Published in final edited form as:
Stat Comput. 2013 September 1; 23(5): 601–614. doi:10.1007/s11222-012-9333-9.

A Tutorial on Rank-based Coefficient Estimation for Censored Data in Small- and Large-Scale Problems

Matthias Chung,
Department of Mathematics, Texas State University, San Marcos, TX 78666, U.S.A.

Qi Long, and
Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, U.S.A.

Brent A. Johnson
Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, U.S.A.

Matthias Chung: mc85@txstate.edu

Abstract

The analysis of survival endpoints subject to right-censoring is an important research area in statistics, particularly among econometricians and biostatisticians. The two most popular semiparametric models are the proportional hazards model and the accelerated failure time (AFT) model. Rank-based estimation in the AFT model is computationally challenging due to optimization of a non-smooth loss function. Previous work has shown that rank-based estimators may be written as solutions to linear programming (LP) problems. However, the size of the LP problem is $O(n^2 + p)$ subject to n^2 linear constraints, where n denotes sample size and p denotes the dimension of parameters. As n and/or p increases, the feasibility of such solution in practice becomes questionable. Among data mining and statistical learning enthusiasts, there is interest in extending ordinary regression coefficient estimators for low-dimensions into high-dimensional data mining tools through regularization. Applying this recipe to rank-based coefficient estimators leads to formidable optimization problems which may be avoided through smooth approximations to non-smooth functions. We review smooth approximations and quasi-Newton methods for rank-based estimation in AFT models. The computational cost of our method is substantially smaller than the corresponding LP problem and can be applied to small- or large-scale problems similarly. The algorithm described here allows one to couple rank-based estimation for censored data with virtually any regularization and is exemplified through four case studies.

Keywords

Accelerated failure time model; Ill-posed problems; Regularization; Survival analysis

1 Introduction

Survival analysis is a ubiquitous concept in statistics and widely used in biomedical, clinical, and reliability studies. Various semiparametric models and estimators have been proposed for survival analysis. Cox's proportional hazards model (Cox, 1972), for example, has been studied extensively for four decades and is widely used partly due to its ease of computation. While the accelerated failure time (AFT) model (Cox and Oakes, 1984; Kalbeisch and Prentice, 1980) is, as suggested by Sir David Cox, "in many ways more appealing because of its quite direct physical interpretation" as compared to the more popular proportional hazards model (Reid, 1994), it has not been adopted in practice because of theoretical and

computational challenges. Over the past several years, there have been technical and computational advances in this area and we build on this work to present a strategy for simultaneous coefficient estimation and variable selection. The current paper provides a detailed account of a fitting algorithm applied to regularized rank-based coefficient estimation in the semiparametric AFT model for both small- and large-scale problems; that is, where the dimension of the predictors can be smaller or larger than the sample size.

The AFT model asserts that the natural logarithm of the survival endpoint T_i is linearly related to explanatory variables, i.e.,

$$\log T_i = \mathbf{x}_i^\top \beta + \zeta_i, \ (i=1, \ldots, n), \quad (1)$$

where $\mathbf{x}_i$ is a p-vector of fixed predictors for the ith subject, $\boldsymbol{\beta}$ is a p-vector of regression coefficients, and $(\zeta_1, \ldots, \zeta_n)$ are independent and identically distributed errors with an unspecified distribution function. If C_i is a stochastic, subject-specific censoring variable, then the observed data are $\{(U_i, \delta_i, \mathbf{x}_i)\}_{i=1}^n$, where $U_i = \min(T_i, C_i)$, $\delta_i = I(T_i \leq C_i)$ and $I(\cdot)$ is the indicator function. The goal is to estimate the regression coefficients $\boldsymbol{\beta}$ using the observed data. Rank-based coefficient estimation was first proposed by Prentice (1978) but the first general asymptotic theory was not developed for more than decade later (Tsiatis, 1990; Wei et al, 1990) and the most general theory under the weakest conditions appeared a few years later (Ying, 1993). A detailed history of early rank-based methods for censored data is provided elsewhere (Kalbeisch and Prentice, 1980).

From the beginning, rank-based coefficient estimation in the AFT model has been difficult and statistical inference even more challenging. The difficulty in estimation arises from the non-smooth nature of the estimating function. The difficulty in inference arises because the asymptotic slope matrix of the estimating function depends on the hazard function of the errors and cannot be directly estimated from the observed data; thus, the sandwich covariance matrix cannot be directly estimated for statistical inference. The earliest coefficient estimation techniques were based on direct search (Tsiatis, 1990; Wei et al, 1990; Lin and Geyer, 1992) and only truly suitable for low-dimensional problems, i.e., small p. Jin et al (2003) provided the the first reliable and accurate estimation procedure through explicit use of linear programming (LP) techniques to compute the Gehan (1965) estimator, a special version of the weighted logrank estimator. Compared to earlier approaches, this was a substantial improvement due to the accuracy of the method and its availability in standard software packages.

Unfortunately, the LP problem in Jin et al (2003) has $O(n^2 + p)$ unknown parameters subject to n^2 linear constraints and the size of optimization problem can quickly overwhelm many standard LP solvers running on desktop computers. Furthermore, the inference procedure by Jin et al (2003) was based on resampling which meant that a perturbed LP problem of the same dimension as the original LP problem had to be solved multiple times (See Section 6). The computational complexity of the inference procedure by Jin et al (2003) prompted investigators to propose other methods. In particular, Heller (2007) proposed to directly approximate the Heaviside function in the Gehan (1965) estimating function with a distribution function while Brown and Wang (2005, 2007) proposed a pseudo-Bayesian approach which effectively estimates the coefficients and sandwich covariance simultaneously, again based on a smooth estimating function. In both cases, statistical inference can be performed immediately after coefficient estimation because the sandwich matrix is directly estimable.

Building on earlier work for smoothed rank-based methods and the need for practical solutions, here we provide a general tutorial for regularized rank-based coefficient

estimation based on smooth approximation. In particular, we are interested in the optimization problem,

$$\arg\min_{\beta} \{f(\beta)+\mathscr{S}(\lambda,\beta)\}, \quad (2)$$

where $f(\boldsymbol{\beta})$ is a smooth rank-based loss function, λ is a regularization parameter, and $\mathscr{S}(\lambda, \boldsymbol{\beta})$ is a generic penalty function or regularization term. Optimizing (2) for general loss functions is currently a hot topic in the areas of data mining, machine learning, engineering, and computational statistics. One reason for this is that the minimizer of (2) leads to a sparse solution for some convex but non-differentiable penalty functions with singularities at the origin. As a result, the minimizer of (2) also serves the role of a variable selection and model construction procedure at the same time. No author has tackled the specific problem here for $f(\boldsymbol{\beta})$ pertaining to smoothed rank-based loss functions and only three authors have considered (2) for non-smooth $f(\boldsymbol{\beta})$ (Johnson, 2008, 2009a; Xu et al, 2010; Cai et al, 2009).

The objective of the current paper is to provide a tutorial on a general numerical algorithm for smoothed rank-based loss functions $f(\boldsymbol{\beta})$ and various penalty functions $\mathscr{S}(\lambda, \boldsymbol{\beta})$. Where earlier proposals for ℓ_1-regularized rank-based coefficient estimation provided exact solutions (Johnson, 2009a; Xu et al, 2010; Cai et al, 2009), the motivation behind our current approach is to provide a practical numerical solution of low computational complexity. In order to maximize the efficiency of our procedure, we adopt gradient-based Newton methods for minimizing smooth objective functions. In order to minimize computational complexity for ill-posed problems, we adopt limited-memory quasi-Newton algorithms. The algorithm outlined here applies to general smooth rank-based loss functions $f(\boldsymbol{\beta})$ (Heller, 2007; Brown and Wang, 2005, 2007; Johnson and Strawderman, 2009) and current regularizations (Hoerl and Kennard, 1970; Tibshirani, 1996; Zou and Hastie, 2005; Yuan and Lin, 2006; Tibshirani et al, 2005; Zou, 2006; Johnson et al, 2008; Candes and Tao, 2007; Wu et al, 2009).

The contribution of the current paper is two-fold. First, after a brief history of the problem in Section 2, we provide in Section 3 a detailed tutorial on how to compute rank-based coefficient estimates in the AFT model using a smoothed loss function. In addition to reviewing recent trends in this area, we also propose a new estimator derived from polynomial-based smoothing and complements other estimators based on a smoothed loss function (Heller, 2007; Brown and Wang, 2005, 2007; Johnson and Strawderman, 2009). Second, in Section 4, we review regularized rank-based coefficient estimation and provide a tutorial on how to implement these procedures for rank-based estimators in a computationally efficient manner. We demonstrate unregularized and regularized coefficient estimation through three examples in Section 6. This tutorial is comprehensive for the topic and the framework described here may be applied to other loss functions and regularizations.

2 Background

Tsiatis (1990) proposed coefficient estimation through the weighted logrank estimating function,

$$\boldsymbol{\Psi}_{\vartheta}(\beta)=\frac{1}{n}\sum_{i=1}^{n}\delta_i\,\vartheta\,(e_i,\beta)\left\{\mathbf{x}_i-\frac{\sum_{j=1}^{n}I(e_j\geq e_i)\mathbf{x}_j}{\sum_{j=1}^{n}I(e_j\geq e_i)}\right\},$$

where $e_i = \log U_i - \mathbf{x}_i^\top \beta$ and $\vartheta(\cdot, \boldsymbol{\beta})$ is a user-specified, data-dependent, non-negative weight function. Due to the discrete nature of the estimating function, the weighted logrank estimator $\hat{\boldsymbol{\beta}}_\vartheta$ is defined as a zero-crossing of $\boldsymbol{\Psi}_\vartheta(\boldsymbol{\beta})$; that is, $\hat{\boldsymbol{\beta}}_\vartheta$ satisfies,

$$\Psi_{\vartheta,j}(\hat{\beta}-)\Psi_{\vartheta,j}(\hat{\beta}+) \leq 0,$$

for all $j = 1, \ldots, p$. The class of weighted logrank coefficient estimators has been studied extensively in the statistics literature. It is well-known that the weighted logrank coefficient estimator $\hat{\boldsymbol{\beta}}_\vartheta$ is consistent and asymptotically normal, under certain regularity conditions (Tsiatis, 1990; Wei et al, 1990; Ying, 1993). Unfortunately, the estimating function $\boldsymbol{\Psi}_\vartheta(\boldsymbol{\beta})$ is not monotone, in general, and may contain multiple roots, thus, making parameter estimation troublesome. Fygenson and Ritov (1994) showed that the weighted logrank estimating function with Gehan (1965) weight, i.e.,

$$\vartheta(t,\beta) = n^{-1} \sum_{j=1}^{n} I(e_j \geq t), \quad (3)$$

is monotone. In this case, it can be shown that the weighted logrank estimating function simplifies to

$$\boldsymbol{\Psi}_G(\beta) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i (\mathbf{x}_i - \mathbf{x}_j) I(e_i \leq e_j), \quad (4)$$

which is often referred to as the Gehan estimating function. The Gehan estimating function $\boldsymbol{\Psi}_G(\boldsymbol{\beta})$ in (4) is the p-dimensional quasi-gradient of the following convex loss function,

$$f_G(\beta) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i (e_j - e_i) I(e_i \leq e_j), \quad (5)$$

and the Gehan estimator is defined as the minimizer of the objective function $f_G(\boldsymbol{\beta})$,

$$\hat{\beta}_G = \arg\min_{\beta} f_G(\beta). \quad (6)$$

The function $f_G(\boldsymbol{\beta})$ in (5) is a piecewise-linear convex function and the global minimizer $\hat{\boldsymbol{\beta}}_G$ lies in a p-dimensional polytope. Hence, although the objective function is convex, its minimizer may not be unique. Since $f_G(\boldsymbol{\beta})$ is a non-differentiable function, gradient-based optimization methods cannot be applied directly to solve for $\hat{\boldsymbol{\beta}}_G$. In order for the rank-based estimator to be adopted in practice, efficient numerical methods are needed to solve the optimization problem. There are basically three ways to solve (6): direct search, linear programming, or smoothing.

