Research Article

Statistics in Medicine

Received 21 June 2012, Accepted 18 March 2013 Published online 28 April 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5817

Large-scale parametric survival analysis

Sushil Mittal,[a*†] David Madigan,[a] Jerry Q. Cheng[b] and Randall S. Burd[c]

Survival analysis has been a topic of active statistical research in the past few decades with applications spread across several areas. Traditional applications usually consider data with only a small numbers of predictors with a few hundreds or thousands of observations. Recent advances in data acquisition techniques and computation power have led to considerable interest in analyzing very-high-dimensional data where the number of predictor variables and the number of observations range between 10^4 and 10^6. In this paper, we present a tool for performing large-scale regularized parametric survival analysis using a variant of the cyclic coordinate descent method. Through our experiments on two real data sets, we show that application of regularized models to high-dimensional data avoids overfitting and can provide improved predictive performance and calibration over corresponding low-dimensional models. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: survival analysis; parametric models; regularization; penalized regression; pediatric trauma

1. Introduction

Regression analysis of time-to-event data occupies a central role in statistical practice [1, 2] with applications spread across several fields including biostatistics, sociology, economics, demography, and engineering [3–6]. Newer applications often gather high-dimensional data that present computational challenges to existing survival analysis methods. As an example, new technologies in genomics have led to high-dimensional microarray gene expression data where the number of predictor variables is of the order of 10^5. Other large-scale applications include medical adverse event monitoring, longitudinal clinical trials, and business data mining tasks. All of these applications require methods for analyzing high-dimensional data in a survival analysis framework.

In this paper, we consider high-dimensional parametric survival regression models involving both large numbers of predictor variables and large numbers of observations. Although Cox models continue to attract much attention [7], parametric survival models have always been a popular choice among statisticians for analyzing time-to-event data [4, 6, 8, 9]. Parametric survival models feature prominently in commercial statistical software, are straightforward to interpret, and can provide competitive predictive accuracy. To bring all these advantages to high-dimensional data analysis, these methods need to be scaled to data involving 10^4–10^6 predictor variables and even larger numbers of observations.

Computing the maximum likelihood fit of a parametric survival model requires solving a nonlinear optimization problem. Standard implementations work well for small-scale problems. Because these approaches typically require matrix inversion, solving large-scale problems using standard software is typically impossible. One possible remedy is to perform feature selection as a preprocessing step. Although feature selection does reduce memory and computational requirements and also serves as a practical solution to overfitting, it introduces new problems. First, the statistical consequences of most feature selection methods remain unclear, making it difficult to choose the number of features for a

[a] *Department of Statistics, Columbia University, 1255 Amsterdam Avenue, New York, NY 10027, U.S.A.*
[b] *Cardiovascular Institute, UMDNJ-Robert Wood Johnson Medical School, 125 Patterson Street, Clinical Academic Building, New Brunswick, NJ 08901, U.S.A.*
[c] *Division of Trauma and Burn Surgery, Joseph E. Robert Jr. Center for Surgical Care, Children's National Medical Center, Washington, DC, U.S.A.*
**Correspondence to: Sushil Mittal, Department of Statistics, Columbia University, 1255 Amsterdam Avenue, New York, NY 10027, U.S.A.*
[†] *E-mail: mittal@stat.columbia.edu*

given task in a principled way. Second, the most efficient feature selection methods are greedy and may choose redundant or ineffective combinations of features. Finally, it is often unclear how to combine heuristic feature selection methods with domain knowledge. Even when standard software does produce estimates, numerical ill conditioning can result in a lack of convergence, large estimated coefficient variances, and poor predictive accuracy or calibration.

In this paper, we describe a regularized approach to parametric survival analysis [10]. The main idea is the use of a regularizing prior probability distribution for the model parameters that favors sparseness in the fitted model, leading to point estimates being zero for many of the model parameters. To solve the optimization problem, we use a variation of the cyclic coordinate descent method [11–14]. We show that application of this type of model to high-dimensional data avoids overfitting, can provide improved predictive performance and calibration over corresponding low-dimensional models, and is efficient both during fitting and at prediction time.

In Section 2, we review some of the other works that address similar problems. In Section 3, we describe the basics of a regularized approach to parametric survival analysis. In Section 4, we provide a brief overview of the four parametric models used for survival analysis using a high-dimensional formulation. We also describe our optimization technique, which is tailored for each specific model for computing point estimates of model parameters. More involved algorithmic details of the method pertaining to each of the four parametric models can be found in Appendix A. We describe the data sets and methods that we use in our experiments in Section 5 and provide experimental results in Section 6. The first application uses a large data set of hospitalized injured children for developing a model for predicting survival. Through our experiments, we establish that an analysis that uses our proposed approach can add significantly to predictive performance as compared with the traditional low-dimensional models. In the second application, we apply our method to a publicly available breast cancer gene expression data set and show that the high-dimensional parametric model can achieve performance similar to a low-dimensional Cox model while being better calibrated. Finally, we conclude in Section 8 with directions for future work.

We have publicly released the C++ implementation of our algorithm, which can be downloaded from http://code.google.com/p/survival-analysis-cmake/. This code has been derived from the widely used BBR/BXR software for performing large-scale Bayesian logistic regression (http://www.bayesianregression.org). All four parametric models discussed in this paper are included in the implementation; however, it is fairly straightforward to also extend it to other parametric models. The computation of various evaluation metrics discussed in Section 5 is also integrated into the code.

2. Related work

Survival analysis is an old subject in statistics that continues to attract considerable research attention. Traditionally, survival analysis is concerned with the study of survival times in clinical and health-related studies [4, 15]. However, over the past several decades, survival analysis has found an array of applications ranging from reliability studies in industrial engineering to analyses of inter-child birth times in demography and sociology [3]. Other application areas have also benefited from the use of these methods [6, 8].

Earlier applications usually consisted of a relatively small number of predictors (usually less than 20) with a few hundred or sometimes a few thousand examples. Recently, there has been considerable interest in analyzing high-dimensional time-to-event problems. For instance, a large body of work has focused on methodologies to handle the overwhelming amount of data generated by new technologies in biology such as gene expression microarrays and single nucleotide polymorphism data. The goal of the work in high-dimensional survival analysis has been both to develop new statistical methods [16–21] and to extend the existing methods to handle new data sets [13, 22–24]. For example, recent work [16, 18] has extended the traditional support vector machines used for regression to survival analysis by additionally penalizing discordant pairs of observations. Another method [19, 20] extends the use of random forests for variable selection for survival analysis. Similarly, other methods [22, 25] have used an elastic net approach for variable selection both under the Cox proportional hazards model and under an accelerated failure time model. This method is similar to another work [24] that applies an efficient method to compute L1-penalized parameter estimates for Cox models. A recent review [26] provides a survey of the existing methods for variable selection and model estimation for high-dimensional data.

Some more recent tools such as `coxnet` [27] and `fastcox` [25] adopt optimization approaches that can scale to the high-dimensional high-sample-size data that we focus on. Although other models

provided by the `R` package `glmnet` do support sparse formats, neither `coxnet` nor `fastcox` currently supports a sparse matrix format for the input data. However, both `coxnet` and `fastcox` provide estimates for the Cox proportional hazards model, and `fastcox` supports elastic net regularization.

3. Regularized survival analysis

Denote by n the number of individuals in the training data. We represent their survival times by $y_i = \min(t_i, c_i),\ i = 1, \ldots, n$, where t_i and c_i are the time to event (failure time) and right-censoring time for each individual, respectively. Let $\delta_i = I(t_i \leqslant c_i)$ be the indicator variable such that δ_i is 1 if the observation is not censored and 0 otherwise. Further, let $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]^\top$ be a p-vector of covariates. We assume that t_i and c_i are conditionally independent given $\mathbf{x}_i$ and that the censoring mechanism is noninformative. The observed data comprise triplets $D = \{(y_i, \delta_i, \mathbf{x}_i) : i = 1, \ldots, n\}$.

