Biometrics 65, 394–404
June 2009

DOI: 10.1111/j.1541-0420.2008.01074.x

Regularized Estimation for the Accelerated Failure Time Model

T. Cai,[1,*] J. Huang,[2,] and L. Tian[2,***]**

[1]Department of Biostatistics, Harvard University, Boston, Massachusetts 02115, U.S.A.
[2]Department of Preventive Medicine, Northwestern University, Chicago, Illinois 60611, U.S.A.
[*]*email:* tcai@hsph.harvard.edu
[**]*email:* jjhuang@northwestern.edu
[***]*email:* lutian@northwestern.edu

Summary. In the presence of high-dimensional predictors, it is challenging to develop reliable regression models that can be used to accurately predict future outcomes. Further complications arise when the outcome of interest is an event time, which is often not fully observed due to censoring. In this article, we develop robust prediction models for event time outcomes by regularizing the Gehan's estimator for the accelerated failure time (AFT) model (Tsiatis, 1996, *Annals of Statistics* **18,** 305–328) with least absolute shrinkage and selection operator (LASSO) penalty. Unlike existing methods based on the inverse probability weighting and the Buckley and James estimator (Buckley and James, 1979, *Biometrika* **66,** 429–436), the proposed approach does not require additional assumptions about the censoring and always yields a solution that is convergent. Furthermore, the proposed estimator leads to a stable regression model for prediction even if the AFT model fails to hold. To facilitate the adaptive selection of the tuning parameter, we detail an efficient numerical algorithm for obtaining the entire regularization path. The proposed procedures are applied to a breast cancer dataset to derive a reliable regression model for predicting patient survival based on a set of clinical prognostic factors and gene signatures. Finite sample performances of the procedures are evaluated through a simulation study.

Key words: AFT model; LASSO regularization; Linear programming.

1. Introduction

Global gene expression profiling using microarrays has the potential to lead to a better understanding of the molecular features corresponding to different phenotypic disease outcomes. Prediction of disease outcomes using genomic markers is challenging because the number of covariates could be large relative to the sample size. To incorporate such high-dimensional data, dimension reduction methods such as principal component regression (Jolliffe, 1986) and partial least square methods (Martens and Naes, 1989) have been proposed. These methods essentially search for low-dimensional projections of the covariates to optimize the tradeoff between bias and variance and thus achieve reduced mean squared errors (MSEs; Park, 1981). An alternative approach to handle high-dimensional predictors is through penalized estimation. Penalized methods have been studied extensively in the literature for noncensored outcomes. Examples include the ridge regression (Hoerl and Kennard, 1970), the support vector machines (Vapnik, 1995), the LASSO (Tibshirani, 1996), the gradient directed regularization method (Friedman and Popescu, 2004), and the adaptive LASSO (Zou, 2006; Zhang and Lu, 2007). These methods modify ("regularize") the minimization of a usual empirical risk function $\hat{R}_n(\boldsymbol{\theta})$ by adding a penalty $\lambda P(\boldsymbol{\theta})$ to the risk and thus instead minimize

$$\hat{R}_n(\boldsymbol{\theta}) + \lambda P(\boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$ is the unknown parameter associated with marker effects and $\lambda \geqslant 0$ is the penalty parameter that controls the degree of regularization. Larger values of λ provide increased regularization producing more stable estimates (Friedman and Popescu, 2004). For any given λ, one may obtain a regularized estimator for $\boldsymbol{\theta}$. Typically, both the empirical risk function $\hat{R}_n(\boldsymbol{\theta})$ and the penalty function $P(\boldsymbol{\theta})$ are convex and thus the penalized risk function remains convex, which ensures the existence and uniqueness of the minimizer for any given λ. The LASSO-type regression procedures based on the L_1 penalty have become useful tools to incorporate high-dimensional data because these methods achieve the shrinkage and variable selection simultaneously by producing sparse solutions. Another attractive feature of these procedures is that efficient numerical algorithms such as the gradient LASSO (Kim and Kim, 2004) and least angle regression (LARS) (Efron et al., 2004) have become available for implementation.

Regularized methods for combining high-dimensional markers to predict failure time outcomes are less well developed. Tibshirani (1997) and Gui and Li (2005) developed regularized Cox regression methods by adding an L_1 LASSO penalty to the partial likelihood. A more comprehensive review of related literatures can be found in Li (2008). However, the proportional hazards assumption may not be appropriate for certain applications. A useful alternative to the Cox model is the accelerated failure time (AFT) model (Wei,

1992), which has been studied extensively in recent years for the standard regression setting. Inference procedures for the regression parameters under the AFT model include the inverse probability weighting (IPW) method, Buckley–James iterative method, and rank-based method (Buckley and James, 1979; Koul, Susarla, and Van Ryzin, 1981; Tsiatis, 1996). To incorporate high-dimensional covariates, the LASSO regularization has been applied to the IPW and Buckley–James (BJ) estimators for the AFT model (Huang, Ma, and Xie, 2006; Datta, Le-Rademacher, and Datta, 2007; Wang et al., 2008). The LASSO regularized IPW estimator inherits most of the good properties of the common LASSO regularization in linear models. However, the validity of the IPW approach relies on the correct specification of the conditional censoring distribution, which may be difficult in practice. Furthermore, it requires the support of the censoring time to contain the entire support the failure time. This assumption is likely to be violated in practice, because most of the clinical studies have prespecified duration of follow-up, which may not be sufficient to observe all failures. The BJ procedure needs less stringent conditions on the censoring time. However, it also relies on the identifiability of the entire residual distribution, which may not be available in the presence of censoring. Furthermore, the LASSO-regularized BJ estimator cannot be interpreted as a constrained minimizer of a convex objective function, because it is based on an estimating equation motivated from the self-consistency principle. Lastly, the computation of the BJ estimator is based on an unstable iterative algorithm, which may lead to multiple limiting values (Currie, 1996; Huang and Harrington, 2005; Huang et al., 2006). To overcome such difficulties, we propose to regularize the rank-based estimation procedure with LASSO-type penalty to develop parsimonious prediction models for failure time outcomes. The proposed regularization methods have several advantages. First, we only require the censoring to be independent of the event time conditional on the covariates without any additional specification of the censoring distribution. Furthermore, the resulting estimator is well defined for any given penalty parameter and can be viewed as a solution to a linear programming optimization problem. In fact, the exact LASSO path can be easily identified through an efficient numerical algorithm detailed in Web Appendix C. Another advantage of the proposed method lies in its robustness with respect to model misspecification. When the AFT model is correctly specified, our proposed estimator provides a consistent estimator for the regression coefficients and thus would be the optimal prediction rules asymptotically. On the other hand, when the AFT model is only an approximation to the true model, the convexity of the proposed objective function would ensure the convergence of the estimation procedure and thus lead to a stable regression model for prediction.

The rest of the article is organized as follows. In Section 2, we detail the estimation method and procedures for making inference about the estimator. We discuss the selection of the regularization parameter in Section 3. Procedures for computing the exact LASSO path are shown in Section 4. We illustrate the proposed procedure with the data from a breast cancer study in Section 5. Simulation results are summarized in Section 6 and some final remarks are provided in Section 7. The detailed algorithm for computing the exact LASSO path is given in Web Appendix C.