Direct search methods are widely used in fields such as computational biology. Methods such as evolutionary algorithms and the Nelder-Mead algorithm are easy to implement and, therefore, very popular. In addition, bisection can be applied rather straightforwardly and effectively for low-dimensional problems. However, these methods lack of a comprehensive convergence theory and are well known to perform poorly on medium- to high-dimensional problems (Nocedal and Wright, 2006).

A second way to address optimization problem (6) is to reformulate it as a linear programming (LP) problem. Jin et al (2003) were the first authors to make successful use of the LP formulation,

$$\min_{u,\beta} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i u_{ij} \quad (7)$$

subject to: $u_{ij} = (e_j - e_i)$,
$u_{ij} \leq 0$,
for $i, j = 1, \ldots, n$.

Either simplex or interior point methods may be engaged to solve the LP problem in (7). The advantage of the simplex method is that this method provides an exact solution after a finite number of iterations. However, a major drawback is the rate at which the dimension of the optimization problem increases. While the optimization problem (6) deals with p unknown parameters, the LP problem (7) has $O(n^2 + p)$ parameters: one for every u_{ij} pair and one for each coefficient parameter, β_j, $j = 1, \ldots, p$. Furthermore, the LP problem has n^2 linear constraints. When n and/or p are large, the complexity of the method increases and the convergence rate drops dramatically (Nocedal and Wright, 2006). Interior point methods belong to a class of inexact methods and, unlike simplex, they utilize gradient information. While interior point methods are better than simplex for moderately-sized problems (in terms of the sample size n and dimension of predictors p), they are computationally costly for large n and moderate to large p.

3 Smooth Gehan Loss Functions

3.1 Induced Loss Functions

Several authors have noted practical challenges in inferential procedures for estimators derived from non-smooth loss functions. Two recent germane contributions include the monotone estimating function by Heller (2007) and the pseudo-Bayesian method by Brown and Wang (2005, 2007). Both Heller (2007) and Brown and Wang (2005, 2007) cite simplified standard error estimation as a principal motivation for their smoothing procedures.

3.1.1 Brown and Wang (2005, 2007)—Brown and Wang (2005) proposed an intriguing pseudo-Bayesian method of simultaneous coefficient and standard error estimation in non-smooth parameter estimation problems. Brown and Wang (2007) considered the same parameter estimation discussed here and is directly relevant. Recently, Johnson and Strawderman (2009) reviewed the work by Brown and Wang (2005, 2007), provided theoretical justification for the censored data problem (Brown and Wang, 2007), and extended the method to clustered failure time data. Let $\mathbf{Z} \sim N(0, I_p)$ and $\mathbf{\Gamma}$ be a p-dimensional matrix, such that $\|\mathbf{\Gamma}\| = O(1)$, $\mathbf{\Gamma}^2 = \mathbf{\Omega}$, and $\mathbf{\Omega}$ is a symmetric, positive definite matrix. Then, the Brown and Wang (2007) estimating function is the perturbed Gehan estimating function, $\mathbf{\Psi}_B(\boldsymbol{\beta}) = E_Z\{\mathbf{\Psi}_G(\boldsymbol{\beta}) + \mathbf{\Gamma Z}\}$; that is,

$$\mathbf{\Psi}_B(\beta) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i (\mathbf{x}_i - \mathbf{x}_j) \bar{\Phi}\left(\frac{e_i - e_j}{r_{ij}}\right), \quad (8)$$

$\Phi(t)$ is the standard normal cumulative distribution function, $\bar{\Phi}(t) = 1 - \Phi(t)$, and $r_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} \mathbf{\Omega} (\mathbf{x}_i - \mathbf{x}_j)$. Furthermore, the Brown and Wang (2007) coefficient estimator, say

$\hat{\boldsymbol{\beta}}_B$, is consistent and $\sqrt{n}(\hat{\beta}_B - \beta_0)$ converges in distribution to a mean-zero random vector with asymptotic covariance that is automatically computed as part of the estimation procedure (Brown and Wang, 2005, 2007; Johnson and Strawderman, 2009).

The estimator by Brown and Wang (2007) can be shown to minimize a convex loss function as well. Using integration by parts and facts about normal distribution functions, Johnson and Strawderman (2009) showed that the estimating function $\boldsymbol{\Psi}_B(\boldsymbol{\beta})$ has an associated convex loss function for which $\nabla f_B(\boldsymbol{\beta}) = \boldsymbol{\Psi}_B(\boldsymbol{\beta})$; in particular,

$$f_B(\beta) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\left[(e_j - e_i)\Phi\left(\frac{e_j - e_i}{r_{ij}}\right) + r_{ij}\phi\left(\frac{e_j - e_i}{r_{ij}}\right)\right]. \quad (9)$$

3.1.2 Heller (2007)—Heller (2007) proposed a estimating function by smoothing the indicator function in $\boldsymbol{\Psi}_G(\boldsymbol{\beta})$, i.e.,

$$\boldsymbol{\Psi}_H(\beta) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i(\mathbf{x}_i - \mathbf{x}_j)\bar{\mathrm{G}}\left(\frac{e_i - e_j}{a}\right), \quad (10)$$

where $G(t)$ is a cumulative distribution function, $\bar{G}(t) = 1 - G(t)$, and 'a' is a tuning parameter. Heller proved that, under suitable regularity conditions, the solution to $0 = \boldsymbol{\Psi}_H(\boldsymbol{\beta})$, say $\hat{\boldsymbol{\beta}}_H$, was a consistent estimator of $\boldsymbol{\beta}_0$. Moreover, he showed that $\sqrt{n}(\hat{\beta}_H - \beta_0)$ converges in distribution to a normal random vector with mean zero and whose covariance could be directly estimated.

A common and convenient choice of the distribution function is the standard normal distribution, i.e., $G(t) \equiv \Phi(t)$. With this distribution function, one can again use integration by parts to show that $\boldsymbol{\Psi}_H(\boldsymbol{\beta})$ is the p-dimensional gradient of the following convex loss function,

$$f_H(\beta) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\left[(e_j - e_i)\Phi\left(\frac{e_j - e_i}{a}\right) + a\phi\left(\frac{e_j - e_i}{a}\right)\right]. \quad (11)$$

A straightforward calculation confirms that $\nabla f_H(\boldsymbol{\beta}) = \boldsymbol{\Psi}_H(\boldsymbol{\beta})$.

3.2 The Polynomial-smoothed Gehan Loss Function

A third approach for the optimization problem (6), is to approximate the objective function $f_G(\boldsymbol{\beta})$ directly by a *smooth* approximating function. This is a common technique in applied mathematics and has been used for at least six decades (Huber, 1964). The gain of this approach is that we can adopt computationally efficient gradient-based methods to minimize a surrogate loss function.

Define the following smooth approximation to the Gehan loss function,

$$f_{G,\varepsilon}(\beta) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\, c_\varepsilon(e_i - e_j), \quad (12)$$

where c_ε is a sufficiently smooth real-valued function. Here, we choose a polynomial smoothing function

$$c_\varepsilon(z)=\begin{cases} -z, & \text{if } z \le -\varepsilon, \\ -\frac{1}{16\varepsilon^3}(\varepsilon - z)^4+\frac{1}{4\varepsilon^2}(\varepsilon - z)^3 & \text{if } z \in (-\varepsilon, \varepsilon], \\ 0, & \text{if } z>\varepsilon, \end{cases}$$

with sufficiently small but strictly positive ε (e.g., $\varepsilon = 10^{-4}$). As shown in Figure 1, the function $c_\varepsilon(z) = -z$ for all $z \le -\varepsilon$ and $c_\varepsilon(z) = 0$ for all $z > \varepsilon$ and, hence, matches the Gehan loss function exactly for $|z| > \varepsilon$. The smoothing takes place within the interval $(-\varepsilon, \varepsilon]$. Straightforward calculations reveal that the function c_ε and its first two derivatives $c_\varepsilon^{'}, c_\varepsilon^{''}$ are continuous in the points $-\varepsilon$ and ε for any $\varepsilon > 0$. Hence the loss function $f_{G,\varepsilon}(\boldsymbol{\beta})$ is twice-differentiable in $\boldsymbol{\beta}$ for $\varepsilon > 0$. Note, $f_{G,\varepsilon}(\boldsymbol{\beta})$ also inherits convexity from c_ε for every $\varepsilon \ge 0$. Given our definition of c_ε, it is evident that $\lim_{\varepsilon \to 0} f_{G,\varepsilon}(\boldsymbol{\beta}) = f_G(\boldsymbol{\beta})$.

The estimator is defined as the minimizer of the polynomial-smoothed objective function, i.e.,

$$\hat{\beta}_{G,\varepsilon}=\arg\min_{\beta} f_{G,\varepsilon}(\beta). \quad (13)$$

The following theorem establishes the main consistency result.

Theorem 1 *Under Conditions A1–A4 in* Johnson and Strawderman (2009, *p.586),* $\hat{\boldsymbol{\beta}}_{G,\varepsilon}$ *is a strongly consistent estimator of* $\boldsymbol{\beta}_0$.

The proof of Theorem 1 as well as other large sample results are outlined in the Appendix. The result is a direct consequence of a strong law of large numbers for *U*-statistics.