Let θ be the set of unknown, underlying model parameters. We assume that the survival times $y_1, y_2, \ldots, y_n$ arise in an independent and identically distributed fashion from density and survival functions $f(y|\theta)$ and $S(y|\theta)$, respectively, parametrized by θ. We are interested in the likelihood $L(\theta|D)$ of the parametric model, where

$$L(\theta|D) = \prod_{i=1}^{n} f(y_i|\theta)^{\delta_i} S(y_i|\theta)^{(1-\delta_i)}. \tag{1}$$

We analyze and compare the performance of four different parametric models by modeling the distributions of the survival times using exponential, Weibull, log-logistic, or lognormal distributions. Each of these distributions can be fully parametrized by the parameter pair $\theta = (\lambda, \alpha)$. Typically, the parameter λ is reparametrized in terms of the covariates $\mathbf{x} = [x_1, x_2, \ldots, x_p]^\top$ and the vector $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_p]^\top$ such that $\lambda_i = \phi\left(\boldsymbol{\beta}^\top \mathbf{x}_i\right),\ i = 1, \ldots, n$. In general, the mapping function $\phi(\cdot)$ is different for each model, and standard choices exist. The likelihood function of (1) in terms of the new parameters can be written as

$$L(\boldsymbol{\beta}, \alpha|D) = \prod_{i=1}^{n} f(y_i|\lambda_i, \alpha)^{\delta_i} S(y_i|\lambda_i, \alpha)^{(1-\delta_i)}. \tag{2}$$

The parameters $\boldsymbol{\beta}$ and α are estimated by maximizing their joint posterior density

$$L(\boldsymbol{\beta}, \alpha) \propto L(\boldsymbol{\beta}, \alpha|D)\pi(\boldsymbol{\beta})\pi(\alpha). \tag{3}$$

The joint posterior distribution of $(\boldsymbol{\beta}, \alpha)$ does not usually have a closed-form solution, but it can be shown that the conditional posterior distributions $\pi(\boldsymbol{\beta}|\alpha, D)$ and $\pi(\alpha|\boldsymbol{\beta}, D)$ are concave and therefore can be solved efficiently. In practice, it is sufficient to estimate $\boldsymbol{\beta}$ and α by maximizing the conditional posterior of just $\boldsymbol{\beta}$

$$L(\boldsymbol{\beta}) \propto L(\boldsymbol{\beta}, \alpha|D)\pi(\boldsymbol{\beta}). \tag{4}$$

For a model to generalize well to unseen test data, it is important to avoid overfitting the training data. In the Bayesian paradigm, this goal can be achieved by specifying appropriate prior distribution on $\boldsymbol{\beta}$ such that each $\boldsymbol{\beta}_j$ is likely to be near 0. As we will observe in the following sections, both Gaussian and Laplacian priors fall under this category. Because we focus on posterior mode estimation in this paper, one can view our procedure as Bayesian or simply as a form of regularization or penalization. We use the Bayesian terminology in part because we view fully Bayesian computation as the next desirable step in large-scale survival analysis. We return to this point in the conclusion section.

3.1. *Gaussian priors and ridge regression*

For L2 regularization, we assume a Gaussian prior for each β_j with 0 mean and variance τ_j, that is,

$$\pi(\beta_j|\tau_j) = \mathrm{N}(0, \tau_j) = \frac{1}{\sqrt{2\pi\tau_j}} \exp\left(-\frac{\beta_j^2}{2\tau_j}\right). \tag{5}$$

The mean of 0 encodes a prior preference for values of β_j that are close to 0. The variances τ_j are positive constants that control the degree of regularization and are typically chosen through cross-validation.

Smaller values of τ_j imply a stronger belief that β_j is being close to 0 whereas larger values impose a less-informative prior. In the simplest case, we assume that $\tau_1 = \tau_2 = \ldots = \tau_p$. Assuming that the components of $\boldsymbol{\beta}$ are independent *a priori*, the overall prior for $\boldsymbol{\beta}$ can be expressed as the product of the priors of each individual β_j, that is, $\pi(\boldsymbol{\beta}|\tau_1, \ldots, \tau_p) = \prod_{j=1}^{p} \pi(\beta_j|\tau_j)$. Finding the maximum a posteriori estimate of $\boldsymbol{\beta}$ with this prior is equivalent to performing a ridge regression [28]. Note that although the Gaussian prior favors the value of β_j close to 0, posterior modes are generally not exactly equal to 0 for any $\boldsymbol{\beta}_j$.

3.2. *Laplacian prior and lasso regression*

For L1 regularization, we again assume that each β_j follows a Gaussian distribution with mean 0 and variance τ_j. However, instead of fixed τ_j's, we assume that each τ_j arises from an exponential distribution parametrized using γ_j's and having a density of

$$\pi(\tau_j|\gamma_j) = \frac{\gamma_j}{2} \exp\left(-\frac{\gamma_j}{2}\tau_j\right). \tag{6}$$

Integrating out τ_j gives an equivalent nonhierarchical, double-exponential (Laplace) distribution with a density of

$$\pi(\beta_j|\gamma_j) = \frac{\sqrt{\gamma_j}}{2} \exp\left(-\sqrt{\gamma_j}|\beta_j|\right). \tag{7}$$

Again, in the simplest case, we assume that $\gamma_1 = \gamma_2 = \ldots = \gamma_p$. Similar to the Gaussian prior, assuming that the components of $\boldsymbol{\beta}$ are independent, $\pi(\boldsymbol{\beta}|\gamma_1, \ldots, \gamma_p) = \prod_{j=1}^{p} \pi(\beta_j|\gamma_j)$. Finding the maximum *a posteriori* estimate of $\boldsymbol{\beta}$ with the Laplacian prior is equivalent to performing a Lasso regression [29]. With this approach, a sparse solution typically ensues, meaning the posterior mode for many components of the $\boldsymbol{\beta}$ vector will be 0.

4. Parametric models

It can be shown that for both Gaussian and Laplacian regularization, the negated log-posterior of (4) for all the four parametric models considered in our work are log-convex. Therefore, a wide range of optimization algorithms can be used. Because of the high dimensionality of our target applications, the usual methods such as Newton–Raphson cannot be used because of their high memory requirements. Many alternate optimization approaches have been proposed for maximum a posteriori estimation for high-dimensional regression problems [30, 31]. We use the Combined Local and Global (CLG) algorithm of [30], which is a type of cyclic coordinate descent algorithm, because of its favorable property of scaling to high-dimensional data and ease of implementation. This method was successfully adapted to the lasso [14] for performing large-scale logistic regression and implemented in the widely used BBR/BXR software (http://www.bayesianregression.org).

A cyclic coordinate descent algorithm begins by setting all p parameters β_j, $j = 1, \ldots, p$, to some initial value. It then sets the first parameter to a value that minimizes the objective function, holding all other parameters constant. This problem then is a one-dimensional optimization. The algorithm finds the minimizing value of a second parameter, while holding all other values constant (including the new value of the first parameter). The third parameter is then optimized, and so on. When all variables have been traversed, the algorithm returns to the first parameter and starts again. Multiple passes are made over the parameters until some convergence criterion is met. We note that similar to previous work [14], instead of iteratively updating each parameter till convergence, we do it only once before proceeding on to the next parameter. Because the optimal values of the other parameters are themselves changing, tuning a particular parameter to very high precision, in each pass of the algorithm, is not necessary. For more details of the CLG method, we refer readers to relevant publications in this area [14, 30]. The details of our algorithm for different parametric models are described in Appendix A. We also note that for the case of the Laplacian prior, the derivative of the negated log-posterior is undefined for $\beta_j = 0$, $j = 1, \ldots, p$. Section 4.3 of [14] describes the modification of the CLG algorithm that we utilized to address this issue.

5. Experiments

We tested our algorithm for large-scale parametric survival analysis on two different data sets. In the following, we briefly describe the data sets and also motivate their choice for our work. In both cases, we compare the performance of the high-dimensional parametric model trained on all predictors with that of a low-dimensional model trained on a small subset of predictors.

5.1. Pediatric trauma data

Trauma is the leading cause of death and acquired disability in children and adolescents. Injuries result in more deaths in children than all other causes combined [32]. To provide optimal care, children at high risk for mortality need to be identified and triaged to centers with the resources to manage these patients. The overall goal of our analysis is to develop a model for predicting mortality after pediatric injury. For prediction within the domain of trauma, the literature has traditionally focused on low-dimensional analysis, that is, modeling using a small set of features from the injury scene or emergency department [33, 34]. Although approaches that use low-dimensional data to predict outcome may be easier to implement, the trade-off may be poorer predictive performance.

