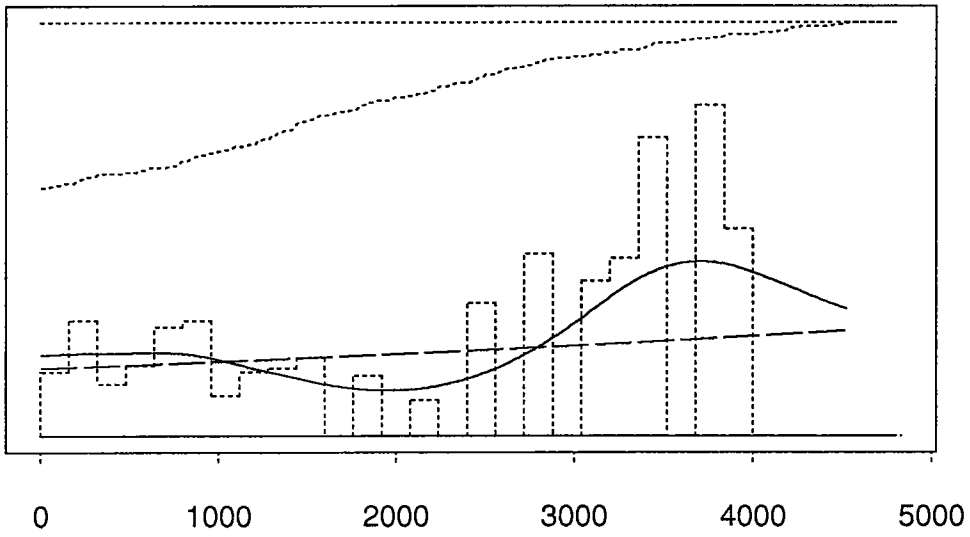
6 Concluding Remarks

In this article, we have studied an implementation of penalized likelihood hazard estimation, where the algorithm, with a cross-validation performance-oriented iteration, is adapted from an earlier similar algorithm on density estimation. Portable code is available from me at `chong@stat.purdue.edu`. The current implementation is generic and is of order $O(N^3)$. Implementations of order $O(N)$ should be possible, for specific configurations separately, by using local basis expressions instead of (2.2) for $\hat{\eta}_n$ (cf. O'Sullivan 1988).

As noticed from the simulation results but not reported in §4, the SKL-optimal λ and the V-optimal λ in hazard estimation are often far apart, which reflects the fact that $(e^{\hat{\eta}} - e^{\eta_0}) \approx e^{\eta_0}(\hat{\eta} - \eta_0)$ is often a rather wild approximation. For the same reason, the proximity argument in §2.3 is



DPCA Treatment (N=65/n=158)



Placebo Treatment (N=60/n=154)

Figure 5.1: Hazard estimation of PBC patients. Estimated hazards are in solid lines. Data are in dotted lines. Parametric placebo hazard is in dashes.

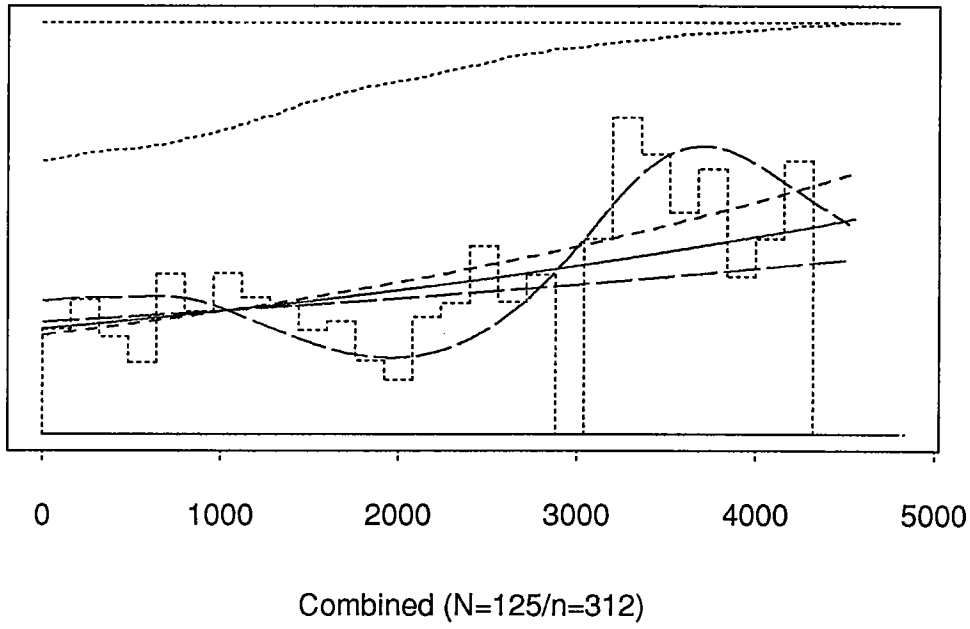


Figure 5.2: Combined hazard of PBC patients and treatment comparison. Combined hazard is in solid line. Combined data are in dotted lines. The DPCA hazard is in short dashes. The placebo hazards are in long dashes.

less plausible than its counterpart in density estimation. Consequently, the automatic smoothing parameter works less effectively in hazard estimation than in density estimation, as can be seen by comparing Figure 4.1 with similar figures in Gu (1991a). Hazard estimation might just be a more difficult problem, and the current algorithm appears serviceable. However, much better empirical results, even better than the corresponding density estimation results, have been reported in O'Sullivan (1988) on an AIC score applied to hazard estimation. It remains to be understood how O'Sullivan's AIC score works.

Of more interest than plain hazard estimation is censored data regression, for which proportional hazard models remain the prime tool in applications (cf. Kalbfleisch and Prentice, 1980; Fleming and Harrington, 1991). Hazard proportionality may sometimes be violated in practice, however, so more flexible models are needed. Inserting tensor product splines into the current structure, it is possible to construct nonparametric models more general than the proportional hazard models for censored data regression. Details on this line are to be developed in future works.

References

Anderson, J. A. and Senthilselvan, A. (1980). Smooth estimates for the hazard function. *J. Roy. Statist. Soc.* B **42**, 322 – 327.

Bartoszynski, R., Brown, B. W., McBride, C. M., and Thompson, J. R. (1981). Some nonparametric techniques for estimating the intensity function of a cancer related nonstationary Poisson process. *Ann. Statist.* **9**, 1050 – 1060.

Cox, D. D. and O'Sullivan, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *Ann. Statist.* **18**, 1676 – 1695.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31**, 377 – 403.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis.* Wiley, New York.

Gu, C. (1991a). Smoothing spline density estimation: A dimensionless automatic algorithm. Technical Report 91-41, Purdue University, Dept. of Statistics.

—— (1991b). Penalized likelihood hazard estimation. Technical Report 91-58, Purdue University, Dept. of Statistics.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* Wiley, New York.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM J. Sci. Statist. Comput.* **9**, 363 – 379.

Wahba, G. (1990). *Spline Models for Observational Data.* CBMS–NSF Regional Conference Series in Applied Mathematics, Vol. 59. SIAM, Philadelphia.