Remark 1. Although asymptotic analysis suggests ε decreases as n increases, here, we simply view ε as a tuning constant that weighs two objectives: numerical accuracy versus the speed of algorithmic convergence. We used $\varepsilon = 10^{-4}$ in numerous real and simulated examples and found this value to a suitable rule-of-thumb. In statistical computing, it is not uncommon that algorithms include fixed tuning constants; see, for example, MM algorithms (Hunter and Lange, 2004).

3.3 Connections, Contrasts

To facilitate comparisons to other estimators, the first-order partial derivatives of $f_{G,\varepsilon}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ leads to the monotone estimating function,

$$\boldsymbol{\Psi}_{G,\varepsilon}(\beta)=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i(\mathbf{x}_i - \mathbf{x}_j)K_\varepsilon(e_i - e_j), \quad (14)$$

where $K_\varepsilon(z)= - c_\varepsilon^{'}(z)$,

$$K_\varepsilon(z)=\begin{cases} 1, & \text{if } z \le -\varepsilon, \\ -\frac{1}{4\varepsilon^3}(\varepsilon - z)^3+\frac{3}{4\varepsilon^2}(\varepsilon - z)^2 & \text{if } z \in (-\varepsilon, \varepsilon], \\ 0, & \text{if } z>\varepsilon. \end{cases}$$

Evidently, $K_\varepsilon(z)$ is a weight function operating on the differences in residuals $(e_i - e_j)$: the weight is 1 if $(e_i - e_j) < -\varepsilon$, 0 if $(e_i - e_j) > \varepsilon$, and values between 0 and 1 if $-\varepsilon < (e_i - e_j) \le \varepsilon$.

When $\varepsilon = 0$, the polynomial-smoothed Gehan estimating function is exactly the Gehan estimating function, $\boldsymbol{\Psi}_{G,\varepsilon}(\boldsymbol{\beta}) = \boldsymbol{\Psi}_{G}(\boldsymbol{\beta})$.Written in this way, the difference between $\boldsymbol{\Psi}_{H}(\boldsymbol{\beta})$ and the polynomial-smoothed estimating function $\boldsymbol{\Psi}_{G,\varepsilon}(\boldsymbol{\beta})$ is how the weight is assigned to the difference $(e_i - e_j)$. This fact leads to a useful heuristic for standard error estimation for the polynomial-smoothed estimator $\hat{\boldsymbol{\beta}}_{G,\varepsilon}$ and is outlined in the Appendix.

Compared with $f_H(\boldsymbol{\beta})$ or $f_{G,\varepsilon}(\boldsymbol{\beta})$, $f_B(\boldsymbol{\beta})$ is self-contained in the sense that there is no independent tuning parameter a or ε, respectively. The price one pays for the automatic data-dependent bandwidth r_{ij} is mostly computational: a sandwich matrix must be computed at every iteration to update $\boldsymbol{\Omega}$. However, simultaneous coefficient and covariance estimation is a principal motivation behind the method of Brown and Wang (2005, 2007) and one expects a proportional increase in the computational burden.

3.4 Inference Procedures

Tsiatis (1990) showed that, under suitable regularity conditions, the Gehan estimator is asymptotically normal, i.e. $n^{1/2}(\hat{\boldsymbol{\beta}}_G - \boldsymbol{\beta}_0)$ converges in distribution to a mean-zero normal random vector with covariance

$$\boldsymbol{\Omega}_G = \mathbf{A}_G^{-1}\mathbf{B}_G\mathbf{A}_G^{-1},$$

where

$$\mathbf{A}_G = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\left[\delta_i \vartheta(e_i,\beta)\frac{h'_{\zeta}(e_i)}{h_{\zeta}(e_i)}\left\{\mathbf{x}_i - \frac{\sum_{j=1}^{n} I(e_j \geq e_i)\mathbf{x}_j}{\sum_{j=1}^{n} I(e_j \geq e_i)}\right\}^{\otimes 2}\right],$$

$$\mathbf{B}_G = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\left[\delta_i \{\vartheta(e_i,\beta)\}^2\left\{\mathbf{x}_i - \frac{\sum_{j=1}^{n} I(e_j \geq e_i)\mathbf{x}_j}{\sum_{j=1}^{n} I(e_j \geq e_i)}\right\}^{\otimes 2}\right],$$

$\vartheta(t, \boldsymbol{\beta})$ is the Gehan weight in (3), $h_{\zeta}(t)$ is the hazard function of the errors ζ_i in (1), $h'_{\zeta}(t) = (d/dt)h_{\zeta}(t)$, and $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}^{\top}$. For the inference procedures here, it is assumed that $\mathbf{A}_G$ is full-rank. Because the matrix $\mathbf{A}_G$ involves the hazard function of the errors, it cannot be directly evaluated without non-parametric smoothing or numerical differentiation, but these techniques can incur substantial instability in finite samples. This was the impetus for the resampling technique by Jin et al (2003). The idea is to generate n independent, exponentially-distributed random variables $Z_i \sim \text{Exp}(1)$, $i = 1, \ldots, n$, and define

$$\hat{\beta}_G^* = \arg\min_{\beta} f_G^*(\beta)$$

where $f_G(\boldsymbol{\beta})^*$ is the perturbed loss function,

$$f_G^*(\beta) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i (e_j - e_i) I(e_i \leq e_j) Z_i.$$

The perturbation is repeated a large number of times, say M, and an estimate of var($\hat{\boldsymbol{\beta}}_G$) is the sample covariance of the M resampled vectors $\hat{\beta}^*_G$. The key to the success of perturbing the minimand is that $E(Z_i) = \text{var}(Z_i) = 1$ and the mechanism that generates Z_i is completely independent of the data-generating mechanism for $\{(\mathbf{x}_i, U_i, \delta_i), i = 1, \ldots, n\}$. Note, that perturbing the minimand is a general technique and it applies to any of the smooth loss functions, $f_{G,\varepsilon}(\boldsymbol{\beta})$, $f_B(\boldsymbol{\beta})$ or $f_H(\boldsymbol{\beta})$.

Of course, resampling is computationally demanding and a direct solution is preferable. Similar to the Gehan estimator, the asymptotic covariance of the smooth estimators $\hat{\boldsymbol{\beta}}$. takes the usual sandwich form,

$$\boldsymbol{\Omega}_\bullet = \mathbf{A}_\bullet^{-1}\mathbf{B}_\bullet\mathbf{A}_\bullet^{-1},$$

where **B.** is the asymptotic covariance of $n^{1/2}\boldsymbol{\Psi}_\bullet(\boldsymbol{\beta}_0)$ and **A.** is the asymptotic slope matrix of $\lim_{n\to\infty} \boldsymbol{\Psi}_\bullet(\boldsymbol{\beta}_0)$. However, unlike the original Gehan estimator whose asymptotic slope matrix could not be directly evaluated, the derivative of the smoothened estimating function may be evaluated analytically and $\mathbf{A}_\bullet = \lim_{n\to\infty} -(\partial/\partial\boldsymbol{\beta})\boldsymbol{\Psi}_\bullet(\boldsymbol{\beta}_0)$. The sample estimator for **A.** is

$$\hat{\mathbf{A}}_\bullet = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\,(\mathbf{x}_i - \mathbf{x}_j)^{\otimes 2} w_{ij},$$

where w_{ij} is a weight on the difference $(e_i - e_j)$. If $\phi(z)$ is the normal probability density function evaluated at z, then for Brown and Wang, $w_{ij} = \phi\{(e_i - e_j)/r_{ij})\}/r_{ij}$; Heller, $w_{ij} = \phi\{(e_i - e_j)/a)\}/a$; polynomial-smoothed, w_{ij} is obtained by straightforward differentiation of $K_\varepsilon(z)$. We also need an estimator for **B.**. Due to the asymptotic equivalence of $n^{1/2}\boldsymbol{\Psi}_\bullet(\boldsymbol{\beta}_0)$ and $n^{1/2}\boldsymbol{\Psi}_G(\boldsymbol{\beta}_0)$, we may use the sample estimator of the asymptotic covariance of $n^{1/2}\boldsymbol{\Psi}_G(\boldsymbol{\beta}_0)$ for the smoothened estimators $\hat{\boldsymbol{\beta}}$. (cf. Brown and Wang, 2005, 2007; Johnson and Strawderman, 2009); namely,

$$\hat{\mathbf{B}}_\bullet = \frac{1}{n}\sum_{i=1}^{n}\left[\delta_i\{\vartheta(e_i,\beta)\}^2\left\{\mathbf{x}_i - \frac{\sum_{j=1}^{n} I(e_j \geq e_i)\mathbf{x}_j}{\sum_{j=1}^{n} I(e_j \geq e_i)}\right\}^{\otimes 2}\right].$$

Both Heller (2007) and Brown and Wang (2005, 2007) provide other estimators for **B.** based a theory of U-statistics. The estimator by Johnson and Strawderman (2009) reduces to $\hat{\mathbf{B}}$. when the survival times are independent as we have here. Consequently, our estimator of the asymptotic covariance $\boldsymbol{\Omega}$. is

$$\hat{\boldsymbol{\Omega}}_\bullet = \hat{\mathbf{A}}_\bullet^{-1}\hat{\mathbf{B}}_\bullet\hat{\mathbf{A}}_\bullet^{-1}.$$

4 Regularized Rank-based Estimation in AFT Models

Regularized regression has drawn substantial interest among researchers in statistics and biostatistics in recent years because it achieves simultaneous model selection and parameter estimation (Tibshirani, 1996; Zou and Hastie, 2005; Tibshirani et al, 2005; Yuan and Lin, 2006). A second reason regularized regression has gained popularity is due to high-dimensionality of today's data sets. In particular, when the dimension of the predictors p

exceeds the sample size n, the estimation problem is said to be *ill-posed*. In 1902, Hadamard defined a mathematical problem to be *well-posed* if a solution exists, the solution is unique, and depends continuously on the data in some reasonable topology (Hadamard, 1902). In general, every least-squares-based or likelihood-based estimation problem will be *ill-posed* with no unique solution when $n < p$. In order to construct a well-posed parameter estimation problem for high-dimensional data, one needs to incorporate prior knowledge to overcome the ambiguity of the global minimizers. In the Bayesian framework, the prior knowledge is reflected by a-priori information on the estimators, leading naturally to additive regularization terms (Vogel, 2002; Kaipio and Somersalo, 2005).