We obtained our data set from the National Trauma Data Bank (NTDB), a trauma database maintained by the American College of Surgeons. The data set includes 210,555 patient records of injured children aged <15 years collected over 5 years (2006–2010). We divided these data into a training data set (153,402 patients for years 2006–2009) and a testing data set (57,153 patients for year 2010). The mortality rate of the training set is 1.68%, whereas that of the test set is 1.44%. There are a total of 125,952 binary predictors indicating the presence or absence of a particular attribute (or interaction among various attributes). The high-dimensional model was trained using all the 125,952 predictors, whereas the low-dimensional model used only 41 predictors. The information about various predictors used for both high-dimensional and low-dimensional models is summarized in Table I.

5.2. Breast cancer gene expression data

For our second application, we analyzed the well-known breast cancer gene expression data set [35, 36]. This data set is publicly available and consists of cDNA expression profiles of 295 tumor samples of patients with breast cancer. These patients were diagnosed with breast cancer between 1984 and 1995 at the Netherlands Cancer Institute and were aged 52 years or younger at the time of diagnosis. Overall, 79 (26.78%) patients died during the follow-up time, and the remaining 216 are censored. The total number of predictors (number of genes) is 24,885, each of which represents the log ratio of the intensities of the two color dyes used for a specific gene. The high-dimensional model was trained using all the 24,885 predictors. For the low-dimensional model, we used the `glmpath` (`coxpath`) package of `R` [37] and generated the entire regularized paths and output predictors in the order of their relative importance.

Table I. Description of the predictors used for high-dimensional and low-dimensional models for pediatric trauma data.

Predictor type	# Predictors	Description	High-dim	Low-dim
Main effects				
ICD-9 codes	1890	International Classification of Disease, Ninth Revision	✓	✗
AIS codes (predots)	349	Abbreviated Injury Scale codes that include body region, anatomic structure associated with the injury, and the level of injury	✓	✗
Interactions/combinations				
ICD-9, ICD-9	102,284	Co-occurrences of two ICD-9 injury codes	✓	✗
AIS code, AIS code	20,809	Co-occurrences of two AIS codes	✓	✗
Body region, AIS score	41	Combinations of any of the nine body regions associated to an injury with injury severity score (between 1 and 6) determined according to the AIS coding scheme	✓	✓
[Body region, AIS score], [Body region, AIS score]	579	Co-occurrences of two [body region, AIS score] combinations	✓	✗

We picked the top five predictors from the list and trained a low-dimensional model for comparisons. The data were then randomly split into training (67%) and testing (33%) sets such that the mortality in both parts was equal to the original mortality rate.

5.3. *Hyperparameter selection*

The Gaussian and Laplacian priors both require a prior variance, $\sigma_j^2,\ j = 1, \ldots, p$, for parameter values. The actual hyperparameters are $\tau = \tau_j = \sigma_j^2$ for Gaussian and $\sqrt{\gamma} = \sqrt{\gamma_j} = \sqrt{2}/\sigma_j$ for Laplacian. For both the applications, the regularization parameter was selected using a fourfold cross-validation on the training data. The variance σ_j^2 was varied between 10^{-5} and 10^6 by multiples of 10. For the Gaussian loss, this amounts to varying the actual regularization parameter between 10^{-5} and 10^6 by multiples of 10, whereas that for the Laplacian was between 0.0014 and 447.21 in multiples of $\sqrt{10}$. For each choice of the hyperparameter, we computed the sum of log-likelihoods for the patients in all the four validation sets and chose the hyperparameter value that maximized this sum.

5.4. *Performance evaluation*

There are several ways to compare the performance of fitted parametric models. These evaluation metrics can be divided into two categories—ones that assess the discrimination quality of the model and others that assess the calibration. As for any other regression model, the log-likelihood on the test data is an obvious choice for measuring the performance of penalized parametric survival regression. In the past several years, survival analysis-specific metrics that take censoring into account have been proposed. Among these metrics, the area under the ROC curve (AUC), also known as Harrell's c-statistic [38, 39], a measure of discriminative accuracy of the model, has become one of the important criteria used to evaluate the performance of survival models [40]. Motivated from a similar test for logistic regression model proposed by Hosmer and Lemeshow, we evaluated calibration using an overall goodness-of-fit test that has been proposed for survival data [41, 42]. Although the original test was proposed for Cox models, its use for parametric models has also been described [6, Chapter 8]. In the following, we briefly describe these two metrics.

5.4.1. Harrell's c-statistic. Harrell's c-statistic is an extension of the traditional area under the ROC curve statistic but is more suited for time-to-event data because it is independent of any thresholding process. The method is based on the comparison of estimated and ground truth ordering of risks between pairs of *comparable* subjects. Two subjects are said to be comparable if at least one of the subjects in the pair has developed the event (e.g., death) and if the follow-up time duration for that subject is less than that of the other. By using the notation of Section 3, the comparability of an ordered pair of subjects (i, j) can be represented by the indicator variable $\zeta_{ij} = I(y_i < y_j,\ \delta_i = 1)$, such that $\zeta_{i,j}$ is 1 when the two subjects are comparable and 0 otherwise. The total number of comparable pairs in the test data containing n subjects can then be computed as

$$n_\zeta = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \zeta_{i,j}. \tag{8}$$

To measure the predicted discrimination, the *concordance* of the comparable pairs is then estimated. A pair of comparable subjects (as defined earlier) is said to be concordant if the estimated risk of the subject who developed the event earlier is more than that of the other subject. Therefore, the concordance of the ordered subject pair (i, j) can also be represented using the indicator variable $\xi_{ij} = I(\zeta_{ij} = 1,\ r_i > r_j)$, where $r_i = -\boldsymbol{\beta}^\top \mathbf{x}_i$ and $r_j = -\boldsymbol{\beta}^\top \mathbf{x}_j$ are the relative risk scores of the ith and jth subjects. Therefore, ξ_{ij} is 1 when the two subjects are concordant and 0 otherwise. The total number of concordant pairs can then be written as

$$n_\xi = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \xi_{i,j}. \tag{9}$$

Finally, the c-statistic is given by

$$c\text{-statistic} = \frac{n_\xi}{n_\zeta}. \tag{10}$$

5.4.2. Hosmer–Lemeshow statistic. To assess the overall goodness of fit using this test, the subjects are first sorted in order of their relative risk score ($r_i = -\boldsymbol{\beta}^\top \mathbf{x}_i,\ i = 1, \ldots, n$) and divided into G equal-sized groups. The observed number of events o_g of the gth group is obtained by summing the number of noncensored observations in that group, whereas the number of expected events e_g is computed by summing the cumulative hazard of all the subjects in the same group. The χ^2 statistic for the overall goodness of fit is then given by

$$\chi^2 = \sum_{g=1}^{G} \frac{(o_g - e_g)^2}{e_g}. \tag{11}$$

Table II. Comparison of low-dimensional and high-dimensional models for pediatric trauma data with Gaussian penalization.

	Predictors				
Model type	Overall	Selected	Log-likelihood	c-Statistic	χ^2
Exponential					
Low-dim	41	41	−4270.01	0.88	**124.39**
High-dim	125,952	101,733	**−4372.41**	**0.94**	543.28
Weibull					
Low-dim	41	41	−4242.38	0.88	**131.60**
High-dim	125,952	101,794	**−4557.10**	**0.94**	749.22
Log-logistic					
Low-dim	41	41	−4120.66	0.89	95.37
High-dim	125,952	100,889	**−3765.45**	**0.94**	**95.02**
Lognormal					
Low-dim	41	41	−3234.00	0.89	**76.95**
High-dim	125,952	88,244	**−3129.02**	**0.93**	165.68

The number of selected predictors refers to the number of significant predictors ($|\beta_j| > 10^{-4}$). Bold emphases represent superior performance of one model over the other.

Table III. Comparison of low-dimensional and high-dimensional models for pediatric trauma data with Laplacian penalization.