2. Method

Let $\{(T_i, C_i, \mathbf{Z}_i), i = 1, \ldots, n,\}$ be n independent and identically distributed random vectors, where T and C are the log-transformed survival and censoring times, respectively, and $\mathbf{Z}$ is the p-dimensional covariate vector. Due to censoring, the observed data consists of $\{(Y_i, \delta_i, \mathbf{Z}_i), i = 1, \ldots, n\}$, where $Y_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leqslant C_i)$. We assume that the distribution of T given $\mathbf{Z}$ follows an AFT model:

$$T_i = \boldsymbol{\beta}_0' \mathbf{Z}_i + \epsilon_i,$$

where $\{\epsilon_i, i = 1, \ldots, n\}$ are n independent random errors with an unspecified common distribution function.

2.1 *Estimation Procedure*

To estimate the covariate effect $\boldsymbol{\beta}_0$, Tsiatis (1996) proposed a set of estimating equations motivated from inverting a class of rank tests. Specifically, a consistent estimator of $\boldsymbol{\beta}_0$ can be obtained as the solution to the estimating equation

$$n^{-1} \sum_{i=1}^{n} W_n\{e_i(\boldsymbol{\beta}); \boldsymbol{\beta}\} \left[\mathbf{Z}_i - \frac{\sum_{j=1}^{n} \mathbf{Z}_j I\{e_j(\boldsymbol{\beta}) \geqslant e_i(\boldsymbol{\beta})\}}{\sum_{j=1}^{n} I\{e_j(\boldsymbol{\beta}) \geqslant e_i(\boldsymbol{\beta})\}} \right] \delta_i = o_p(n^{-1/2}), \tag{1}$$

where $e_i(\boldsymbol{\beta}) = Y_i - \boldsymbol{\beta}' \mathbf{Z}_i$ and $W_n(\cdot; \cdot)$ is a known weight function convergent to a deterministic limit. A common choice of $W_n(\cdot; \cdot)$ is the Gehan's weight function $W_n(t; \boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} I\{e_i(\boldsymbol{\beta}) \geqslant t\}$. This corresponds to a simple monotone estimating function,

$$\mathbf{S}(\boldsymbol{\beta}) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\mathbf{Z}_j - \mathbf{Z}_i) I\{e_i(\boldsymbol{\beta}) \geqslant e_j(\boldsymbol{\beta})\} \delta_j,$$

which is the "quasiderivative" of a convex objective function,

$$L(\boldsymbol{\beta}) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \{e_i(\boldsymbol{\beta}) - e_j(\boldsymbol{\beta})\} I\{e_i(\boldsymbol{\beta}) \geqslant e_j(\boldsymbol{\beta})\} \delta_j.$$

Therefore, a valid estimator of $\boldsymbol{\beta}_0, \hat{\boldsymbol{\beta}}_G$ may be obtained as a minimizer of the convex function $L(\boldsymbol{\beta})$ (Ritov, 1990). In practice, $\hat{\boldsymbol{\beta}}_G$ may be obtained through linear programming (Jin, Lin, and Wei, 2003) by minimizing $\sum_{i=1}^{n} \sum_{j=1}^{n} e_{ij}^{+} \delta_j$ subject to the constraints

$$(Y_i - Y_j) - \boldsymbol{\beta}'(\mathbf{Z}_i - \mathbf{Z}_j) = e_{ij}^{+} - e_{ij}^{-}, \quad e_{ij}^{+} \geqslant 0, e_{ij}^{-} \geqslant 0, \quad \text{for } 1 \leqslant i, j \leqslant n. \tag{2}$$

When the sample size is not large relative to the dimension of $\boldsymbol{\beta}_0$, the conventional estimator $\hat{\boldsymbol{\beta}}_G$ may have poor performance (Huang et al., 2006). To obtain a more accurate estimator for $\boldsymbol{\beta}_0$ in such settings, we propose the LASSO-regularized Gehan's estimator,

$$\hat{\boldsymbol{\beta}}_{LG} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ L(\boldsymbol{\beta}) + \lambda_n \sum_{k=1}^{p} |\beta_k| \right\},$$

where $\lambda_n > 0$ is the penalty parameter and for any vector $\boldsymbol{a}$, we use notation a_k to denote the kth component of

$\boldsymbol{a}$. Equivalently, $\hat{\boldsymbol{\beta}}_{LG}$ is the minimizer of $L(\boldsymbol{\beta})$ under the constraint

$$\sum_{k=1}^{p} |\beta_k| \leqslant s_n,$$

for some s_n that corresponds to the penalty parameter λ_n. As for $\hat{\boldsymbol{\beta}}_G$, the computation of the regularized estimator $\hat{\boldsymbol{\beta}}_{LG}$ can be achieved through linear programming by minimizing $\sum_{i=1}^{n}\sum_{j=1}^{n} e_{ij}^{+}\delta_j$ subject to constraints

$$\beta_k = \beta_k^{+} - \beta_k^{-}, \quad \beta_k^{+} \geqslant 0,\ \beta_k^{-} \geqslant 0, \quad \text{for } k = 1, \ldots, p,$$

$$\text{and} \quad \sum_{k=1}^{p} (\beta_k^{+} + \beta_k^{-}) \leqslant s_n,$$

in addition to the constraints given in equation (2).

The LASSO method penalizes all the regression coefficients in the same way and may not be consistent in identifying all the nonzero regression coefficients. Recently, Zou (2006) and Zhang and Lu (2007) proposed the adaptive LASSO method, which penalizes the regression coefficients according to an initial estimator. The adaptive LASSO approach penalizes the regression coefficients based on the magnitude of their initial estimators. It entails more stringent constraints or equivalently larger penalties on coefficients that are likely to be zero according to the initial estimator. The adaptive LASSO has been used for variable selection in different models (Leng and Ma, 2007; Lu and Zhang, 2007; Wang, Li, and Jiang, 2007) and can be easily applied to regularize the Gehan's estimator. Specifically, one may estimate $\boldsymbol{\beta}_0$ by minimizing $L(\boldsymbol{\beta}) + n^{-\frac{1}{2}}\tilde{\lambda}_n \sum_{k=1}^{p} |\beta_k| / |\hat{\beta}_{Gk}|$ or equivalently minimizing $L(\boldsymbol{\beta})$ under the constraints

$$\sum_{k=1}^{p} \frac{|\beta_k|}{|\hat{\beta}_{Gk}|} \leqslant \tilde{s}_n.$$

Let $\hat{\boldsymbol{\beta}}_{AG}$ denote the corresponding adaptive LASSO regularized Gehan's estimator. In practice, we may obtain $\hat{\boldsymbol{\beta}}_{AG}$ based on the same procedure as for $\hat{\boldsymbol{\beta}}_{LG}$ with the rescaled predictors $\tilde{\mathbf{Z}}_i = (Z_{i1}|\hat{\beta}_{G1}|, \ldots, Z_{ip}|\hat{\beta}_{Gp}|)'$.