4.1 Exact Solutions

Few authors have offered algorithms for regularized rank-based coefficient estimation for censored outcomes. To clarify the challenges, the familiar Lagrangian form of the ℓ_1-regularized Gehan estimator is:

$$\arg\min_{\beta} \left\{ f_G(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \quad (15)$$

Building on the earlier LP problem for the unregularized optimization in (7) and noting that $|\beta_j| = \beta_j^+ + \beta_j^-$, the optimization problem in (15) can be rewritten as the following LP problem:

$$\min_{u,\beta} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i u_{ij}$$

subject to: $u_{ij} = (e_j - e_i)$,

$u_{ij} \le 0$,

$$\sum_{k} (\beta_k^+ + \beta_k^-) \le \tau,$$

$$\beta_k^+ \ge 0, \beta_k^- \ge 0,$$

for $i, j = 1, \ldots, n$,

for $k = 1, \ldots, p$,

where τ is a regularization parameter. Johnson (2008) was the first to attempt to solve a class of general problems related to (15) through direct search methods. In that paper, he noted that the optimization in (15) was of the form, ℓ_1 loss plus ℓ_1 penalty, and could be written as another LP problem but provided no algorithm to produce an exact solution. Then, Johnson (2009a) developed a practical solution for the ℓ_1-regularized Gehan estimator that made explicit use of linear programming. In independent work, Xu et al (2010) offered the same algorithm as in Johnson (2009a) and extended it to correlated survival times.

Both Johnson (2009a) and Xu et al (2010) extended an earlier algorithm by Jin et al (2003) to accommodate the ℓ_1 penalty. Their procedures use the ubiquitous `quantreg` package in `R`, developed by Roger Koenker and coauthors (Koenker and Bassett Jr, 1978; Koenker and D'Orey, 1987; Koenker and Ng, 2005), and were primarily developed for problems where $n > p$. As mentioned earlier in Section 1, *it is possible* to use interior point methods to solve large-scale LP problems with $p > n$. But our experience is that procedures built on `quantreg` in `R` will not suffice and another solution is needed. Two alternatives are (a) to

use another LP solver, or (b) to write new code to solve the specific LP problem. A caveat to (a) is that many off-the-shelf LP solvers used to solve large-scale problems are unfamiliar to most statisticians and would be rarely adopted in practice. Cai et al (2009) chose the second route (b) and developed a path-finding algorithm that computes the entire ℓ_1-regularized coefficient path. Unfortunately, this path-finding algorithm cannot be easily extended to general penalty functions.

4.2 Newton-type Solutions

For large-scale problems, it becomes almost imperative to utilize efficient algorithmic methods to solve optimization problem (15) and gradient-based optimization algorithms are efficient. However, for the same reason that gradient-based algorithms cannot be applied to minimize $f_G(\boldsymbol{\beta})$ alone, gradient-based optimization cannot be applied directly to minimize (15) because of the non-differentiability of ℓ_1-norm at zero. A natural way to proceed here is to approximate the absolute value function with a piecewise quadratic function. For example, the Huber function (Huber, 1964) is often used to smoothen the ℓ_1-norm and is given by

$$h_\varepsilon(z)=\begin{cases}\frac{1}{2\varepsilon}z^2, & \text{if } |z|<\varepsilon,\\ |z|-\frac{\varepsilon}{2}, & \text{if } |z|\geq\varepsilon,\end{cases} \quad (16)$$

as illustrated in Figure 2. Let $f_\bullet(\boldsymbol{\beta})$ be short-hand for any smoothed rank-based loss function: $f_B(\boldsymbol{\beta})$, $f_H(\boldsymbol{\beta})$, or $f_{G,\varepsilon}(\boldsymbol{\beta})$. So, by substituting $f_\bullet(\boldsymbol{\beta})$ for $f_G(\boldsymbol{\beta})$ and the Huber ε-approximation $h_\varepsilon(\beta_j)$ for $|\beta_j|$ in (15), we arrive at a new optimization problem,

$$\arg\min_{\beta}\left\{f_\bullet(\beta)+\lambda\sum_{j=1}^{p}h_\varepsilon(\beta_j)\right\}. \quad (17)$$

Because the convex objective function in (17) is twice-differentiable in $\boldsymbol{\beta}$, we may use gradient-based algorithms to minimize it. Newton-type algorithms possess quadratic convergence rates and are, thus, very efficient.

The current literature provides various types of regularization methods and penalty functions, say $\mathcal{S}(\lambda, \boldsymbol{\beta})$, where $\mathcal{S}$ is a convex function and $\lambda \geq 0$ is a regularization parameter. Many methods depend on the ℓ_1-norm and Huber's ε-approximation provides a recipe for smoothing. Along the lines discussed above, we will replace a non-differentiable function $\mathcal{S}(\lambda, \boldsymbol{\beta})$ with a smooth approximation $\mathcal{S}_\varepsilon(\lambda, \boldsymbol{\beta})$. Then, we characterize the family of regularized smooth Gehan estimators as

$$\arg\min_{\beta}\{f_\bullet(\beta)+\mathcal{S}_\varepsilon(\lambda,\beta)\}. \quad (18)$$

While it is impossible to provide a comprehensive list of all available penalty functions, we enumerate below a list of seven common regularization methods, i.e., $\mathcal{S}(\lambda, \boldsymbol{\beta})$, which have been proposed for penalized least squares and penalized likelihood problems. In Table 1, we exemplify the smoothed penalty function $\mathcal{S}_\varepsilon(\lambda, \boldsymbol{\beta})$ alongside the original function $\mathcal{S}(\lambda, \boldsymbol{\beta})$ for all seven penalty functions.

1. The ℓ_2^2 regularization (Hoerl and Kennard, 1970) with $\mathcal{S}(\lambda,\beta)=\lambda\|\beta\|_2^2$ penalizes $\boldsymbol{\beta}$ quadratically, leading to greater penalties for large values of $\boldsymbol{\beta}$, and straightforward numerical algorithms can be used to handle the optimization problem (18).

2. The ℓ_1 regularization, a. k. a. *lasso* (Tibshirani, 1996), with $\mathcal{S}(\lambda, \boldsymbol{\beta}) = \lambda \|\boldsymbol{\beta}\|_1$ penalizes all parameters $\boldsymbol{\beta}$ linearly and leads to sparse representations (Boyd and Vandenberghe, 2004). ℓ_1 regularization methods are broadly used such as in signal processing, statistics and geophysical applications. Since ℓ_1 regularization is not differentiable at 0 (i.e., for any $\beta_j = 0$), computationally this approach needs special attention. Applicable methods to solve the optimization problem (18) include linear programming, interior point, and iterative re-weighted least squares methods.

3. The *elastic net* (Zou and Hastie, 2005) $\mathcal{S}(\lambda,\beta)=\lambda_1\|\beta\|_1+\lambda_2\|\beta\|_2^2$, with $\lambda = (\lambda_1, \lambda_2)^\top$ combines the ℓ_1 and ℓ_2^2 regularization. Dependent on the sizes of λ_1 and λ_2 this regularizer penalizes large values quadratically and small values linearly.

4. A regularizer of "opposite" behavior to the elastic net is the Berhu regularizer (Owen, 2006), $\mathcal{S}(\lambda, \boldsymbol{\beta}) = \sum_j \mathcal{S}_j(\lambda, \beta_j)$, and

$$\mathcal{S}_j(\lambda,\beta_j)=\lambda \begin{cases} |\beta_j|, & \text{if } |\beta_j| \geq \varepsilon, \\ \frac{\beta_j^2+\varepsilon^2}{2\varepsilon}, & \text{if } |\beta_j|<\varepsilon, \end{cases}$$

penalizing small values of $\boldsymbol{\beta}$ quadratically and large values linearly.

5. The discrete *total variation* type regularizer $\mathcal{S}(\lambda,\beta)=\lambda\sum_{j=1}^{p-1}|\beta_{j+1}-\beta_j|$ induces smoothness in the parameters $\boldsymbol{\beta}$. Note, this regularizer presumes a neighboring/ordering structure of the β_j's, $j = 1, \ldots, p$.

6. To induce sparsity and smoothness at the same time the two dimensional *fused lasso* regularization $\mathcal{S}(\lambda,\beta)=\lambda_1\|\beta\|_1+\lambda_2\sum_{j=1}^{p-1}|\beta_{j+1}-\beta_j|$ has been introduced by Tibshirani et al (2005).

7. Let the G_k's be the mutually exclusive subsets of $\{1, \ldots, p\}$. The *group lasso* can now be seen as the grouped ℓ_2 regularization,

$$\mathcal{S}(\lambda,\beta)=\lambda\sum_{k=1}^{K}\sqrt{\aleph(G_k)}\left\|\beta_{G_k}\right\|_2,$$

where $\aleph(G_k)$ is the cardinality of G_k (Yuan and Lin, 2006; Meier et al, 2008) and reduces to ℓ_1 regularization when $\aleph(G_k) = 1$. Note, $\boldsymbol{\beta}_{G_k}$ is the vector of elements β_k for which $k \in G_k$. Group lasso requires prior knowledge on the grouping of the parameters $\boldsymbol{\beta}_k$. Note, $\|\cdot\|_2$ is not differentiable in a singular point, i.e., 0. Typically, this is neglected in numerical investigations using gradient-based methods.

5 Algorithm

Every regularization procedure consists of two parts: (a) estimating the regression parameters $\boldsymbol{\beta}$ for fixed regularization parameter λ, and (b) tuning λ for optimal performance. For coefficient estimation in part (a), the computational advantage of our approximation lies in the smoothness and convexity of $f_{G,\varepsilon}(\boldsymbol{\beta})$ and $\mathcal{S}_\varepsilon(\lambda, \boldsymbol{\beta})$. Due to these properties, we may use gradient-based optimization algorithms for which the optimization theory provides numerous iterative methods. One of the foremost gradient-based optimization algorithm is Newton's method, which converges locally at a quadratic rate and uses the gradient and Hessian to form the search direction and step length at each iteration. In our experiments, we observe that the numerical calculations of the Hessian of the

smoothed Gehan estimator and solving the inner system are at unreasonable costs (in particular for large scale problems) and we try to avoid utilizing Hessian information, i.e., curvature information; as a result, we prefer *quasi-Newton* methods in our investigations. The fundamental concept behind quasi-Newton methods is to provide curvature information of a loss function $f_{G,\varepsilon}(\boldsymbol{\beta})$ and the regularizer $\mathscr{S}_\varepsilon(\lambda, \boldsymbol{\beta})$ in order to calculate an efficient search direction at each iteration without calculating the Hessian matrix explicitly and solving the inner system.