	Predictors				
Model type	Overall	Selected	Log-likelihood	c-Statistic	χ^2
Exponential					
Low-dim	41	41	−4271.23	0.88	122.58
High-dim	125,952	153	**−4034.67**	**0.92**	**94.34**
Weibull					
Low-dim	41	41	−4243.57	0.88	126.84
High-dim	125,952	151	**−3997.28**	**0.92**	**107.99**
Log-logistic					
Low-dim	41	41	−4122.07	0.89	94.73
High-dim	125,952	432	**−3777.83**	**0.94**	**83.00**
Lognormal					
Low-dim	41	41	−3236.79	0.89	**80.71**
High-dim	125,952	168	**−2974.49**	**0.93**	89.36

Bold emphases represent superior performance of one model over the other.

6. Results

We now compare the performance of the low-dimensional and high-dimensional parametric models on both data sets. For both applications, the hyperparameter σ^2 was selected by performing a fourfold cross-validation on the training data.

6.1. Pediatric trauma data

Tables II and III summarize the results for the pediatric trauma data set for low-dimensional and high-dimensional models using all the four parametric models with Gaussian and Laplacian penalization. Note that under the L2 prior, the estimate for any β_j is never exactly 0, and thus, all variables contribute towards the final model. The number of selected predictors in Table II just refers to the number of significant predictors ($|\beta_j| > 10^{-4}$). The Hosmer–Lemeshow χ^2 index was computed using the value of $G = 50$. Both the discriminative measures (log-likelihood and c-statistic) are always significantly better for high-dimensional models than for the corresponding low-dimensional models. In many cases, the high-dimensional model is also better calibrated.

Table IV. Comparison of number of events in low-risk, medium-risk, and high-risk groups for various low-dimensional and high-dimensional models using Gaussian penalization for pediatric trauma data.

			Low-dim		High-dim	
Model type	Risk group	# Subjects	# Events	MST	# Events	MST
Exponential	Low	18,860	10	2.00	8	3.38
	Medium	18,860	41	4.10	18	4.33
	High	19,433	774	3.66	799	3.65
Weibull	Low	18,860	11	2.00	7	3.43
	Medium	18,860	34	3.44	18	3.67
	High	19,433	780	3.69	800	3.66
Log-logistic	Low	18,860	10	2.00	11	3.73
	Medium	18,860	32	4.13	21	4.14
	High	19,433	783	3.66	793	3.65
Lognormal	Low	18,860	10	2.00	16	6.56
	Medium	18,860	31	4.10	39	3.23
	High	19,433	784	3.66	770	3.62

MST, mean observed survival time in days.

Table V. Comparison of number of events in low-risk, medium-risk, and high-risk groups for various low-dimensional and high-dimensional models using Laplacian penalization for pediatric trauma data.

			Low-dim		High-dim	
Model type	Risk group	# Subjects	# Events	MST	# Events	MST
Exponential	Low	18,860	10	2.00	10	3.00
	Medium	18,860	40	3.95	28	7.75
	High	19,433	775	3.67	787	3.52
Weibull	Low	18,860	10	2.00	10	3.00
	Medium	18,860	37	3.32	20	9.80
	High	19,433	778	3.70	795	3.51
Log-logistic	Low	18,860	10	2.00	4	4.50
	Medium	18,860	32	4.13	23	3.65
	High	19,433	783	3.66	798	3.66
Lognormal	Low	18,860	10	2.00	12	3.00
	Medium	18,860	32	4.10	25	2.96
	High	19,433	783	3.67	788	3.69

MST, mean observed survival time in days.

Table VI. Comparison of low-dimensional and high-dimensional models for gene expression data with Gaussian penalization.

	Predictors				
Model type	Overall	Selected	Log-likelihood	*c*-Statistic	χ^2
Exponential					
Low-dim	5	5	**−86.26**	0.71	**16.83**
High-dim	24,496	24,344	−98.72	**0.75**	112.69
Weibull					
Low-dim	5	5	−85.64	**0.71**	12.79
High-dim	24,496	22,299	**−85.56**	0.70	**9.34**
Log-logistic					
Low-dim	5	5	**−85.65**	0.70	14.74
High-dim	24,496	22,090	−86.14	0.70	**5.37**
Lognormal					
Low-dim	5	5	−52.10	**0.70**	**16.02**
High-dim	24,496	24,154	**−65.14**	0.66	23.61

The number of selected predictors refers to the number of significant predictors ($|\beta_j| > 10^{-4}$). Bold emphases represent superior performance of one model over the other.

Table VII. Comparison of low-dimensional and high-dimensional models for gene expression data with Laplacian penalization.

	Predictors				
Model type	Overall	Selected	Log-likelihood	*c*-Statistic	χ^2
Exponential					
Low-dim	5	5	**−86.27**	**0.71**	17.14
High-dim	24,496	13	−87.51	0.67	**2.74**
Weibull					
Low-dim	5	5	**−85.63**	**0.71**	12.79
High-dim	24,496	13	−86.80	0.66	**3.75**
Log-logistic					
Low-dim	5	5	**−85.65**	**0.70**	14.71
High-dim	24,496	9	−86.20	0.68	**3.66**
Lognormal					
Low-dim	5	5	**−52.10**	**0.70**	15.98
High-dim	24,496	9	−53.46	0.66	**8.57**

Bold emphases represent superior performance of one model over the other.

To provide further insight, we grouped the subjects into low-risk, medium-risk, and high-risk groups by sorting their relative risk scores in increasing order and using threshold values at the 33rd and 66th percentiles. For each group, we counted the number of events (number of noncensored observations). Tables IV and V summarize the results for Gaussian and Laplacian penalizations, respectively. The results show that in almost all cases, the subjects assigned to the high-risk group by the high-dimensional models had more events than the ones assigned to the high-risk group by the low-dimensional models. Although both kinds of models assign a similar number of subjects to the low-risk group, the mean observed survival time of the subjects having events is more for the high-dimensional models than for the corresponding low-dimensional ones. These findings also establish that in most cases, the high-dimensional models are better calibrated than their low-dimensional counterparts.

6.2. *Breast cancer gene expression data*

The results of the gene expression data set for low-dimensional and high-dimensional models using all the four parametric models with Gaussian and Laplacian penalizations are summarized in Tables VI and VII, respectively. The Hosmer–Lemeshow χ^2 statistic was computed using $G = 10$ because of very few

Table VIII. Computation time (s) taken for training various low-dimensional and high-dimensional models using Gaussian and Laplacian penalizations for pediatric trauma and gene expression data.

	Pediatric trauma				Gene expression			
	Gaussian		Laplacian		Gaussian		Laplacian	
Model type	Low-dim	High-dim	Low-dim	High-dim	Low-dim	High-dim	Low-dim	High-dim
Exponential	1	4453	1	2588	1	4	1	8
Weibull	3	3406	2	2600	1	12	1	30
Log-logistic	2	2794	2	3278	1	13	1	38
Lognormal	6	3250	6	3101	1	51	1	52

observations in the test set. Although the discriminative performance of both methods is similar for most of the cases, the high-dimensional models are almost always better calibrated.

7. Computation time

Table VIII summarizes the training time taken for fitting different parametric models for low-dimensional and high-dimensional data sets. All the experiments were performed on a system with an Intel 2.4-GHz processor with 8 GB of memory. Note that even though the time taken to fit high-dimensional models to pediatric trauma data set is much more than that taken to fit low-dimensional models, given the scale of the problem (153,402 patients with 125,952 predictors), the performance of the methods may be acceptable in many applications.

8. Conclusions

We present a method to perform regularized parametric survival analysis on data with 10^4–10^6 predictor variables and a large number of observations. Through our experiments in the context of two different applications, we have demonstrated the advantage of using high-dimensional survival analysis over the corresponding low-dimensional models. We have provided a freely available software tool that implements our proposed algorithm. Future work will provide the extension to Cox proportional hazards models in addition to accelerated failure time models and Aalen's additive hazard model. We have also developed software for high-dimensional regularized generalized linear models that utilizes inexpensive massively parallel devices known as graphics processing units (GPU's) [43]. This provides more than an order-of-magnitude speedup and in principle could be further developed to include survival analysis. Fully Bayesian extensions to our current work could explore the hierarchical framework to simultaneously model multiple time-to-event endpoints, model multilevel structure such as patients nested within hospitals, and incorporate prior information when available.

Appendix A

Here, we describe the details of our algorithm for exponential, Weibull, log-logistic, or lognormal distributions of the survival times.