2.2 *Large Sample Properties and Inference Procedures*

In Web Appendix A, we show that $\hat{\boldsymbol{\beta}}_{LG} \to \boldsymbol{\beta}_0$ almost surely if $\lambda_n \to 0$. Because $L(\boldsymbol{\beta})$ can be locally approximated by a quadratic function, we show in Web Appendix A that if $n^{1/2}\lambda_n \to \lambda_0 \geqslant 0$, then as $n \to \infty, n^{1/2}(\hat{\boldsymbol{\beta}}_{LG} - \boldsymbol{\beta}_0)$ converges in distribution to $\text{argmin}_{\mathbf{u}} V(\mathbf{u}, \mathbf{W})$, where

$$V(\mathbf{u}, \mathbf{W}) = \mathbf{u}'\mathbf{W} + \frac{1}{2}\mathbf{u}'\mathbb{A}_0\mathbf{u} + \lambda_0 \sum_{j=1}^{p} \{u_j \text{sgn}(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0)\},$$

$\mathbf{W}$ follows a multivariate normal distribution with mean zero and a variance covariance matrix of $\mathbb{B}_0, \mathbb{A}_0 = \partial^2 E\{L(\boldsymbol{\beta})\}/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'$ at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, and $\mathbb{B}_0 = \text{var}\{n^{1/2}\mathbf{S}(\boldsymbol{\beta}_0)\}$.

For the adaptive LASSO estimator, we require $\tilde{\lambda}_n \to 0$ and $n^{\frac{1}{2}}\tilde{\lambda}_n \to \infty$. Under such an assumption and regularity conditions given in Web Appendix A, $\hat{\boldsymbol{\beta}}_{AG} \to \boldsymbol{\beta}_0$ almost surely. Furthermore, in Web Appendix B, we establish its "oracle" property (Zou, 2006; Zhang and Lu, 2007). Specifically, let $\mathcal{A} = \{k : \beta_{0k} \neq 0\}, \hat{\mathcal{A}} = \{k : \hat{\beta}_{AGk} \neq 0\}$, and let $\boldsymbol{\beta}^{\mathcal{A}}$ denote the subvector of $\boldsymbol{\beta}$ that corresponds to $\mathcal{A}$. Using similar arguments as given in Zou (2006), we show that as $n \to \infty, P(\hat{\mathcal{A}} = \mathcal{A}) \to 1$ and $n^{1/2}(\hat{\boldsymbol{\beta}}_{AG}^{\mathcal{A}} - \boldsymbol{\beta}_0^{\mathcal{A}})$ converges in distribution to $N\{0, \mathbb{A}_1^{-1}\mathbb{B}_1(\mathbb{A}_1^{-1})'\}$, where $\mathbb{A}_1 = \partial^2 E\{L(\boldsymbol{\beta})\}/\partial\boldsymbol{\beta}^{\mathcal{A}}\partial\boldsymbol{\beta}^{\mathcal{A}'}$ at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, and $\mathbb{B}_1 = n\text{var}\{\mathbf{S}(\boldsymbol{\beta}_0)^{\mathcal{A}}\}$, where $\mathbf{S}(\boldsymbol{\beta}_0)^{\mathcal{A}}$ is the subvector of $\mathbf{S}(\boldsymbol{\beta}_0)$ corresponds to $\mathcal{A}$.

Although the foregoing theoretical asymptotical properties provide some justifications on the large sample performance of the LASSO and adaptive LASSO regularized estimators in practice, they are not directly applicable to make statistical inference about $\boldsymbol{\beta}_0$ in finite samples. It is generally difficult to approximate the distribution of the LASSO regularized estimators well or to provide desirable interval estimates for $\boldsymbol{\beta}_0$. Routinely used procedures such as the bootstrap may fail in such settings (Knight and Fu, 2000). In practice, we propose to approximate the variance covariance of the proposed estimators based on the local quadratic approximation of the Gehan's objective function (Fan and Li, 2001). Specifically, we suggest estimating the variance of $\hat{\boldsymbol{\beta}}_{LG}$ and $\hat{\boldsymbol{\beta}}_{AG}$ by

$$n^{-1}\{\hat{\mathbb{A}} + \lambda_n \mathbb{D}_{LG}(\hat{\boldsymbol{\beta}}_{LG})\}^{-1}\hat{\mathbb{B}}\{\hat{\mathbb{A}} + \lambda_n \mathbb{D}_{LG}(\hat{\boldsymbol{\beta}}_{LG})\}^{-1} \quad \text{and}$$

$$n^{-1}\{\hat{\mathbb{A}} + n^{-1/2}\tilde{\lambda}_n \mathbb{D}_{AG}(\hat{\boldsymbol{\beta}}_{AG})\}^{-1}\hat{\mathbb{B}}\{\hat{\mathbb{A}} + n^{-1/2}\tilde{\lambda}_n \mathbb{D}_{AG}(\hat{\boldsymbol{\beta}}_{AG})\}^{-1},$$

respectively, where $\hat{\mathbb{A}}$ and $\hat{\mathbb{B}}$ are the respective consistent estimators of $\mathbb{A}_0$ and $\mathbb{B}_0, \mathbb{D}_{LG}(\boldsymbol{\beta}) = \text{diag}\{I(\beta_1 \neq 0)/|\beta_1|, \ldots, I(\beta_p \neq 0)/|\beta_p|\}$ and $\mathbb{D}_{AG}(\boldsymbol{\beta}) = \text{diag}\{I(\beta_1 \neq 0)/\beta_1^2, \ldots, I(\beta_p \neq 0)/\beta_p^2\}$. An extra layer of difficulty arises when obtaining a consistent estimator of $\mathbb{A}_0$ due to the nonsmoothness of the Gehan's objective function. One possible solution is to take advantage of the local linear property of $\mathbf{S}(\boldsymbol{\beta})$,

$$\mathbf{S}(\hat{\boldsymbol{\beta}}^*) - \mathbf{S}(\hat{\boldsymbol{\beta}}) = \mathbb{A}_0(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) + o_p(n^{-1/2}), \quad \text{for } \hat{\boldsymbol{\beta}}^* = \boldsymbol{\beta}_0 + O_p(n^{-1/2}),$$

and approximate the kth column of $\mathbb{A}_0$ based on the estimated regression coefficients $\boldsymbol{a}_k$ in the linear regression model, $S_k(\hat{\boldsymbol{\beta}}^*) - S_k(\hat{\boldsymbol{\beta}}) = \boldsymbol{a}_k'(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) + \epsilon$, where $S_k(\boldsymbol{\beta})$ is the kth component of $\mathbf{S}(\boldsymbol{\beta})$. To estimate the slope for each k, we first obtain B realizations of $\hat{\boldsymbol{\beta}}^*$ by generating B bootstrap samples of $\hat{\boldsymbol{\beta}}_G$, denoted by $\{\hat{\boldsymbol{\beta}}_G^{*(1)}, \ldots, \hat{\boldsymbol{\beta}}_G^{*(B)}\}$; and then fit the linear regression model based on the "responses" $\{S_k(\hat{\boldsymbol{\beta}}_G^{*(1)}) - S_k(\hat{\boldsymbol{\beta}}_G), \ldots, S_k(\hat{\boldsymbol{\beta}}_G^{*(B)}) - S_k(\hat{\boldsymbol{\beta}}_G)\}$ and the corresponding "covariate vectors" $\{\hat{\boldsymbol{\beta}}_G^{*(1)} - \hat{\boldsymbol{\beta}}, \ldots, \hat{\boldsymbol{\beta}}_G^{*(B)} - \hat{\boldsymbol{\beta}}\}$.