Algorithm 1

L-BFGS method

Require:

f_ε {smooth model function}

$\mathscr{S}_\varepsilon$ {smooth regularizer function}

$\boldsymbol{\beta}_0$ {initial guess}

d {data}

λ_j {regularization parameter}

1: **while** $\nabla \mathscr{J}(\boldsymbol{\beta}) \neq 0$ **do**

2: calculate $\mathscr{J}(\boldsymbol{\beta}) = f_\varepsilon(\boldsymbol{\beta}) + \mathscr{S}_\varepsilon(\lambda_j, \boldsymbol{\beta})$ and $\nabla \mathscr{J}(\boldsymbol{\beta})$

3: estimate Hessian inverse approximation $\mathscr{H}$ by last K update steps

4: $\mathbf{s} = -\mathscr{H}^\top \nabla \mathscr{J}(\boldsymbol{\beta})$ {quasi Newton search direction}

5: calculate α via Armijo line search

6: $\boldsymbol{\beta} = \boldsymbol{\beta} + \alpha \mathbf{s}$ {update step}

7: **end while**

8: $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$

Ensure:

$\hat{\boldsymbol{\beta}}$ {optimal parameter}

The limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method is one such quasi-Newton method, which is designed to target large scale optimization problems. The L-BFGS method avoids forming the Hessian and solving the inner linear system and only incorporates curvature information of the last few iterates, as outlined in pseudo-code in Algorithm 1. Throughout this work, we use the L-BFGS with an Armijo line search algorithm. For all our experiments we use a relative tolerance of 10^{-6} for the stopping criteria of the optimization algorithm (Gill et al, 1981).

The choice of the regularization parameter λ in part (b) is also crucial. To get a good estimate of the regularization parameter λ one may utilize information-based rules, cross validation, generalized cross validation, discrepancy principle and statistical learning techniques (Chung et al, 2011). In the interest of space, we refer readers to Hastie et al (2009) for a detailed description of cross-validation but provide a summary in our pseudo-code in Algorithm 2. The whole procedure, including coefficient estimation in part (a) and parameter tuning in part (b), is described in Algorithm 3. Our general framework is implemented in Matlab and is available upon request.

6 Worked Examples

As discussed previously, we divide the rank-based estimation problems for AFT models into two categories. First, the problems deal with low-dimensional predictors **x** relative to the

number of the observation n, where n could be small or large. These problems are typically well-posed. The other case is when we have $p > n$ or even $p \gg n$. These problems are typically ill-posed and a regularization $\mathscr{S}_\varepsilon$, often driven by prior knowledge on $\mathbf{x}$, is used.

In this section we present four data examples where the AFT model relates lifetime or survival to risk factors. The first three examples have low-dimensional predictors and intended to compare accuracy, computational complexity, and inference between the original Gehan estimator $\hat{\boldsymbol{\beta}}_G$ and the smoothened estimators, $\hat{\boldsymbol{\beta}}_\bullet$: $\hat{\boldsymbol{\beta}}_B$, $\hat{\boldsymbol{\beta}}_H$, and $\hat{\boldsymbol{\beta}}_{G,\varepsilon}$. For smoothing in $\Psi_H(\boldsymbol{\beta})$, we use Heller's suggested estimate and rate, $\hat{a} = \hat{\sigma} n^{-0.26}$, where $\hat{\sigma}^2$ is the sample variance of the uncensored residuals based on an initial Gehan fit. We estimate standard errors for $\hat{\boldsymbol{\beta}}_H$ and $\hat{\boldsymbol{\beta}}_B$ directly through a sandwich estimator $\hat{\boldsymbol{\Omega}}_\bullet$ for the asymptotic covariance $\boldsymbol{\Omega}_\bullet$. For the polynomial-smoothed estimator, $\hat{\boldsymbol{\beta}}_{G,\varepsilon}$, our experience is that perturbing the loss function works better in this case and, hence, results from resampling are presented below. Finally, in the fourth example, we present a data analysis of high-dimensional microarray data using the group lasso regularization. In this last example, no standard error estimates are presented because, at the time of this writing, there is no theoretically-justified inference procedure for the class of regularized estimators considered here.

Algorithm 2

Cross Validation

Require:

f_ε	{smooth model function}
$\mathscr{S}_\varepsilon$	{smooth regularizer function}
$\boldsymbol{\beta}_0$	{initial guess}
d	{data}
λ_j	{regularization parameter}

1: choose n cross validation sample sets $\{d_i\}_{i=1,\ldots,n}$

2: **for** $i = 1$ to n **do**

3: extract sub-sample set

4: $\hat{\beta}_j^n = \arg\min_\beta f_\varepsilon(\beta) + \mathscr{S}_\varepsilon(\lambda j, \beta)$ using d_i

{optimization method see Algorithm 1}

5: calculate $f_i = f_\varepsilon(\beta_i^n)$

6: **end for**

7: calculate V_j = mean (f_i's)

Ensure:

V_j

Algorithm 3

Driver for Smooth Statistical Models

Require:

f_ε	{smooth model function}
$\mathscr{S}_\varepsilon$	{smooth regularizer function}

$\boldsymbol{\beta}_0$ {initial guess}

d {data}

$\{\lambda_1, \ldots, \lambda_m\}$ {set of regularization parameters}

1: **for** $j = 1$ to m **do**

2: $\hat{\boldsymbol{\beta}}_j = \arg\min_{\boldsymbol{\beta}} f_e(\boldsymbol{\beta}) + \mathcal{S}_e(\lambda_j, \boldsymbol{\beta})$ {optimization method, see Algorithm 1}

3: calculate f_j {Cross-Validation method, see Algorithm 2}

4: **end for**

5: $\hat{j} = \arg\min_{j=1, \ldots, m} V_j$ {choose minimal cross validation set}

6: set $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_j$

Ensure:

$\hat{\boldsymbol{\beta}}$ {optimal parameter}

6.1 Multiple Myeloma Data

First, we exemplify the methods using multiple myeloma data set, given in the online SAS/STAT User's Guide. This is the primary example in the PHREG procedure and was also used for illustration in Jin et al (2003). The data consist of survival outcomes and two independent variables, hemoglobin (HGB) and the natural logarithm of blood urea nitrogen (BUN), for a total of $n = 65$ patients. The covariates are standardized to have mean zero and unit variance. For HGB and log(BUN), Jin et al (2003) report estimated coefficients $\hat{\boldsymbol{\beta}}_G$ as −0.532 and 0.292 with estimated standard errors 0.146 and 0.169, respectively. Using our polynomial-smoothed estimator $\hat{\boldsymbol{\beta}}_{G,\varepsilon}$, the coefficient estimates are −0.532 and 0.292 with estimated standard errors 0.149 and 0.164. The Brown and Wang coefficient estimates are −0.510 and 0.304 with standard error estimates 0.212 and 0.208. Using a smoothing parameter $\hat{a} = 0.987$, Heller's estimates are −0.510 and 0.302 and standard error estimates 0.190 and 0.196, respectively. In short, the polynomial-smooth estimate $\hat{\boldsymbol{\beta}}_{G,\varepsilon}$ is similar to the original Gehan estimate $\hat{\boldsymbol{\beta}}_G$ in both point estimate and standard error estimate. The point and standard error estimates of $\hat{\boldsymbol{\beta}}_B$ and $\hat{\boldsymbol{\beta}}_H$ are similar to one another, but the point estimates are 3–4% different in absolute magnitude than the Gehan coefficient estimates and the standard error estimates 20–30% larger than resampling. When a different covariance estimator is used, Brown and Wang (2007) find standard error estimates of the same magnitude as in Jin et al (2003).

6.2 Mayo PBC Data

The Mayo primary biliary cirrhosis (PBC) data set (Fleming and Harrington, 1991) contains information about the survival time and prognostic variables for 418 patients who were eligible to participate in a randomized study of the drug D penicillamin. Of 418 patients who met standard eligibility criteria, a total of 312 patients participated in the randomized portion of the study. Using the smaller randomized cohort, the study investigators used stepwise deletion to build a Cox proportional hazards model for the natural history of PBC (Dickson et al, 1989). Of the original ten predictors, stepwise deletion selected five significant variables: age, albumin, bilirubin, edema, and prothrombin time (protime). We take the natural logarithmic transformation of albumin, bilirubin, and prothrombin time to conform to the analysis in Fleming and Harrington (1991). These five variables constitute the natural history model for PBC (Dickson et al, 1989).

We present in Table 2 the coefficient estimates $\hat{\boldsymbol{\beta}}_B$, $\hat{\boldsymbol{\beta}}_H$, $\hat{\boldsymbol{\beta}}_{G,\varepsilon}$ (here we set $\varepsilon = 10^{-4}$) based on (13), and $\hat{\boldsymbol{\beta}}_G$ based on (6) using linear programming. We note that the coefficient and standard error estimates between $\hat{\boldsymbol{\beta}}_G$ and $\hat{\boldsymbol{\beta}}_{G,\varepsilon}$ are nearly identical. The proposed BFGS algorithm runs in less than one-half of one second and the linear programming method of Jin et al (2003) is still reasonable at 5.3 seconds (on our MacBook Pro running R 2.9.1) given the moderate sample size. Heller's smoothing parameter is $\hat{a} = 1.607$ and the resulting coefficient estimates tend to be stronger, i.e. farther from the null; the corresponding standard error estimates are larger. The coefficient estimates from Brown and Wang generally differ from the other estimators and the standard error estimates lie between those computed for $\hat{\boldsymbol{\beta}}_H$ and $\hat{\boldsymbol{\beta}}_{G,\varepsilon}$.

6.3 Nursing Home Data

From 1980–1982, the National Center for Health Services Research conducted a study to determine the effect of financial incentives on variation of patient care in nursing homes. In particular, 18 out of 36 nursing homes from San Diego, California, received higher per diem payments for accepting and admitting Medicaid patients and additional bonuses when the patient's prognosis improved. The study collected data from an additional 18 control nursing homes where no financial incentives were used. A complete description is given in Morris et al (1994). The total sample size from all 36 nursing homes is $n = 1601$. Our data set consists of seven co-variables: treatment (trt), age, sex, marital status, and three health status indicators (h1–h3), ranging from the best health to the worst health. For the polynomial-smoothed estimator, we used $\varepsilon = 10^{-4}$. Our results are presented in Table 3.

In Table 3, coefficient estimates for all four estimators are displayed. For the nursing home data set, we computed Heller's smoothing parameter as $\hat{a} = 0.7056$. In this data set, we found very minor differences among the coefficient and standard error estimates. Among the three smoothened estimators, the polynomial-smoothed estimator took the longest took converge at 35 iterations. Both $\hat{\boldsymbol{\beta}}_H$ and $\hat{\boldsymbol{\beta}}_B$ converged in less than 10 iterations.