A.1. Exponential

The exponential model is the simplest of all and can be parametrized using a single parameter λ, such that the density and survival functions can respectively be written as $f(y|\theta) = f(y|\lambda) = 1/\lambda \exp(-y/\lambda)$ and $S(y|\theta) = S(y|\lambda) = \exp(-y/\lambda)$. The likelihood function can be written as

$$L(\lambda_1, \ldots, \lambda_n | D) = \prod_{i=1}^{n} f(y_i|\lambda_i)^{\delta_i} S(y_i|\lambda_i)^{(1-\delta_i)} = \exp\left(\sum_{i=1}^{n}\left(-\delta_i \log \lambda_i - \frac{y_i}{\lambda_i}\right)\right). \quad (12)$$

A common form for the mapping function $\phi(\cdot)$ is $\phi\left(\boldsymbol{\beta}^\top \mathbf{x}_i\right) = \exp\left(\boldsymbol{\beta}^\top \mathbf{x}_i\right)$. The likelihood function is then

$$L(\boldsymbol{\beta}|D) = \exp\left(-\sum_{i=1}^{n} \delta_i \boldsymbol{\beta}^\top \mathbf{x}_i - \sum_{i=1}^{n} y_i \exp\left(-\boldsymbol{\beta}^\top \mathbf{x}_i\right)\right), \tag{13}$$

and the corresponding log-likelihood is

$$l(\boldsymbol{\beta}|D) = -\sum_{i=1}^{n} \delta_i \boldsymbol{\beta}^\top \mathbf{x}_i - \sum_{i=1}^{n} y_i \exp\left(-\boldsymbol{\beta}^\top \mathbf{x}_i\right). \tag{14}$$

By adding the Gaussian prior with mean 0 and variance τ_j, the posterior density can be written as

$$l_{\mathrm{G}}(\boldsymbol{\beta}) = l(\boldsymbol{\beta}|D) + \log(\pi(\boldsymbol{\beta}|\tau_1, \ldots, \tau_p)) = l(\boldsymbol{\beta}|D) - \sum_{j=1}^{p} \left(\log\sqrt{\tau_j} + \frac{1}{2}\log 2\pi + \frac{\beta_j^2}{2\tau_j}\right). \tag{15}$$

Similarly, for the Laplacian prior, the posterior density can be written as

$$l_{\mathrm{L}}(\boldsymbol{\beta}) = l(\boldsymbol{\beta}|D) + \log(\pi(\boldsymbol{\beta}|\gamma_1, \ldots, \gamma_p)) = l(\boldsymbol{\beta}|D) - \sum_{j=1}^{p} (\log 2 - \log\sqrt{\gamma_j} + \sqrt{\gamma_j}|\beta_j|). \tag{16}$$

With the CLG algorithm, the one-dimensional problem involves finding $\beta_j^{(\mathrm{new})}$, the value of the jth entry of $\boldsymbol{\beta}$ that minimizes $-l(\boldsymbol{\beta})$, assuming that the other β_j's are held at their current values. Therefore, when (14) and (15) are used for Gaussian prior (and the constants $\log\sqrt{\tau_j}$ and $(1/2)\log 2\pi$ are ignored), finding $\beta_j^{(\mathrm{new})}$ is equivalent to finding the z that minimizes

$$g_{\mathrm{G}}(z) = z\sum_{i=1}^{n} x_{ij}\delta_i + \sum_{i=1}^{n} y_i \exp\left(-zx_{ij} - \sum_{\substack{k=1 \\ k\neq j}}^{p} \beta_k x_{ik}\right) + \frac{z^2}{2\tau_j}. \tag{17}$$

The classic Newton method approximates the objective function $g(\cdot)$ by the first three terms of its Taylor series at the current β_j

$$g(z) \approx g(\beta_j) + g'(\beta_j)(z-\beta_j) + \frac{1}{2}g''(\beta_j)(z-\beta_j)^2, \tag{18}$$

where

$$g'_{\mathrm{G}}(\beta_j) = \left.\frac{\mathrm{d}g_{\mathrm{G}}(z)}{\mathrm{d}z}\right|_{z=\beta_j} = \sum_{i=1}^{n} x_{ij}\delta_i - \sum_{i=1}^{n} y_i x_{ij} \exp\left(-\boldsymbol{\beta}^\top \mathbf{x}_i\right) + \frac{\beta_j}{\tau_j}, \tag{19}$$

$$g''_{\mathrm{G}}(\beta_j) = \left.\frac{\mathrm{d}^2 g_{\mathrm{G}}(z)}{\mathrm{d}z^2}\right|_{z=\beta_j} = \sum_{i=1}^{n} y_i x_{ij}^2 \exp\left(-\boldsymbol{\beta}^\top \mathbf{x}_i\right) + \frac{1}{\tau_j}. \tag{20}$$

Similarly, for the Laplacian prior,

$$g_{\mathrm{L}}(z) = z\sum_{i=1}^{n} x_{ij}\delta_i + \sum_{i=1}^{n} y_i \exp\left(-zx_{ij} - \sum_{\substack{k=1 \\ k\neq j}}^{p} \beta_k x_{ik}\right) + \sqrt{\gamma_j}|z|, \tag{21}$$

$$g'_{\mathrm{L}}(\beta_j) = \left.\frac{\mathrm{d}g_{\mathrm{L}}(z)}{\mathrm{d}z}\right|_{z=\beta_j} = \sum_{i=1}^{n} x_{ij}\delta_i - \sum_{i=1}^{n} y_i x_{ij} \exp\left(-\boldsymbol{\beta}^\top \mathbf{x}_i\right) + \sqrt{\gamma_j}\,\mathrm{sign}(\beta_j), \quad \beta_j \neq 0, \tag{22}$$

$$g_{\mathrm{L}}''(\beta_j) = \left.\frac{\mathrm{d}^2 g_{\mathrm{L}}(z)}{\mathrm{d}z^2}\right|_{z=\beta_j} = \sum_{i=1}^{n} y_i x_{ij}^2 \exp\left(-\boldsymbol{\beta}^\top \mathbf{x}_i\right), \quad \beta_j \neq 0. \tag{23}$$

The value of $\beta_j^{(\mathrm{new})}$ for both types of priors can then be computed as

$$\beta_j^{(\mathrm{new})} = \beta_j + \Delta\beta_j = \beta_j - \frac{g'(\beta_j)}{g''(\beta_j)}. \tag{24}$$

A.2. Weibull

The Weibull model is a more general parametric model with density and survival functions given by $f(y|\theta) = f(y|\lambda,\alpha) = (\alpha y^{\alpha-1}/\lambda)\exp(-y^\alpha/\lambda)$ and $S(y|\theta) = S(y|\lambda,\alpha) = \exp(-y^\alpha/\lambda)$, respectively. The corresponding likelihood function can be written as

$$L(\lambda_1,\ldots,\lambda_n,\alpha|D) = \prod_{i=1}^{n} f(y_i|\lambda_i,\alpha)^{\delta_i} S(y_i|\lambda_i,\alpha)^{(1-\delta_i)} = \alpha^d \prod_{i=1}^{n}\left(\frac{y_i^{\alpha-1}}{\lambda_i}\exp\left(-\frac{y_i^\alpha}{\lambda_i}\right)\right)^{\delta_i}\left(\exp\left(-\frac{y_i^\alpha}{\lambda_i}\right)\right)^{(1-\delta_i)} \tag{25}$$

where $d = \sum_{i=1}^{n}\delta_i$. Similar to the previous case, using $\lambda_i = \exp\left(\boldsymbol{\beta}^\top \mathbf{x}_i\right)$, the log-likelihood function can be written as

$$l(\alpha,\boldsymbol{\beta}|D) = d\log\alpha + \sum_{i=1}^{n}\left(\delta_i(\alpha-1)\log y_i - \delta_i\boldsymbol{\beta}^\top\mathbf{x}_i - \frac{y_i^\alpha}{\exp\left(\boldsymbol{\beta}^\top\mathbf{x}_i\right)}\right). \tag{26}$$