3. Selection of the Regularization Parameter

The selection of the regularization parameter is crucial to the performance of the final prediction model. Various methods have been proposed to select the penalty parameter in the LASSO regularization. In general, there are two different objectives one may wish to achieve when selecting the penalty parameter: (i) optimizing the prediction accuracy of the final model; and (ii) identifying the "true" prediction model or, more realistically, the set of nonzero $\boldsymbol{\beta}$ coefficients in the regression model. These two goals are similar, but may be inconsistent and thus require substantially different penalty parameters in the absence of a true model. In this article, we

focus on optimizing the prediction accuracy and suggest using crossvalidation methods for selecting an appropriate penalty parameter.

To optimize the prediction accuracy, one needs to first choose a desirable accuracy measure. A convenient choice is to measure the accuracy by $L(\boldsymbol{\beta})$ itself. Because $E\{L(\boldsymbol{\beta})\} \approx (\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\mathbb{A}_0(\boldsymbol{\beta}-\boldsymbol{\beta}_0)/2$ in a small neighborhood of $\boldsymbol{\beta}_0$, the value of $L(\boldsymbol{\beta})$ can be interpreted as a weighted MSE of using $\boldsymbol{\beta}$ to approximate $\boldsymbol{\beta}_0$. When interest lies in the prediction of t-year survival, one may consider the corresponding c-statistic as a measure of accuracy (Zheng, Cai, and Feng, 2006). For any given accuracy measure, we propose to use the K-fold crossvalidation to select λ that achieves the optimal accuracy. Specifically, we randomly partition the data into K subgroups of approximately equal sizes and for any given λ compute the crossvalidated accuracy based on

$$\mathcal{L}_{cv}(\lambda) = \sum_{k=1}^{K} \mathcal{L}_{(k)}\{\hat{\boldsymbol{\beta}}^{(-k)}(\lambda)\},$$

where $\mathcal{L}_{(k)}(\boldsymbol{\beta})$ is the estimated accuracy function based on the observations in the kth subgroup and $\hat{\boldsymbol{\beta}}^{(-k)}(\lambda)$ is the LASSO or adaptive LASSO regularized Gehan's estimator based on observations not in the kth group with the penalty parameter λ. One may carry out the foregoing K-fold crossvalidation procedure repeatedly and obtain λ_n as the optimizer of the average $\mathcal{L}_{cv}(\lambda)$ among all the replicates.

4. Computation of the Exact Regularization Path

With the linear programming techniques, one may easily obtain $\hat{\boldsymbol{\beta}}_{LG}(s)$, the minimizer of $L(\boldsymbol{\beta})$ under the constraint $\sum_{k=1}^{p}|\beta_k| \leqslant s$, for any given s. However, the data-dependent selection of an optimal penalty λ_n or the corresponding s_n may be time consuming if one directly evaluates the accuracy measure function $\mathcal{L}_{cv}(\lambda)$ via a naive application of the linear programming techniques. To overcome the computational burden, the threshold gradient descending method may be applied to find an approximated solution to the LASSO path (Friedman and Popescu, 2004). However, this approximation may not be accurate in some settings and leads to solutions that are markedly different from the exact LASSO path. In the following section, we propose an efficient numerical algorithm to compute the *exact* entire LASSO regularization path. Because the exact path of $\hat{\boldsymbol{\beta}}_{LG}(s)$ is piecewise linear due to the Karesh–Kuhn–Tucker condition, the entire path is determined by all the changing points $\{\boldsymbol{\beta}^{[1]}, \boldsymbol{\beta}^{[2]}, \ldots\}$ in the path.

To illustrate the algorithm, we consider a simple example by taking the first seven observations from the well-known Mayo Clinic Primary Biliary Cirrhosis study (Fleming and Harrington, 2001) with age and log(albumin) as the only two predictors. With $p = 2$, Figure 1 plots part of the two-dimensional parameter space divided by lines $Y_i - Y_j - \boldsymbol{\beta}'(\mathbf{Z}_i - \mathbf{Z}_j) = 0$, $1 \leqslant i < j \leqslant 7$. Because $L(\boldsymbol{\beta})$ is linear within each segment bounded by straight lines, the fastest descending direction has to be along the edges in the figure. To illustrate how the algorithm explores the regularization path, suppose that $\boldsymbol{\beta}^{[2]}$ and $\boldsymbol{\beta}^{[3]}$ are the points "A" and "B" in the figure, respectively. After reaching point "B," we need to determine the next optimal direction. First, the optimal direction must be

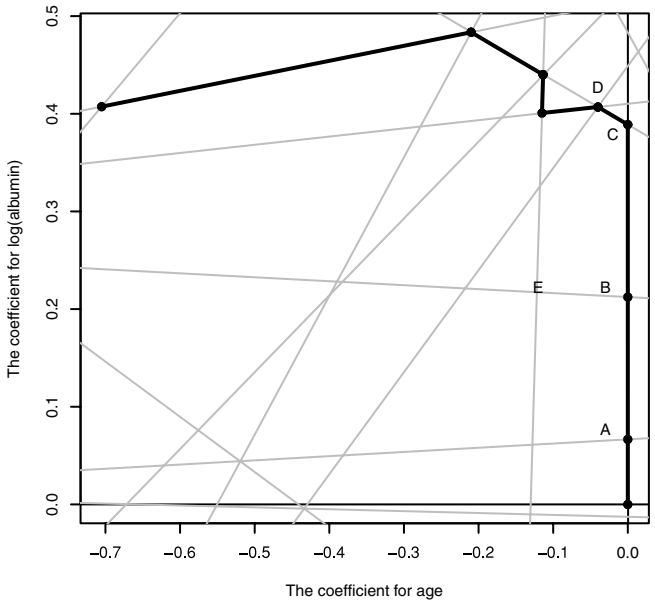


Figure 1. The simple example explaining the algorithm for computing the exact LASSO regularization path.

along the edges "BE" or "BC" due to the convexity and local linearity of the objective function. To decide which direction to take, we need to calculate and compare the descending rate of $L(\boldsymbol{\beta})$ along these two directions relative to the increasing rate of $\|\boldsymbol{\beta}\|$. Specifically, from point "B" to "E," $L(\boldsymbol{\beta})$ decreases from 11.04 to 9.67, while $\|\boldsymbol{\beta}\| = \sum_{k=1}^{2}|\beta_k|$ increases from 0.21 to 0.34. This suggests a relative descending rate of $(9.67 - 11.04)/(0.34 - 0.21) \approx -10.77$. Similarly, from point "B" to "C," maintaining the same direction as from "A" to "B," $L(\boldsymbol{\beta})$ decreases from 11.04 to 8.64, while $\|\boldsymbol{\beta}\|$ increases from 0.21 to 0.39. This implies a faster relative descending rate of -13.73. Therefore, the optimal direction should go from "B" to "C." After reached point "C," we face a similar question: keep the same direction or turn left to the point "D." With the same method, it is straightforward to obtain the relative descending rates along these two directions: the relative descending rates are -8.96 and -11.60 along the y-axis and edge "CD," respectively. Therefore, the joint after "C" is "D." In this manner, one may progressively explore the entire LASSO regularization paths highlighted in the figure. Note that in the foregoing discussion, we purposely ignored the possibility of turning right into the other half of the parameter space for simplicity. The same calculation for the relative descending rate can be easily carried out to confirm that those directions are indeed suboptimal. This algorithm is similar to the procedure given by Zhu et al. (2003) for solving the regularized support vector machine. We present the detailed algorithm for computing the exact path in Web Appendix C.