The sample size of the nursing home data is sufficiently "large" where the smoothing algorithm makes a significant impact. Using the algorithm of Jin et al (2003) along with Barrodale-Roberts simplex optimization (Koenker and D'Orey, 1987) via `quantreg` in `R`, the computation fails. However, the improved Frisch-Newton (Koenker and Ng, 2005) algorithm performs better and finishes in just under two minutes (i.e., 1.75 minutes on our MacBook Pro running R 2.9.1). For the nursing home data set, our quasi-Newton algorithm runs in five seconds. To highlight the differences in CPU times, consider computing standard error estimates using the resampling scheme by Jin et al (2003) with $M = 1000$ resamples. In this case, their resampling procedure would take more than one day on our desktop computer. In order to compute the standard error estimates for $\hat{\boldsymbol{\beta}}_G$ in Table 3, we submitted our job to the Emory University Rollins School of Public Health high performance computing cluster. On our cluster, the resampling procedure of Jin et al (2003) applied to the nursing home data took 4 hours for $M = 500$ resamples. All of the other standard error estimates can be computed on an ordinary desktop computer in a matter of seconds.

6.4 CAMDA Data

We investigate a large scale data analysis using data from the Critical Assessment of Microarray Data Analysis (CAMDA) 2003 program, the details of which can be found on their website (CAMDA, 2003). For this analysis, we use gene expression data that were obtained through microarray experiments and the outcome of interest is a survival endpoint measured as time-to-death (in months) due to lung andenocarcinoma. The sample includes $n = 200$ subjects and gene expression data for 1036 gene probe sets; we refer to the probe sets

as “gene biomarkers.” Our goal is to identify the gene biomarkers that are associated with survival of patients with lung andenocarcinoma using the AFT model with the group lasso penalty (see Table 1).

To identify the group structure, we first perform k-mean clustering to divide the gene biomarkers into 50 groups and rearrange the biomarkers so that the biomarkers in the same group have consecutive indices; of note, the group size ranges from 1 to 66. When fitting the AFT model with $f_\bullet(\beta) \equiv f_{G,e}(\beta)$ and the group lasso penalty, we use a five-fold cross validation technique as presented in Algorithm 2 to select the optimal regularization parameter $\lambda_j = 0.01773$. The resulting sparse model includes 78 biomarkers with nonzero regression coefficient estimates from 13 different groups and the regression coefficient estimates are presented in Figure 3.

7 Conclusions

In this paper, we present a general framework to efficiently compute rank-based coefficient estimates for semiparametric AFT models in small- or large-scale problems. Exact rank-based estimates are computed by optimization a linear programming (LP) problem. Although computing exact solutions is a laudable goal and may be required in some unusual settings, it is rarely needed in practical work. For those instances when an accurate Gehan estimate is required, our polynomial-based smoothing yields coefficient estimates nearly identical to Gehan estimates but only take a fraction of the computational resources compared to solving the LP problem. Moreover, the same ideas to smoothen the non-smooth loss function can be applied to the smoothing non-differentiable regularizations as well.

In addition to the polynomial-smooth Gehan estimator, Brown and Wang (2005, 2007) and Heller (2007) have each offered alternative rank-based coefficient estimators that are consistent, asymptotically normal, and whose asymptotic covariance matrix is directly estimable. Hence, one can estimate standard errors easily for large data set without resorting to computationally-intensive resampling. We reviewed standard error estimation for non-regularized Gehan estimators but not for regularized Gehan estimators. At the time of this writing, this is still an open question. If future researchers find that perturbing the minimand is a theoretically-sound resampling technique for regularized estimators, then our description in Section 3.4 will be germane for all estimators reviewed in this paper.

The fact that many constrained optimization problems can be closely approximated by an unconstrained optimization problem with a smooth objective function is an old idea. However, the application to regularized rank-based estimation for censored data is new and relevant to emerging data sets. We can apply our algorithm to penalty functions that have been proposed in the least squares framework but not yet extended to rank-based estimators, e.g. large-scale rank-based coefficient estimation with group lasso penalty. The general form of the computational framework makes it applicable for a wide range of optimization problems beyond the survival analysis applications discussed here.

Finally, as with most scientific problems, there is more than one approach, more than one technique to achieve the scientific objective. Rank-based estimation in the AFT model is but one technique and competing methods include those based on least squares, imputation, or inverse weighting. Over the past decade, several authors have advanced these competing methods to the regularized estimation setting yet were not discussed here (cf. Huang et al, 2006; Johnson, 2008; Johnson et al, 2008; Johnson, 2009b; Johnson et al, 2011). This omission is unabashedly self-serving and partly reflects our bias for rank-based estimators in the AFT model. Compared with least-squares estimators, rank-based estimators lose only a small amount of efficiency for (log)-normal errors but are more efficient for skewed and

heavy-tailed error distributions. Inverse weighting is a powerful and convenient technique but whose finite sample behavior can be tied closely to the magnitude of the weights and hence the tail of the censoring distribution. This tutorial aims to be more or less comprehensive for unregularized and regularized rank-based estimation in the AFT model but falls well short as a comprehensive review of small- or large-scale coefficient estimation in the AFT model, in general. Nevertheless, we hope researchers interested in robust coefficient estimation in the AFT model find this tutorial helpful.

Acknowledgments

This work was supported in part by US NIH PHS Grant UL1 RR025008 from the Clinical and Translational Science Award program.

Appendix

Operating Characteristics of Polynomial-smoothed Gehan Estimator

In this section, we outline the large sample properties of the estimator $\hat{\boldsymbol{\beta}}_{G,\varepsilon}$. Let the parameter $\boldsymbol{\beta}$ belong to a parameter space $\mathbb{B}$ a compact subset of $\Re^p$ and let $f_0(\boldsymbol{\beta})$ be a convex function for $\boldsymbol{\beta} \in \mathbb{B}$ The proof of Theorem 1 relies on the following two facts regarding the loss functions $f_G(\boldsymbol{\beta})$ and $f_{G,\varepsilon}(\boldsymbol{\beta})$.

Lemma 1 *Under Conditions A1–A3 in* Johnson and Strawderman (2009, *p.586)*,

$$\sup_{\beta\in\mathbb{B}}|f_G(\beta) - f_0(\beta)| \to 0 \quad almost \quad surely.$$

Lemma 2 *Under Conditions A1–A3 in* Johnson and Strawderman (2009, *p.586)*,

$$\sup_{\beta\in\mathbb{B}}|f_{G,\varepsilon}(\beta) - f_0(\beta)| \to 0 \quad almost \quad surely.$$

Lemma 1 is also Lemma 1 in Johnson and Strawderman (2009) under exactly the same conditions and stated without proof.

Outline proof of Lemma 2. By the triangle inequality, we have

$$|f_{G,\varepsilon}(\beta) - f_0(\beta)| \le |f_{G,\varepsilon}(\beta) - f_G(\beta)| + |f_G(\beta) - f_0(\beta)|. \quad (19)$$

By Lemma 1, the second term in (19) can be made arbitrarily small, uniformly for all $\boldsymbol{\beta} \in \mathbb{B}$ except on a set of probability measure zero. The first term in (19) is

$$|f_{G,\varepsilon}(\beta) - f_G(\beta)| \le \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|(c_\varepsilon(-u_{ij}) - c_0(-u_{ij}))| \le \max|(c_\varepsilon(-u_{ij}) - c_\varepsilon(-u_{ij}))| = \frac{3}{16}\varepsilon.$$

Hence, the absolute difference between the Gehan loss and its smooth approximation can be made arbitrarily small, for every $\boldsymbol{\beta} \in \mathbb{B}$ The conclusion then follows.

Proof of Theorem 1. Under Conditions A1–A3 of Johnson and Strawderman, $f_G(\boldsymbol{\beta})$ and $f_{G,\varepsilon}(\boldsymbol{\beta})$ converge uniformly to the convex function $f_0(\boldsymbol{\beta})$ by Lemmas 1 and 2, respectively.

By Condition A4, $f_0(\boldsymbol{\beta})$ is strictly convex at its unique minimizer, $\boldsymbol{\beta}_0$. Thus, the minimizers of the random convex functions $f_{G,\varepsilon}(\boldsymbol{\beta})$ and $f_G(\boldsymbol{\beta})$ converge almost surely to $\boldsymbol{\beta}_0$.

Asymptotic Distribution The polynomial-smoothed Gehan estimator bears a close similarity to Heller's (2007) estimator and one expects the asymptotic distribution theory follows similarly. A straightforward calculation confirms that $K_\varepsilon(z)$ in $\boldsymbol{\Psi}_{G,\varepsilon}(\boldsymbol{\beta})$ in (14) is a survivor function and $k_\varepsilon(z) = (d/dz)K_\varepsilon(z)$ is symmetric about zero with finite second moment (that is, Heller's, 2007, Condition C3, p. 553). Define the asymptotic slope matrix $\mathbf{A}_\varepsilon(\boldsymbol{\beta})$ and asymptotic covariance $\mathbf{B}_\varepsilon(\boldsymbol{\beta})$,

$$\mathbf{A}_\varepsilon(\beta) = \lim_{n\to\infty} E\left[(\partial/\partial\beta)\boldsymbol{\Psi}_{G,\varepsilon}(\beta)\right]|_{\beta=\beta_0},$$

$$\mathbf{B}_\varepsilon(\beta) = \lim_{n\to\infty} \operatorname{var}\left\{n^{1/2}\boldsymbol{\Psi}_{G,\varepsilon}(\beta_0)\right\}.$$

Then, assuming the covariate matrix has finite second moment and the non-singularity of $\mathbf{A}_\varepsilon(\boldsymbol{\beta})$ in a neighborhood of the true value $\boldsymbol{\beta}_0$, one can show $\sqrt{n}(\hat{\beta}_{G,\varepsilon} - \beta_0)$ converges in distribution to a mean-zero normal random vector with asymptotic covariance

$$\{\mathbf{A}_\varepsilon(\beta_0)\}^{-1}\mathbf{B}_\varepsilon(\beta_0)\{\mathbf{A}_\varepsilon(\beta_0)\}^{-1},$$

(see Heller, 2007, Appendix). As with Heller's estimator, both $\mathbf{A}_\varepsilon(\boldsymbol{\beta})$ and $\mathbf{B}_\varepsilon(\boldsymbol{\beta})$ are directly estimable from the data, the latter derived from a theory of *U*-statistics.

References

Boyd, SP.; Vandenberghe, L. Convex optimization. Cambridge Univ Pr; 2004.

Brown BM, Wang YG. Standard errors and covariance matrices for smoothed rank estimators. Biometrika. 2005; 92:149–158.