Similar to Equations (15) and (16), conditional posterior corresponding to the Gaussian and Laplacian priors can be written as

$$l_{\mathrm{G}}(\alpha,\boldsymbol{\beta}) = l(\alpha,\boldsymbol{\beta}|D) + \log(\pi(\boldsymbol{\beta}|\tau_1,\ldots,\tau_p)) = l(\alpha,\boldsymbol{\beta}|D) - \sum_{j=1}^{p}\left(\log\sqrt{\tau_j} + \frac{1}{2}\log 2\pi + \frac{\beta_j^2}{2\tau_j}\right), \tag{27}$$

$$l_{\mathrm{L}}(\alpha,\boldsymbol{\beta}) = l(\alpha,\boldsymbol{\beta}|D) + \log(\pi(\boldsymbol{\beta}|\gamma_1,\ldots,\gamma_p)) = l(\alpha,\boldsymbol{\beta}|D) - \sum_{j=1}^{p}(\log 2 - \log\sqrt{\gamma_j} + \sqrt{\gamma_j}|\beta_j|). \tag{28}$$

With the CLG algorithm, the one-dimensional problem involves finding $\alpha_j^{(\mathrm{new})}$ and $\beta_j^{(\mathrm{new})}$, the value of the jth entry of $\boldsymbol{\beta}$ that gives the minimum value of $-l(\alpha,\boldsymbol{\beta})$, assuming that the other α_j's and β_j's are held at their current values. Therefore, for both Gaussian and Laplacian priors, finding $\alpha^{(\mathrm{new})}$ is equivalent to finding the z that minimizes

$$g(z) = -d\log z - \sum_{i=1}^{n}\left(\delta_i(z-1)\log y_i - y_i^z\exp\left(-\boldsymbol{\beta}^\top\mathbf{x}_i\right)\right). \tag{29}$$

By writing the Taylor series expansion for $g(z)$ around z,

$$g(z) \approx g(\alpha) + g'(\alpha)(z-\alpha) + \frac{1}{2}g''(\alpha)(z-\alpha)^2, \tag{30}$$

where

$$g'(\alpha) = \left.\frac{\mathrm{d}g(z)}{\mathrm{d}z}\right|_{z=\alpha} = -d/\alpha - \sum_{i=1}^{n}\left(\delta_i\log y_i - y_i^\alpha\exp\left(-\boldsymbol{\beta}^\top\mathbf{x}_i\right)\log y_i\right), \tag{31}$$

$$g''(\alpha) = \left.\frac{\mathrm{d}^2 g(z)}{\mathrm{d}z^2}\right|_{z=\alpha} = d/\alpha^2 + \sum_{i=1}^{n} y_i^\alpha(\log y_i)^2\exp\left(-\boldsymbol{\beta}^\top\mathbf{x}_i\right). \tag{32}$$

The value of $\alpha_j^{(\text{new})}$ for both types of priors can then be computed as

$$\alpha^{(\text{new})} = \alpha + \Delta\alpha = \alpha - \frac{g'(\alpha)}{g''(\alpha)}. \tag{33}$$

The stepwise update for $\boldsymbol{\beta}$ for both Gaussian and Laplacian priors are similar to that of the exponential model and can be obtained by replacing y_i with y_i^{α} in Equations (18)−(24).

A.3. Log-logistic

For the log-logistic model, the density and survival functions are given by

$$f(y|\theta) = f(y|\lambda,\alpha) = \frac{\alpha y^{\alpha-1}}{\lambda\left(1+\frac{y^{\alpha}}{\lambda}\right)^2} \quad \text{and} \quad S(y|\theta) = S(y|\lambda,\alpha) = \frac{1}{\left(1+\frac{y^{\alpha}}{\lambda}\right)}, \tag{34}$$

The corresponding likelihood function can be written as

$$L(\lambda_1,\ldots,\lambda_n,\alpha|D) = \prod_{i=1}^{n} f(y_i|\lambda_i,\alpha)^{\delta_i} S(y_i|\lambda_i,\alpha)^{(1-\delta_i)} = \alpha^d \prod_{i=1}^{n}\left(\frac{y_i^{\alpha-1}}{\lambda_i\left(1+\frac{y_i^{\alpha}}{\lambda_i}\right)^2}\right)^{\delta_i}\left(\frac{1}{\left(1+\frac{y_i^{\alpha}}{\lambda_i}\right)}\right)^{(1-\delta_i)}. \tag{35}$$

where $d = \sum_{i=1}^{n}\delta_i$. By reparameterizing, $\lambda_i = \exp\left(\boldsymbol{\beta}^{\top}\mathbf{x}_i\right)$, the log-likelihood function can be written as

$$l(\alpha,\boldsymbol{\beta}|D) = d\log\alpha + \sum_{i=1}^{n}\left(\delta_i(\alpha-1)\log y_i - \delta_i\boldsymbol{\beta}^{\top}\mathbf{x}_i - (1+\delta_i)\log\left(1+y_i^{\alpha}\exp\left(-\boldsymbol{\beta}^{\top}\mathbf{x}_i\right)\right)\right). \tag{36}$$

The posterior distributions corresponding to the Gaussian and Laplacian priors are similar to those for the Weibull model in (27) and (28). With the CLG algorithm, for both types of priors, finding the update $\alpha^{(\text{new})}$ is equivalent to finding the z that minimizes

$$g(z) = -d\log z - \sum_{i=1}^{n}\left(\delta_i(z-1)\log y_i - (1+\delta_i)\log\left(1+y_i^{z}\exp\left(-\boldsymbol{\beta}^{\top}\mathbf{x}_i\right)\right)\right). \tag{37}$$

The value of $\alpha_j^{(\text{new})}$ for both types of priors can then be computed using (33), where

$$g'(\alpha) = \left.\frac{\mathrm{d}g(z)}{\mathrm{d}z}\right|_{z=\alpha} = -d/\alpha - \sum_{i=1}^{n}\left(\delta_i\log y_i - \frac{(1+\delta_i)y_i^{\alpha}\exp\left(-\boldsymbol{\beta}^{\top}\mathbf{x}_i\right)\log y_i}{1+y_i^{\alpha}\exp\left(-\boldsymbol{\beta}^{\top}\mathbf{x}_i\right)}\right), \tag{38}$$

$$g''(\alpha) = \left.\frac{\mathrm{d}^2g(z)}{\mathrm{d}z^2}\right|_{z=\alpha} = d/\alpha^2 + \sum_{i=1}^{n}\frac{(1+\delta_i)y_i^{\alpha}\exp\left(-\boldsymbol{\beta}^{\top}\mathbf{x}_i\right)(\log y_i)^2}{\left(1+y_i^{\alpha}\exp\left(-\boldsymbol{\beta}^{\top}\mathbf{x}_i\right)\right)^2}. \tag{39}$$

For the Gaussian prior, finding $\beta_j^{(\text{new})}$ is equivalent to finding the z that minimizes

$$g_{\mathrm{G}}(z) = z\sum_{i=1}^{n}x_{ij}\delta_i + \sum_{i=1}^{n}(1+\delta_i)\log\left(1+y_i^{\alpha}\exp\left(-zx_{ij} - \sum_{\substack{k=1\\k\neq j}}^{p}\beta_k x_{ik}\right)\right) + \frac{z^2}{2\tau_j}. \tag{40}$$

Similar to the previous cases, the updated value $\beta_j^{(\text{new})}$ can be computed using (24), where

$$g'_{\mathrm{G}}(\beta_j) = \left.\frac{\mathrm{d}g_{\mathrm{G}}(z)}{\mathrm{d}z}\right|_{z=\beta_j} = \sum_{i=1}^{n}x_{ij}\delta_i - \sum_{i=1}^{n}\frac{(1+\delta_i)y_i^{\alpha}x_{ij}\exp\left(-\boldsymbol{\beta}^{\top}\mathbf{x}_i\right)}{1+y_i^{\alpha}\exp\left(-\boldsymbol{\beta}^{\top}\mathbf{x}_i\right)} + \frac{\beta_j}{\tau_j}, \tag{41}$$

$$g''_{\mathrm{G}}(\beta_j) = \left.\frac{\mathrm{d}^2 g_{\mathrm{G}}(z)}{\mathrm{d}z^2}\right|_{z=\beta_j} = \sum_{i=1}^{n} \frac{(1+\delta_i) y_i^{\alpha} x_{ij}^2 \exp\left(-\boldsymbol{\beta}^{\top}\mathbf{x}_i\right)}{\left(1+y_i^{\alpha}\exp\left(-\boldsymbol{\beta}^{\top}\mathbf{x}_i\right)\right)^2} + \frac{1}{\tau_j}. \tag{42}$$