5. Example

In this example, we are interested in constructing optimal prediction models for survival time using patient level clinical and genetic information based on a breast cancer study. The study involves 295 patients with primary breast carcinomas

from the Netherlands Cancer Institute (Chang et al., 2005). The survival time information was extracted from the medical registry of the Netherlands Cancer Institute. Potential clinical predictors include age, tumor size, lymph node status, tumor grade, vascular invasion status, estrogen receptor status, National Institutes of Health (NIH) risk grade, the use of breast conserving therapy, and the use of adjuvant therapy. Available also are gene signatures that represent distinct analytic strategies and have been validated in independent studies. Specifically, there are seven potential gene signatures: basal-like, ErbB2, luminal A, luminal B, normal-like, a 70-gene, and the wound response gene signatures. The basal-like, ErbB2, luminal A, luminal B, and normal-like gene signatures were identified by an unsupervised clustering method (Perou et al., 2000). The 70-gene signature was constructed based on the association between the gene expression level and the risk of metastasis (van de Vijver et al., 2002). The wound response gene signature was a hypothesis-driven signature proposed by Chang et al. (2005).

To identify the optimal prediction model, we fit the data with the AFT model with the logarithm of the survival time as the response variable and 18 potential predictors: age, tumor size (diameter, cm), the number of lymph nodes, tumor grade (grade 2 versus 1, grade 3 versus 1), vascular invasion (1–3 vessels versus 0 vessel, >3 vessels versus 0 vessel), estrogen receptor status (positive versus negative), NIH risk status (high versus intermediate or low), the use of breast conserving therapy (mastectomy versus breast conserving therapy), the use of adjuvant therapy (no adjuvant therapy versus chemotherapy or hormonal therapy), as well as the seven gene signatures. All the genetic signatures used in the model are continuous correlation measures. In the analysis, to avoid potential biases we excluded a subset of 61 patients, which was used to construct the 70-gene signature. Among the remaining 234 patients, the median follow-up time was 7.2 years and the number of observed deaths is 55.

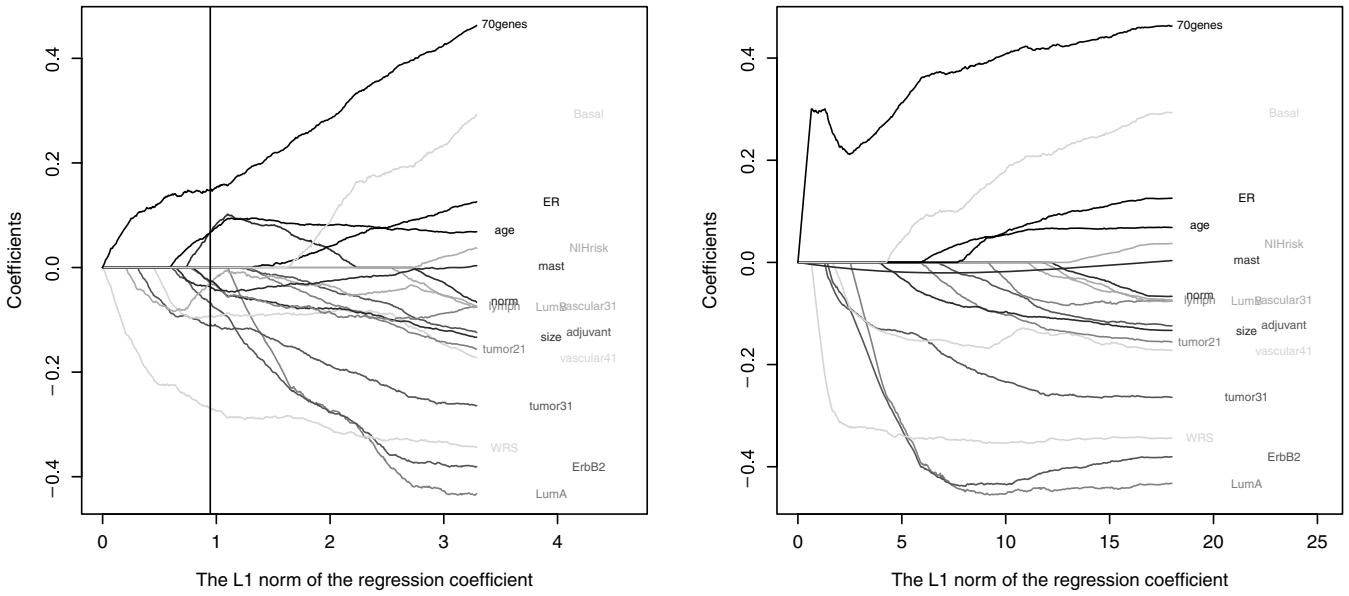
To construct prediction models based on these 18 predictors, we considered three aforementioned estimators for $\boldsymbol{\beta}$: (i) the standard Gehan estimator; (ii) the LASSO estimator; and (iii) the adaptive LASSO estimator. The entire LASSO and adaptive LASSO regularized paths of the proposed estimators are shown in Figure 2a and b. To examine how well the path obtained based on the threshold gradient descending algorithm approximates the exact path, we also obtained the approximated LASSO path in Figure 2c. Although the overall patterns of the two sets of paths are fairly similar as anticipated, there are subtle differences in paths of some individual predictors and in general the threshold gradient descending paths are much "smoother" than the exact paths. To determine the penalty parameter λ, we use $L(\cdot)$ as the accuracy measure and choose λ by optimizing the average of five independent fivefold crossvalidated estimators of the accuracy function as shown in Web Figure 1. The corresponding point estimators of the regression coefficients are summarized in Table 1. It is interesting to note that according to adaptive LASSO, the only nonzero regression coefficients in the AFT model are tumor size, tumor grade (grade 3 versus 1 or 2), vascular invasion (>3 vessels versus 0–3 vessels), basal-like, luminal A, ErbB2, 70-gene, and wound response gene signatures, which suggests that the conventional clinical prognostic factors and selected gene expression signatures are complementary to each other in terms of predicting future survival time. Reported also in Table 1 are the standard error estimates for both LASSO and adaptive LASSO regularized point estimators. However, as discussed in Section 2.2, the finite sample distribution of the proposed regularized estimators may be far from normal and the standard error may not be a good summary for the precision of these estimators.