Brown BM, Wang YG. Induced smoothing for rank regression with censored survival times. Statist Med. 2007; 26:828–836.

Cai T, Huang J, Tian L. Regularized estimation for the accelerated failure time model. Biometrics. 2009; 65:394–404. [PubMed: 18573133]

CAMDA. Critical assessment of microarray data analysis. 2003 http://www.camda.duke.edu/camda03.html.

Candes E, Tao T. The dantzig selector: Statistical estimation when p is much larger than n. The Annals of Statistics. 2007; 35(6):2313–2351.

Chung J, Chung M, O'Leary D. Designing optimal filters for ill-posed inverse problems. SIAM Journal on Scientific Computing. 2011; 33(6):3132–3152.

Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society Series B. 1972; 34:187–220.

Cox, DR.; Oakes, D. Analysis of Survival Data. London: Chapman and Hall; 1984.

Dickson ER, Grambsch PM, Fleming TR, Fisher LD, Langworthy A. Prognosis in primary biliary cirrhosis: model for decision making. Hepatology. 1989; 10(1):1–7. [PubMed: 2737595]

Fleming, TR.; Harrington, DP. Counting processes and survival analysis. Vol. vol 8. New York: Wiley; 1991.

Fygenson M, Ritov Y. Monotone estimating equations for censored data. The Annals of Statistics. 1994; 22:732–746.

Gehan EA. A generalized wilcoxon test for comparing arbitrarily single-censored samples. Biometrika. 1965; 52:203–223. [PubMed: 14341275]

Gill, PE.; Murray, W.; Wright, MH. Practical optimization. Academic press; 1981.

Hadamard J. Sur les problèmes aux dèrivèes partielles et leur signification physique. 1902

Hastie, T.; Tibshirani, R.; J, F. 2nd Edition. New York: Springer; 2009. The Elements of Statistical Learning.

Heller G. Smoothed rank regression with censored data. Journal of the American Statistical Association. 2007; 102(478):552–559.

Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970:55–67.

Huang J, Ma S, Xie H. Regularized estimation in the accelerated failure time model with high-dimensional covariates. Biometrics. 2006:813–820. [PubMed: 16984324]

Huber PJ. Robust estimation of a location parameter. The Annals of Mathematical Statistics. 1964; 35(1):73–101.

Hunter DR, Lange K. A tutorial on mm algorithms. The American Statistician. 2004:30–37.

Jin Z, Lin DY, Wei LJ, Ying Z. Rank-based inference for the accelerated failure time model. Biometrika. 2003; 90(2):341–353.

Johnson BA. Variable selection in semiparametric linear regression with censored data. J R Statist Soc Ser B. 2008; 70:351–370.

Johnson BA. Rank-based estimation in the ℓ_1-regularized partly linear model model with application to integrated analyses of clinical predictors and gene expression data. Biostatistics. 2009a; 10:659–666. [PubMed: 19553356]

Johnson BA. On lasso for censored data. Electronic Journal of Statistics. 2009b; 3:485–506.

Johnson BA, Lin D, Zeng D. Penalized estimating functions and variable seleciton in semiparametric regression models. Journal of the American Statistical Association. 2008; 103:672–680. [PubMed: 20376193]

Johnson BA, Long Q, Chung M. On path restoration for censored outcomes. Biometrics. 2011; 67:1379–1388. [PubMed: 21457193]

Johnson LM, Strawderman RL. Induced smoothing for the semiparametric accelerated failure time model: asymptotics and extensions to clustered data. Biometrika. 2009; 96(3):577–590. [PubMed: 23049117]

Kaipio, JP.; Somersalo, E. Springer Science+ Business Media, Inc.; 2005. Statistical and computational inverse problems.

Kalbeisch, JD.; Prentice, RL. 2nd edn.. Vol. vol 5. New York: Wiley; 1980. The statistical analysis of failure time data.

Koenker R, Bassett G Jr. Regression quantiles. Econometrica: Journal of the Econometric Society. 1978:33–50.

Koenker R, Ng P. A Frisch-Newton algorithm for sparse quantile regression. Acta Mathematicae Applicatae Sinica (English Series). 2005; 21(2):225–236.

Koenker RW, D'Orey V. Algorithm as 229: Computing regression quantiles. Journal of the Royal Statistical Society Series C (Applied Statistics). 1987; 36(3):383–393.

Lin DY, Geyer CJ. Computational methods for semiparametric linear regression with censored data. Journal of Computational and Graphical Statistics. 1992; 1(1):77–90.

Meier L, Van De Geer S, Bühlmann P. The group lasso for logistic regression. group. 2008; 70(Part 1): 53–71.

Morris C, Norton E, Zhou X. Parametric duration analysis of nursing home usage. Case Studies in Biometry. 1994:231–248.

Nocedal, J.; Wright, SJ. 2nd edn. Springer verlag; 2006. Numerical optimization.

Owen, AB. Palo Alto, CA: Department of Statistics, Stanford University; 2006. A robust hybrid of lasso and ridge regression., technical report.

Prentice RL. Linear rank tests with right censored data. Biometrika. 1978; 65(1):167–179.

Reid N. A conversation with sir david cox. Statistical Science. 1994; 9:439–455.

Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996; 58(1):267–288.

Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005; 67(1):91–108.

Tsiatis AA. Estimating regression parameters using linear rank tests for censored data. The Annals of Statistics. 1990; 18(1):354–372.

Vogel CR. Computational methods for inverse problems. Society for Industrial Mathematics. 2002; vol 23

Wei LJ, Ying Z, Lin DY. Linear regression analysis of censored survival data based on rank tests. Biometrika. 1990; 77(4):845–851.

Wu S, Shen X, Geyer CJ. Adaptive regularization using the entire solution surface. Biometrika. 2009; 96(3):513–527.

Xu J, Leng C, Ying Z. Rank-based variable selection with censored data. Statistics and Computing. 2010; 20:165–176.

Ying Z. A large sample study of rank estimation for censored regression data. Annals of Statistics. 1993; 21:76–99.

Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2006; 68(1):49–67.

Zou H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association. 2006; 101:1418–1429.

Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005; 67(2):301–320.

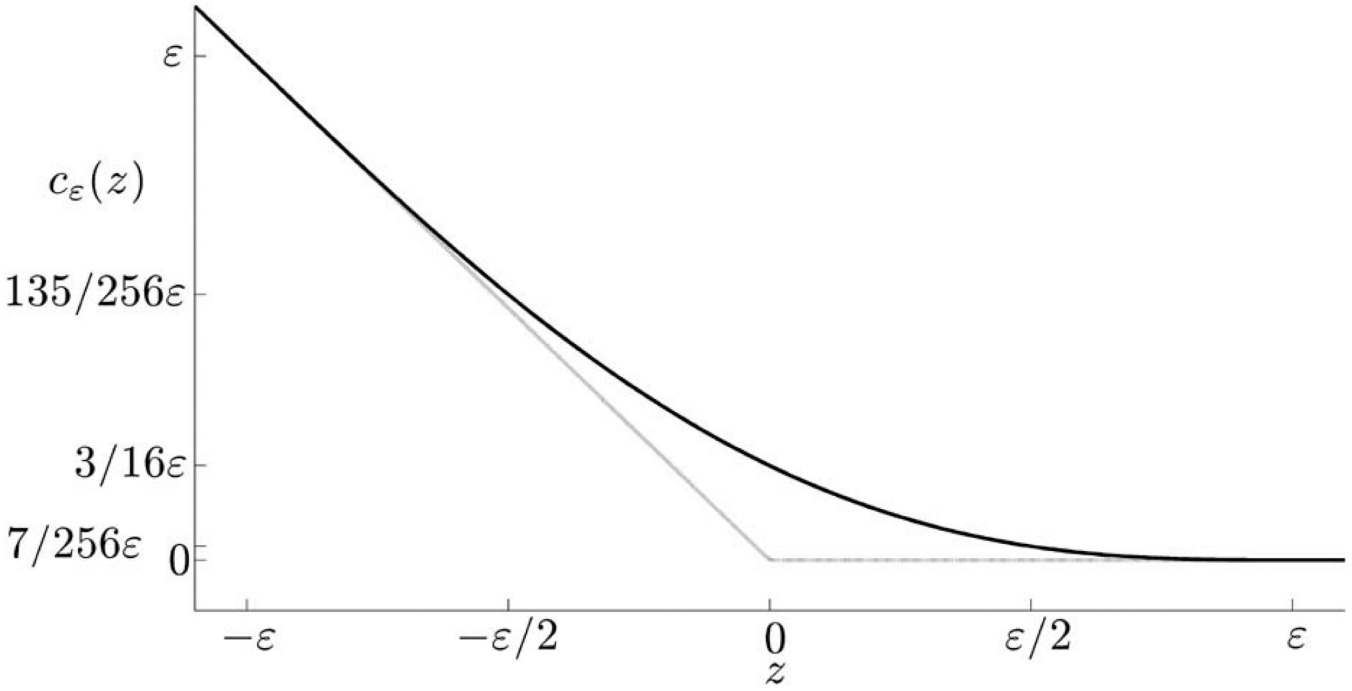

Fig. 1.
Graph of the smoothing function c_ε. Compared with the function $[\cdot]^-$ the error $c_\varepsilon(z) - [z]^-$ is largest at $z = 0$ with an absolute error of $3/16\varepsilon$.