Similarly, for the Laplacian prior,

$$g_{\mathrm{L}}(z) = z\sum_{i=1}^{n} x_{ij}\delta_i + \sum_{i=1}^{n}(1+\delta_i)\log\left(1+y_i^{\alpha}\exp\left(-zx_{ij} - \sum_{\substack{k=1\\k\neq j}}^{p}\beta_k x_{ik}\right)\right) + \sqrt{\gamma_j}|z|, \tag{43}$$

$$g'_{\mathrm{L}}(\beta_j) = \left.\frac{\mathrm{d}g_{\mathrm{L}}(z)}{\mathrm{d}z}\right|_{z=\beta_j} = \sum_{i=1}^{n} x_{ij}\delta_i - \sum_{i=1}^{n}\frac{(1+\delta_i) y_i^{\alpha} x_{ij}\exp\left(-\boldsymbol{\beta}^{\top}\mathbf{x}_i\right)}{1+y_i^{\alpha}\exp\left(-\boldsymbol{\beta}^{\top}\mathbf{x}_i\right)} + \sqrt{\gamma_j}\,\mathrm{sign}(\beta_j), \quad \beta_j \neq 0, \tag{44}$$

$$g''_{\mathrm{L}}(\beta_j) = \left.\frac{\mathrm{d}^2 g_{\mathrm{L}}(z)}{\mathrm{d}z^2}\right|_{z=\beta_j} = \sum_{i=1}^{n} \frac{(1+\delta_i) y_i^{\alpha} x_{ij}^2 \exp\left(-\boldsymbol{\beta}^{\top}\mathbf{x}_i\right)}{\left(1+y_i^{\alpha}\exp\left(-\boldsymbol{\beta}^{\top}\mathbf{x}_i\right)\right)^2}, \quad \beta_j \neq 0. \tag{45}$$

A.4. Lognormal

Assuming that the survival times $y_i,\ i = 1,\ldots,n$, follow a lognormal distribution is equivalent to assuming that their logarithms $w_i = \log y_i,\ i = 1,\ldots,n$, follow a normal $\mathrm{N}(\mu,\sigma^2)$ distribution with density and survival functions given by $f(y|\theta) = f(w|\mu,\sigma) = (1/\sqrt{2\pi\sigma^2})\exp\left(-(w-\mu)^2/2\sigma^2\right)$ and $S(y|\theta) = S(w|\mu,\sigma) = 1-\Phi((w-\mu)/\sigma)$, respectively, where $\Phi(\cdot)$ is the Gaussian cumulative distribution function. Following the convention, we replace λ and α with μ and σ, respectively. The likelihood function of $\mu_i, \mu_j, \ldots, \mu_n$ and σ can be written as

$$\begin{aligned} L(\mu_1,\ldots,\mu_n,\sigma|D) &= \prod_{i=1}^{n} f(y_i|\mu_i,\sigma)^{\delta_i} S(y_i|\mu_i,\sigma)^{(1-\delta_i)} \\ &= (2\pi\sigma^2)^{-d/2}\prod_{i=1}^{n}\left(\exp\left(-\frac{(w-\mu_i)^2}{2\sigma^2}\right)\right)^{\delta_i}\left(1-\Phi\left(\frac{w-\mu_i}{\sigma}\right)\right)^{(1-\delta_i)}, \end{aligned} \tag{46}$$

where $d = \sum_{i=1}^{n}\delta_i$. By reparameterizing $\mu_i = \boldsymbol{\beta}^{\top}\mathbf{x}_i$, the log-likelihood function can be written as

$$l(\sigma,\boldsymbol{\beta}|D) = \log 2\pi - d\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\delta_i\left(w_i - \boldsymbol{\beta}^{\top}\mathbf{x}_i\right)^2 + \sum_{i=1}^{n}(1-\delta_i)\log\left(1-\Phi\left(\frac{w-\boldsymbol{\beta}^{\top}\mathbf{x}_i}{\sigma}\right)\right). \tag{47}$$

The conditional posterior densities corresponding to the Gaussian and Laplacian priors can be written as

$$l_{\mathrm{G}}(\sigma,\boldsymbol{\beta}) = l(\sigma,\boldsymbol{\beta}|D) + \log(\pi(\boldsymbol{\beta}|\tau_1,\ldots,\tau_p)) = l(\sigma,\boldsymbol{\beta}|D) - \sum_{j=1}^{p}\left(\log\sqrt{\tau_j} + \frac{1}{2}\log 2\pi + \frac{\beta_j^2}{2\tau_j}\right), \tag{48}$$

$$l_{\mathrm{L}}(\sigma,\boldsymbol{\beta}) = l(\sigma,\boldsymbol{\beta}|D) + \log(\pi(\boldsymbol{\beta}|\gamma_1,\ldots,\gamma_p)) = l(\sigma,\boldsymbol{\beta}|D) - \sum_{j=1}^{p}(\log 2 - \log\sqrt{\gamma_j} + \sqrt{\gamma_j}|\beta_j|). \tag{49}$$

Assuming that the other parameters are held at their current values, the one-dimensional problems involve finding $\sigma^{(\mathrm{new})}$ and $\beta_j^{(\mathrm{new})}$ that minimize the posterior. With the CLG algorithm for both types of priors, finding $\sigma^{(\mathrm{new})}$ is equivalent to finding the z that minimizes

$$g(z) = d\log\sigma + \frac{1}{2\sigma^2}\sum_{i=1}^{n}\delta_i\left(w_i - \boldsymbol{\beta}^{\top}\mathbf{x}_i\right)^2 - \sum_{i=1}^{n}(1-\delta_i)\log\left(1-\Phi\left(\frac{w-\boldsymbol{\beta}^{\top}\mathbf{x}_i}{\sigma}\right)\right). \tag{50}$$

The value of $\sigma^{(\text{new})}$ can then be computed as

$$\sigma^{(\text{new})} = \sigma + \Delta\sigma = \sigma - \frac{g'(\sigma)}{g''(\sigma)}. \tag{51}$$

Here,

$$g'(\sigma) = \left.\frac{\mathrm{d}g(z)}{\mathrm{d}z}\right|_{z=\sigma} = \frac{d}{\sigma} - \frac{1}{\sigma}\sum_{i=1}^{n}\delta_i p_i^2 - \frac{1}{\sigma}\sum_{i=1}^{n}(1-\delta_i)p_i q_i, \tag{52}$$

$$g''(\sigma) = \left.\frac{\mathrm{d}^2 g(z)}{\mathrm{d}z^2}\right|_{z=\sigma} = -\frac{d}{\sigma^2} + \frac{3}{\sigma^2}\sum_{i=1}^{n}\delta_i p_i^2 - \frac{1}{\sigma^2}\sum_{i=1}^{n}(1-\delta_i)\left(p_i^3 q_i - p_i^2 q_i^2 - 2p_i q_i\right), \tag{53}$$

where

$$p_i = \frac{w_i - \boldsymbol{\beta}^\top \mathbf{x}_i}{\sigma} \quad \text{and} \quad q_i = \frac{\mathrm{N}\left(\left(w_i - \boldsymbol{\beta}^\top \mathbf{x}_i\right)/\sigma\right)}{1 - \Phi\left(\left(w - \boldsymbol{\beta}^\top \mathbf{x}_i\right)/\sigma\right)}. \tag{54}$$