To internally validate the prediction performance of the models constructed based on the regularization methods, we randomly split the data into a training sample and a validation sample of equal sizes. We fit the training data with the AFT model via the proposed procedures to obtain estimates of the regression coefficients. Based on these estimates, we then predict the risk of failure for subjects in the validation sample and classify them as high or low risk based on whether the predicted risk exceeds the median risk. The selection of the penalty parameters in the training stage was based on the fivefold crossvalidation. This process was repeated 500 times. The results demonstrate the LASSO and adaptive LASSO estimators achieve better risk stratifications compared to the unregularized counterpart with respect to their significance in testing the difference between the two risk groups. For example, the observed proportions of p-values being smaller than 0.05 are 96.6%, 91.2%, and 82.2% for predictions based on LASSO, adaptive LASSO, and unregularized estimators, respectively. The entire empirical cumulative distribution functions of the 500 p-values for comparing the two risk groups identified by the LASSO, adaptive LASSO, and unregularized Gehan's estimators are shown in Web Figure 2. Furthermore, the LASSO and adaptive LASSO procedures led to prediction models with an average of 8 and 7 predictors, respectively.

6. Simulation Study

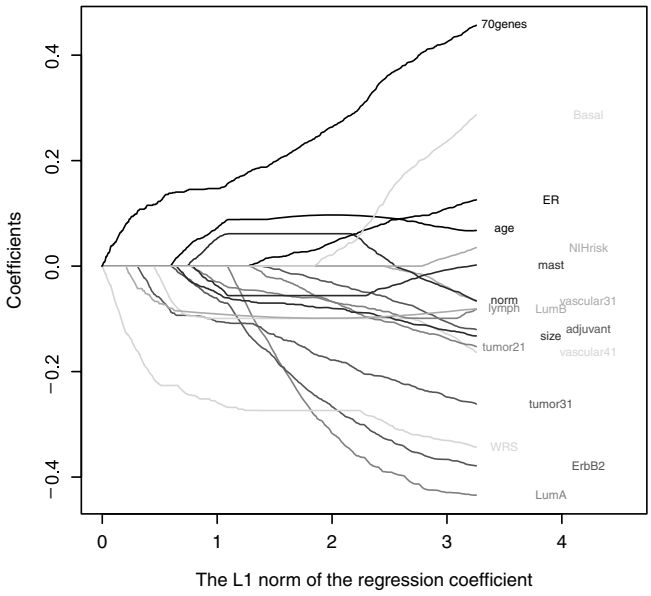
In this section, we examine the finite sample performance of the proposed methods through simulation studies. We mimic the simulation setup considered in Tibshirani (1997). Specifically, we generate the survival time from the exponential distribution with rate $\lambda = \exp(-\boldsymbol{\beta}_0'\mathbf{Z})$, i.e., $\log(T) = \boldsymbol{\beta}_0'\mathbf{Z} + \epsilon$, where $\mathbf{Z} = (Z_1, \ldots, Z_9)'$ is generated from a multivariate normal with mean zero and the variance covariance matrix $\boldsymbol{\Sigma}_Z = (\sigma_{ij}) = (\rho^{|i-j|})$ and ϵ follows the standard extreme value distribution. We considered two sets of regression coefficients: $\boldsymbol{\beta}_0 = (0.35,\ 0.35,\ 0,\ 0,\ 0,\ 0.35,\ 0,\ 0,\ 0)'$ and $(0.7,\ 0.7,\ 0,\ 0,\ 0,\ 0.7,\ 0,\ 0,\ 0)'$, to represent weak and moderate associations between the predictors and the survival time, respectively. Three different values of ρ were considered, $\rho = 0$, 0.5, and 0.9, corresponding to zero, moderate, and strong collinearity among the predictors. The censoring time was generated from a uniform$[0, \xi]$, where ξ was chosen to achieve about 40% of censoring.

For each scenario, the three estimation procedures were evaluated based on 100 simulated datasets at sample sizes of $n = 50$ and 100: the Gehan's rank-based, BJ iterative and IPW methods. For each estimation method, we investigated four regularization procedures: (a) the oracle procedure with β_3, β_4, β_5, β_7, β_8, and β_9 given as 0; (b) the unregularized procedure including all predictors in the model; (c) the LASSO; and (d) the adaptive LASSO. The oracle procedure, while



(a) LASSO

(b) Adaptive LASSO



(c) Approximate LASSO

Figure 2. Paths of the regression coefficients (a) as a function of $\sum_{k=1}^{p} |\beta_k|$ for the exact LASSO (b) as a function of $\sum_{k=1}^{p} |\beta_k|/|\hat{\beta}_{G\,k}|$ for the adaptive LASSO; (c) as a function of $\sum_{k=1}^{p} |\beta_k|$ based on the threshold gradient descending approximation (Approximate LASSO), for the breast cancer example. The vertical lines correspond to the estimated optimal penalty parameters.

not available in practice, may serve as an optimal benchmark for the purpose of comparisons. The penalty parameters used in the LASSO (adaptive LASSO) regularized rank estimator were selected based on the fivefold crossvalidation assisted by the proposed efficient algorithm for computing the exact regularization path. We implemented the procedure proposed by Wang et al. (2008) to compute the LASSO (adaptive LASSO) regularized BJ estimator with regularization applied to the least square fitting within each iteration. The procedure failed to converge for a significant number of datasets due to

Table 1
Estimated regression coefficients and standard errors of all clinical predictors and gene signatures for the breast cancer data example

	Gehan's estimator		LASSO		Adaptive LASSO	
			Clinical predictors			
Age (decades)	0.264	(0.166)[a]	0.123	(0.068)	0.000	(–)
Tumor size (diameter, cm)	−0.132	(0.129)	−0.042	(0.024)	−0.056	(0.033)
Tumor grade						
Grade 3 versus 1	−0.678	(0.396)	−0.225	(0.082)	−0.281	(0.205)
Grade 2 versus 1	−0.405	(0.342)	0.000	(–)	0.000	(–)
Vascular invasion						
1–3 vessels versus 0 vessel	0.268	(0.312)	0.000	(–)	0.000	(–)
>3 vessels versus 0 vessel	−0.122	(0.557)	−0.276	(0.104)	−0.453	(0.047)
Estrogen receptor status (positive versus negative)	−0.261	(0.455)	0.000	(–)	0.000	(–)
Mastectomy versus breast conserving therapy	−0.166	(0.213)	−0.053	(0.035)	0.000	(–)
No adjuvant versus chemo or hormonal therapy	0.096	(0.287)	0.000	(–)	0.000	(–)
The number of lymph nodes	−0.024	(0.078)	−0.013	(0.006)	0.000	(–)
NIH risk status (high versus intermediate or low)	0.208	(0.387)	0.000	(–)	0.000	(–)
			Gene signatures			
Wound response	−1.198	(0.683)	−1.493	(0.371)	−1.915	(0.505)
70-gene	1.181	(0.523)	0.417	(0.119)	0.945	(0.334)
Normal-like	0.460	(1.675)	0.561	(0.297)	0.000	(–)
ErbB2	−2.926	(1.230)	−0.395	(0.151)	−2.161	(0.761)
Luminal A	−1.827	(1.512)	0.000	(–)	−1.478	(0.635)
Luminal B	0.729	(1.577)	−0.251	(0.109)	0.000	(–)
Basal-like	−0.763	(1.619)	0.000	(–)	0.315	(0.120)