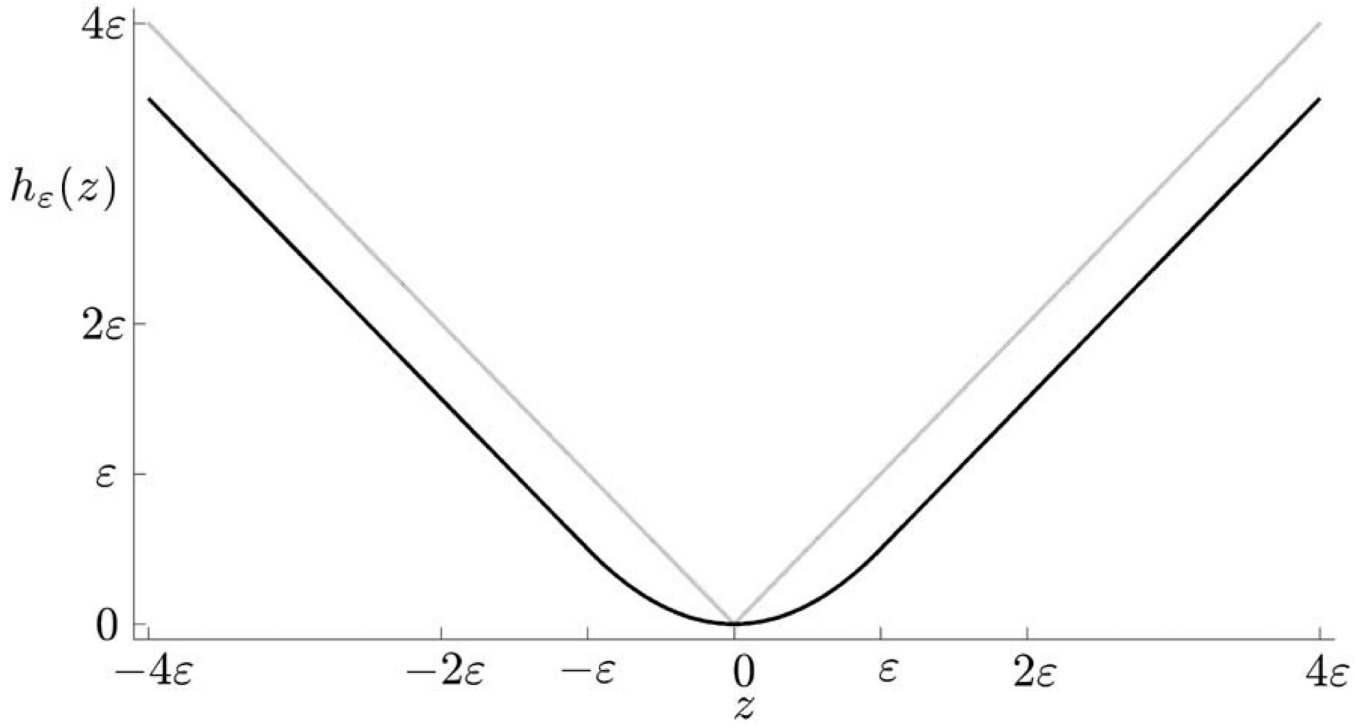

Fig. 2.
Graph of the smoothing function h_ε. Compared with the function $|\cdot|$ the absolute error stays below $\varepsilon/2$ for any z.

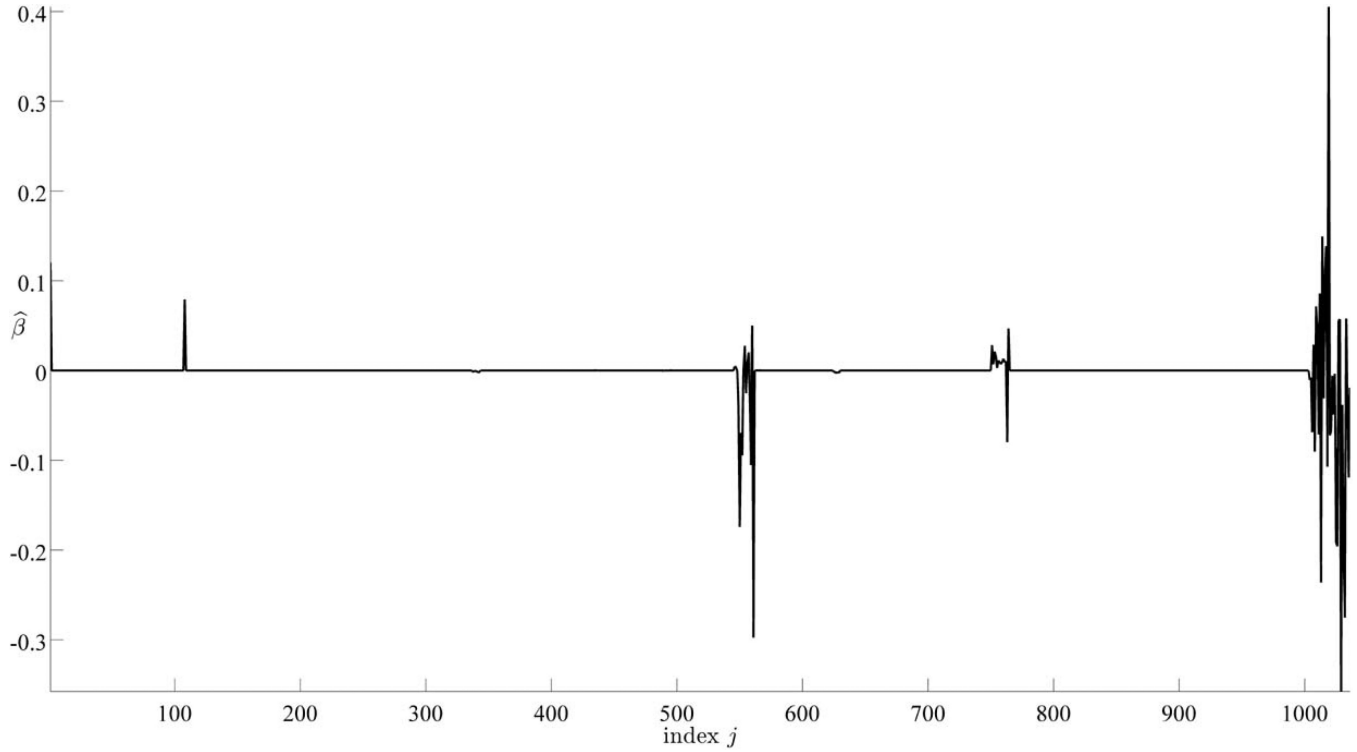

Fig. 3.
Estimated parameters β for the CAMDA data using group lasso regularization.

Table 1

Table of some common penalty functions and their Huber-type approximation.

Penalty	$\mathcal{S}(\boldsymbol{\lambda},\boldsymbol{\beta})$	$\mathcal{S}_{\boldsymbol{\varepsilon}}(\boldsymbol{\lambda},\boldsymbol{\beta})$
ridge	$\lambda \sum_{j=1}^{p} \beta_j^2$	$\lambda \sum_{j=1}^{p} \beta_j^2$
lasso	$\lambda \sum_{j=1}^{p} \lvert\beta_j\rvert$	$\lambda \sum_{j=1}^{p} \begin{cases} \frac{1}{2\varepsilon}\beta_j^2, & \text{if } \lvert\beta_j\rvert<\varepsilon, \\ \lvert\beta_j\rvert - \frac{\varepsilon}{2}, & \text{if } \lvert\beta_j\rvert \geq \varepsilon, \end{cases}$
elastic net	$\lambda_1 \sum_{j=1}^{p} \lvert\beta_j\rvert + \lambda_2 \sum_{j=1}^{p} \beta_j^2$	$\sum_{j=1}^{p} \begin{cases} (\frac{\lambda_1}{2\varepsilon}+\lambda_2)\beta_j^2, & \text{if } \lvert\beta_j\rvert<\varepsilon, \\ \lambda_1(\lvert\beta_j\rvert - \frac{\varepsilon}{2}) + \lambda_2\beta_j^2, & \text{if } \lvert\beta_j\rvert \geq \varepsilon, \end{cases}$
Berhu	$\lambda \sum_{j=1}^{p} \begin{cases} \lvert\beta_j\rvert, & \text{if } \lvert\beta_j\rvert \leq \gamma, \\ \frac{\beta_j^2+\gamma^2}{2\gamma}, & \text{if } \lvert\beta_j\rvert>\gamma, \end{cases}$	$\lambda \sum_{j=1}^{p} \begin{cases} \frac{1}{2\varepsilon}\beta_j^2, & \text{if } \lvert\beta_j\rvert<\varepsilon, \lvert\beta_j\rvert \leq \gamma \\ \lvert\beta_j\rvert - \frac{\varepsilon}{2}, & \text{if } \lvert\beta_j\rvert \geq \varepsilon, \lvert\beta_j\rvert \leq \gamma \\ \frac{\beta_j^2+\gamma^2}{2\gamma}, & \text{if } \lvert\beta_j\rvert>\gamma, \end{cases}$
total variation	$\lambda \sum_{j=1}^{p-1} \lvert\beta_{j+1} - \beta_j\rvert$	$\lambda \sum_{j=1}^{p-1} \begin{pmatrix} \frac{1}{2\varepsilon}(\beta_{j+1} - \beta_j)^2 & \text{if } \lvert\beta_{j+1} - \beta_j\rvert<\varepsilon \\ \lvert\beta_{j+1} - \beta_j\rvert - \frac{\varepsilon}{2}, & \text{if } \lvert\beta_{j+1} - \beta_j\rvert \geq \varepsilon \end{pmatrix}$
fused lasso	$\lambda_1 \sum_{j=1}^{p} \lvert\beta_j\rvert + \lambda_2 \sum_{j=1}^{p-1} \lvert\beta_{j+1} - \beta_j\rvert$	combined lasso and total variation smoothing
group lasso	$\lambda \sum_{k=1}^{K} \sqrt{\aleph(G_k)} \lVert\beta_k\rVert_2$	$\lambda \sum_{k=1}^{J} \sqrt{\aleph(G_k)} \lVert\beta_k\rVert_2$

Table 2

Coefficients estimates for Mayo PBC data.

Parameter	$\hat{\beta}_B$	$\hat{\beta}_H$	$\hat{\beta}_{G,e}$	$\hat{\beta}_G$
age	−0.344 (0.073)	−0.470 (0.102)	−0.270 (0.057)	−0.271 (0.062)
albumin	0.226 (0.089)	0.200 (0.113)	0.205 (0.070)	0.204 (0.069)
bilirubin	−0.721 (0.082)	−0.925 (0.110)	−0.593 (0.068)	−0.594 (0.071)
edema	−0.248 (0.083)	−0.231 (0.099)	−0.223 (0.069)	−0.224 (0.070)
protime	−0.295 (0.086)	−0.347 (0.106)	−0.238 (0.071)	−0.237 (0.080)

Table 3

Coefficient estimates for nursing home data.

Parameter	$\hat{\beta}_B$	$\hat{\beta}_H$	$\hat{\beta}_{G,e}$	$\hat{\beta}_G$
trt	0.141 (0.107)	0.140 (0.108)	0.145 (0.107)	0.144 (0.103)
age	0.096 (0.055)	0.096 (0.055)	0.096 (0.052)	0.096 (0.054)
sex	−0.633 (0.125)	−0.633 (0.125)	−0.628 (0.127)	−0.629 (0.118)
mar stat	−0.249 (0.144)	−0.249 (0.145)	−0.247 (0.142)	−0.252 (0.135)
h1	−0.093 (0.146)	−0.095 (0.148)	−0.092 (0.145)	−0.091 (0.141)
h2	−0.589 (0.130)	−0.591 (0.131)	−0.588 (0.126)	−0.587 (0.131)
h3	−1.073 (0.174)	−1.076 (0.176)	−1.071 (0.169)	−1.071 (0.160)