For the Gaussian prior, finding $\beta_j^{(\text{new})}$ is equivalent to finding the z that minimizes

$$g_{\mathrm{G}}(z) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}\delta_i(r_i - x_{ij}z)^2 - \sum_{i=1}^{n}(1-\delta_i)\log\left(1 - \Phi\left(\frac{r_i - x_{ij}z}{\sigma}\right)\right) + \frac{z^2}{2\tau_j}, \tag{55}$$

where $r_i = w_i - \sum_{\substack{k=1\\k\neq j}}^{p}\beta_k x_{ik}$. The updated value $\beta_j^{(\text{new})}$ can be computed using (24), where

$$g'_{\mathrm{G}}(\beta_j) = \left.\frac{\mathrm{d}g_{\mathrm{G}}(z)}{\mathrm{d}z}\right|_{z=\beta_j} = -\frac{1}{\sigma}\sum_{i=1}^{n}x_{ij}(\delta_i p_i + (1-\delta_i)q_i) + \frac{\beta_j}{\tau_j}, \tag{56}$$

$$g''_{\mathrm{G}}(\beta_j) = \left.\frac{\mathrm{d}^2 g_{\mathrm{G}}(z)}{\mathrm{d}z^2}\right|_{z=\beta_j} = \frac{1}{\sigma^2}\sum_{i=1}^{n}x_{ij}(\delta_i - q_i(1-\delta_i)(p_i - q_i)) + \frac{1}{\tau_j}, \tag{57}$$

and p_i and q_i are same as (54). Similarly, for the Laplacian prior,

$$g_{\mathrm{L}}(z) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}\delta_i(r_i - x_{ij}z)^2 - \sum_{i=1}^{n}(1-\delta_i)\log\left(1 - \Phi\left(\frac{r_i - x_{ij}z}{\sigma}\right)\right) + \sqrt{\gamma_j}|z|, \tag{58}$$

$$g'_{\mathrm{L}}(\beta_j) = \left.\frac{\mathrm{d}g_{\mathrm{L}}(z)}{\mathrm{d}z}\right|_{z=\beta_j} = -\frac{1}{\sigma}\sum_{i=1}^{n}x_{ij}(\delta_i p_i + (1-\delta_i)q_i) + \sqrt{\gamma_j}\,\mathrm{sign}(\beta_j), \quad \beta_j \neq 0, \tag{59}$$

$$g''_{\mathrm{L}}(\beta_j) = \left.\frac{\mathrm{d}^2 g_{\mathrm{L}}(z)}{\mathrm{d}z^2}\right|_{z=\beta_j} = \frac{1}{\sigma^2}\sum_{i=1}^{n}x_{ij}(\delta_i - q_i(1-\delta_i)(p_i - q_i)), \quad \beta_j \neq 0. \tag{60}$$

Acknowledgements

This research was supported by an NIH-NIGMS grant awarded to the Children's National Medical Center (R01GM087600-01).

References

1. Oakes D. Biometrika centenary: survival analysis. *Biometrika* 2001; **88**(1):99–142.
2. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. John Wiley and Sons: New York, 1980.
3. Box-Steffensmeier JM, Jones BS. *Event History Modeling: A Guide for Social Scientists*. Cambridge University Press: Cambridge, UK, 2004.
4. Collett D. *Modelling Survival Data for Medical Research*, 2nd edn. Chapman-Hall: London, UK, 2003.

5. Heckman JJ, Singer B. *Longitudinal Analysis of Labor Market Data*. Cambridge University Press: Cambridge, UK, 1985.
6. Hosmer DW, Lemeshow S, May S. *Applied Survival Analysis: Regression Modeling of Time to Event Data*, 2nd edn, Wiley Series in Probability and Statistics. Wiley-Interscience: New York, NY, 2008.
7. Tibshirani R. The lasso method for variable selection in the Cox model. *Statistics in Medicine* 1997; **16**(4):385–395.
8. Klein JP, Moeschberger ML. *Survival Analysis: Techniques for Censored and Truncated Data*, 2nd edn. John Wiley and Sons: New York, 2003.
9. Lee ET, Wang J. *Statistical Methods for survival Data Analysis*, 3rd edn, Wiley Series in Probability and Statistics. Wiley-Interscience: New York, NY, 2003.
10. Ibrahim JG, Chen M-H, Sinha D. *Bayesian Survival Analysis*. Springer-Verlag: New York, NY, 2001.
11. Tseng P. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications* 2001; **109**(3):475–494.
12. Koh K, Kim S-J, Boyd S. An interior-point method for large-scale L1-regularized logistic regression. *Journal of Machine Learning Research* 2007; **8**:1519–1555.
13. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 2010; **33**(1):1–22.
14. Genkin A, Lewis D, Madigan D. Large-scale Bayesian logistic regression for text categorization. *Technometrics* 2007; **49**(3):291–304.
15. Lawless JF. *Statistical Models and Methods for Lifetime Data*, 2nd edn, Wiley Series in Probability and Statistics. Wiley-Interscience: New York, NY, 2003.
16. Evers L, Messow C-M. Sparse kernel methods for high-dimensional survival data. *Bioinformatics* 2008; **24**(14):1632–1638.
17. Shivaswamy PK, Chu W, Jansche M. A support vector approach to censored targets. *IEEE International Conference on Data Mining*, Omaha, NE, 2007; 655–660.
18. Van Belle V, Pelckmans K, Van Huffel S, Suykens JAK. Improved performance on high-dimensional survival data by application of survival-SVM. *Bioinformatics* 2011; **27**(1):87–94.
19. Ishwaran H, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS. High-dimensional variable selection for survival data. *Journal of the American Statistical Association* 2010; **105**(489):205–217.
20. Ishwaran H, Kogalur UB, Chen X, Minn AJ. Random survival forests for high-dimensional data. *Statistical Analysis and Data Mining* 2011; **4**(1):115–132.
21. Lisboa PJG, Etchells TA, Jarman IH, Arsene CTC, Aung MSH, Eleuteri A, Taktak AFG, Ambrogi F, Boracchi P, Biganzoli E. Partial logistic artificial neural network for competing risks regularized with automatic relevance determination. *IEEE Transactions on Neural Networks* 2009; **20**(9):1403–1416.
22. Engler D, Li Y. Survival analysis with high-dimensional covariates: an application in microarray studies. *Statistical Applications in Genetics and Molecular Biology* 2009; **8**(1):1–22.
23. Gui J, Li H. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics* 2005; **21**(13):3001–3008.
24. Goeman JJ. L1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal* 2010; **52**(1):70–84.
25. Yang Y, Zou H. A cocktail algorithm for solving the elastic net penalized Cox's regression in high dimensions. *Statistics and Its Interface* 2012.
26. Witten DM, Tibshirani R. Survival analysis with high-dimensional covariates. *Statistical Methods in Medical Research* 2010; **19**(1):29–51.
27. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software* 2011; **39**(5):1–13.
28. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 1970; **12**: 55–67.
29. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)* 1996; **58**:267–288.
30. Zhang T, Oles FJ. Text categorization based on regularized linear classification methods. *Information Retrieval* 2000; **4**:5–31.
31. Kivinen J, Warmuth MK. Relative loss bounds for multidimensional regression problems. *Journal of Machine Learning*. MIT Press: Cambridge, MA, 2001; 301–329.
32. National Center for Injury Prevention and Control. *CDC Injury Fact Book, Atlanta (GA): Centers for Disease Control and Prevention*. National Center for Injury Prevention and Control, 2006. http://www.cdc.gov/Injury/publications/FactBook/InjuryBook2006.pdf.
33. Mackersie RC. History of trauma field triage development and the american college of surgeons criteria. *Prehospital Emergency Care* 2006; **10**(3):287–294.
34. Committee on Trauma ACoS. *Resources for optimal care of the injured patient*, 2006. https://web4.facs.org/ebusiness/ProductCatalog/product.aspx?ID=194.
35. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002; **415**(6871):530–536.
36. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine* 2002; **347**(25):1999–2009.
37. Park MY, Hastie T. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2007; **69**(4):659–677.

38. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *The Journal of the American Medical Association* 1982; **247**(18):2543–2546.
39. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 1996; **15**(4):361–387.
40. Chambless LE, Cummiskey CP, Cui G. Several methods to assess improvement in risk prediction models: extension to survival analysis. *Statistics in Medicine* 2011; **30**(1):22–38.
41. Grønnesby JK, Borgan Ø. A method for checking regression models in survival analysis based on the risk score. *Lifetime Data Analysis* 1996; **2**(4):315–328.
42. May S, Hosmer DW. A simplified method of calculating an overall goodness-of-fit test for the Cox proportional hazards model. *Lifetime Data Analysis* 1998; **4**(2):109–120.
43. Suchard M, Simpson SE, Zorych I, Ryan P, Madigan D. Massive parallelization of serial inference algorithms for a complex generalized linear model. *ACM Transactions on Modeling and Computer Simulation* 2013; **23**(1):10:1–10:17.