[a]The estimated standard error for the point estimator.

nonconvergent loops. For such cases, we selected the first member in the loop as the point estimator. The penalty parameter of the regularized BJ estimator was adaptively selected by the generalized crossvalidation method given in Wang et al. (2008). The regularized IPW estimator is constructed as in Datta et al. (2007). Specifically, we replaced the unobservable response $\log(T_i)$ by $\log(T_i^*) = \log(T_i)\delta_i/\hat{G}(T_i)$ and applied the LASSO (adaptive LASSO) regularization method to the synthetic dataset $\{(\log(T_i^*), \mathbf{Z}_i), i = 1, \ldots, n\}$, where $\hat{G}(t)$ is the Kaplan–Meier estimator of $P(C > t)$. The penalty parameters were simply selected using the Mallow's C_p criteria based on the synthetic data.

For each estimate $\hat{\boldsymbol{\beta}}$ obtained from the proposed methods, we generate an independent validation set of size $100n$ and evaluate the prediction performance of $\hat{\boldsymbol{\beta}}'\mathbf{Z}$ based on the mean squared prediction errors, $\text{MSE} = (100n)^{-1}\sum_{j=1}^{100n}\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)'\mathbf{Z}_j^{\text{val}}\}^2$, where $\{\mathbf{Z}_j^{\text{val}}\}$ are covariate vectors in the validation sample. To examine how well the proposed procedures perform with respect to variable selection, we recorded the frequency of the regression coefficients being set to zero correctly and incorrectly for both the LASSO and adaptive LASSO procedures.

The results, summarized in Tables 2 and 3, exhibit several interesting patterns. First, with respect to mean squared prediction errors, the rank-based methods outperform both the BJ and IPW methods. Compared with the MSEs of the regularized BJ estimators, the MSEs of the regularized Gehan estimators are about 30–40% smaller when the signals are weak. The difference between the two procedures is less apparent when the signals are moderate and the correlation ρ is not large. The IPW method performs the worst among the three methods, especially when the signals are moderate (Table 3). This is in part due to the violation of the assumption on the censoring support. Secondly, both the LASSO and adaptive LASSO regularizations can significantly improve the prediction MSE compared to their unregularized counterpart for both rank-based and BJ procedures. Thirdly, both the regularized rank-based and BJ estimators can correctly identify the majority of the zero regression coefficients. Furthermore, the precision increases with the sample size and the signal strength. It is interesting to note that, when the sample size is relatively small and the signal is relatively weak, the adaptive LASSO may *not* be more accurate than the LASSO with respect to correctly identifying the zero coefficients or reduction in MSE. This could in part be attributed to the large variations of the initial estimator used to determine the weights in the adaptive LASSO regularization. Fourthly, for both LASSO and adaptive LASSO, the ability to correctly identify zero coefficients weakens as the collinearity among the predictors becomes higher. This may be explained by the fact that it is more difficult to differentiate two highly correlated predictors, among which only one is truly associated with the

Table 2

Simulation study: Mean squared prediction error, proportion of zero coefficients being set to zero ($\mathcal{P}_{0+}$), and proportion of nonzero coefficients being set to zero ($\mathcal{P}_{0-}$) of model selection results based on the Orcale estimator, the standard unregularized estimator, the LASSO estimator, and the adaptive LASSO (ALASSO) estimator using Gehan's rank based, BJ (BJ) iterative, and IPW procedures, for $\boldsymbol{\beta}_0 = (0.35, 0.35, 0, 0, 0, 0.35, 0, 0, 0)'$

		$\rho = 0$			$\rho = 0.5$			$\rho = 0.9$		
		Gehan	BJ	IPW	Gehan	BJ	IPW	Gehan	BJ	IPW
						MSE				
50	Oracle	0.19	0.22	0.15	0.17	0.21	0.16	0.18	0.22	0.22
	Regular	0.68	0.72	0.41	0.67	0.70	0.44	0.69	0.74	0.50
	LASSO	0.33	0.35	0.31	0.30	0.40	0.33	0.25	0.27	0.31
	ALASSO	0.36	0.37	0.33	0.36	0.43	0.35	0.29	0.30	0.32
100	Oracle	0.08	0.08	0.08	0.08	0.08	0.10	0.08	0.08	0.16
	Regular	0.25	0.26	0.20	0.25	0.26	0.21	0.26	0.27	0.26
	LASSO	0.18	0.23	0.21	0.14	0.22	0.22	0.10	0.14	0.21
	ALASSO	0.17	0.22	0.19	0.16	0.20	0.21	0.12	0.15	0.22
						$\mathcal{P}_{0+}$				
50	Oracle	1	1	1	1	1	1	1	1	1
	Regular	0	0	0	0	0	0	0	0	0
	LASSO	0.70	0.87	0.80	0.69	0.78	0.74	0.69	0.71	0.74
	ALASSO	0.73	0.87	0.78	0.70	0.81	0.76	0.71	0.73	0.76
100	Oracle	1	1	1	1	1	1	1	1	1
	Regular	0	0	0	0	0	0	0	0	0
	LASSO	0.61	0.83	0.72	0.64	0.81	0.73	0.65	0.77	0.72
	ALASSO	0.74	0.87	0.77	0.74	0.84	0.77	0.71	0.82	0.77
						$\mathcal{P}_{0-}$				
50	Oracle	0	0	0	0	0	0	0	0	0
	Regular	0	0	0	0	0	0	0	0	0
	LASSO	0.40	0.72	0.55	0.33	0.55	0.49	0.46	0.49	0.57
	ALASSO	0.38	0.69	0.52	0.39	0.61	0.53	0.55	0.63	0.66
100	Oracle	0	0	0	0	0	0	0	0	0
	Regular	0	0	0	0	0	0	0	0	0
	LASSO	0.18	0.44	0.31	0.10	0.31	0.28	0.32	0.42	0.44
	ALASSO	0.17	0.39	0.28	0.18	0.35	0.35	0.46	0.56	0.59

response. Lastly, the improvements in the prediction performance and the ability to identify the correct model may not be simultaneously realized. For example, with smaller sample size and weaker association, the relative improvement in the prediction MSE over the usual Gehan's estimator is greater, while the probability of correctly detecting the true model is smaller.

Because the support of the censoring is shorter than that of the failure time, the IPW procedure may lead to biased estimators under the current simulation setting. Consequently, in the second set of simulations for examining the finite sample properties of the regularized methods with high-dimensional covariates, we focused on the proposed rank-based and BJ iterative procedures. Specifically, we adopt the similar simulation setup as described above with $\boldsymbol{\beta}_0 = (0.7, 0.7, 0, 0, 0, 0.7, 0, \ldots, 0)'$, and the sample size fixed at $n = 50$. The dimension of β_0, p, was set at 15, $20, \ldots, 45$, and 50. Considering the 40% censoring, the average number of observed failure is 30 and thus $p \geqslant 25$ can be considered as high dimension. For the high-dimensional setup, there is a substantial proportion of simulated datasets, in which the finite unregularized rank or BJ estimator does not exist and thus we only calculate the LASSO-regularized estimators.

The detailed results about the mean squared prediction errors for the second set of simulation are summarized in Figure 3. The patterns shown in the first set of simulations are in general maintained: the LASSO regularization can drastically improve the prediction accuracy especially when p is big relative to n; the performance of the adaptive LASSO estimators deteriorates more rapidly than their LASSO counterpart as p increases; and the regularized rank-based estimators outperform their BJ counterparts, especially when the covariates are highly correlated.

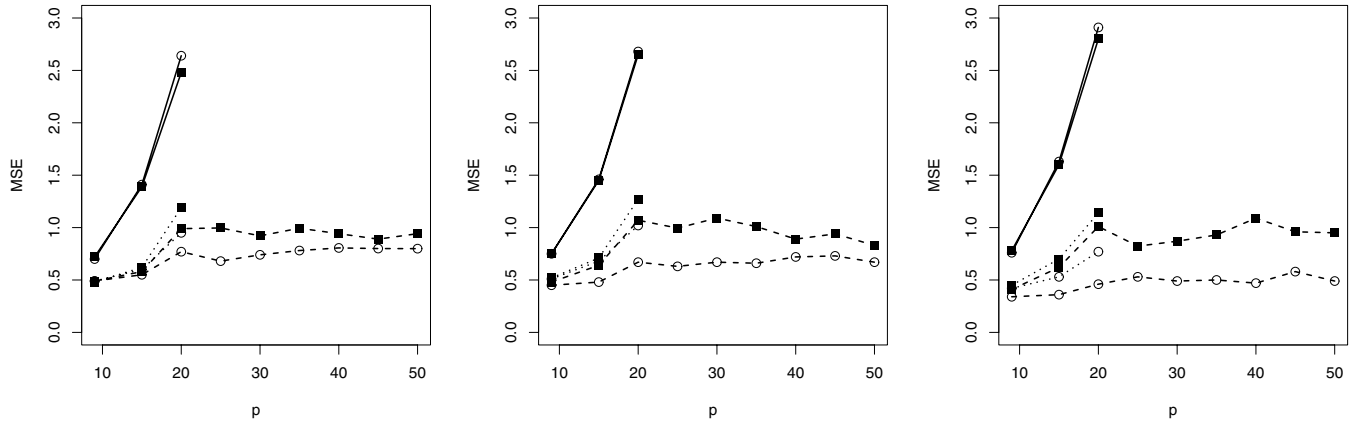
7. Discussion

The proposed regularization methods for the AFT model can be easily extended to incorporate other types of penalty functions such as the L_2 or the more general elastic net regularization (Zou and Hastie, 2005). The entire regularization path with the L_2 or elastic net penalty would also be piecewise linear and can be obtained by modifying the algorithm proposed by Hastie et al. (2004). The Gehan's initial estimator

Table 3

Simulation study: Mean squared prediction error, proportion of zero coefficients being set to zero ($\mathcal{P}_{0+}$), and proportion of nonzero coefficients being set to zero ($\mathcal{P}_{0-}$) of model selection results based on the Orcale estimator, the standard unregularized estimator, the LASSO estimator, and the adaptive LASSO (ALASSO) estimator using Gehan's rank based, BJ (BJ) iterative, and IPW procedures, for $\boldsymbol{\beta}_0 = (0.7, 0.7, 0, 0, 0, 0.7, 0, 0, 0)'$

		$\rho = 0$			$\rho = 0.5$			$\rho = 0.9$		
		Gehan	BJ	IPW	Gehan	BJ	IPW	Gehan	BJ	IPW
					MSE					
50	Oracle	0.19	0.23	0.31	0.21	0.24	0.41	0.23	0.25	0.68
	Regular	0.70	0.72	0.60	0.75	0.75	0.70	0.76	0.78	0.97
	LASSO	0.49	0.49	0.64	0.45	0.48	0.73	0.34	0.41	0.84
	ALASSO	0.49	0.48	0.61	0.51	0.52	0.70	0.42	0.45	0.88
100	Oracle	0.08	0.08	0.23	0.09	0.09	0.32	0.09	0.09	0.60
	Regular	0.25	0.26	0.33	0.26	0.27	0.43	0.27	0.28	0.72
	LASSO	0.18	0.18	0.40	0.17	0.18	0.49	0.14	0.18	0.72
	ALASSO	0.14	0.14	0.34	0.17	0.17	0.45	0.19	0.21	0.73
					$\mathcal{P}_{0+}$					
50	Oracle	1	1	1	1	1	1	1	1	1
	Regular	0	0	0	0	0	0	0	0	0
	LASSO	0.56	0.61	0.65	0.60	0.64	0.66	0.63	0.63	0.68
	ALASSO	0.70	0.75	0.77	0.70	0.73	0.74	0.68	0.72	0.75
100	Oracle	1	1	1	1	1	1	1	1	1
	Regular	0	0	0	0	0	0	0	0	0
	LASSO	0.48	0.70	0.58	0.54	0.70	0.67	0.62	0.70	0.66
	ALASSO	0.70	0.85	0.80	0.75	0.84	0.79	0.73	0.81	0.77
					$\mathcal{P}_{0-}$					
50	Oracle	0	0	0	0	0	0	0	0	0
	Regular	0	0	0	0	0	0	0	0	0
	LASSO	0.06	0.07	0.13	0.07	0.10	0.16	0.23	0.28	0.36
	ALASSO	0.07	0.08	0.18	0.11	0.15	0.24	0.41	0.45	0.54
100	Oracle	0	0	0	0	0	0	0	0	0
	Regular	0	0	0	0	0	0	0	0	0
	LASSO	0	0	0.01	0	0	0.01	0.09	0.12	0.23
	ALASSO	0	0	0.01	0.01	0.02	0.05	0.24	0.33	0.42



(a) $\rho = 0$ (b) $\rho = 0.5$ (c) $\rho = 0.9$

Figure 3. Mean squared prediction errors for the unregularized (solid), LASSO (dotted), and adaptive LASSO (dashed) estimators based on Gehan's rank based (circle) and BJ iterative (solid square) procedures with $p = 9$, 15, 20, ..., 45, and 50.

determining the weights used in the adaptive LASSO may be too unstable or even not available for a high-dimensional $\boldsymbol{\beta}$. For such settings, one may instead use the L_2 regularized Gehan's estimator as the initial estimator.

When fitting the AFT model in the standard setting with a small p, a more efficient estimator may be obtained by using different weight functions in equation (1). In such cases, the root of the estimating equation may be obtained by an iterative algorithm, in which each iteration amounts to minimizing a weighted Gehan's objective function (Jin et al., 2003). Therefore, a simple regularization strategy for the general rank-based estimating equation is to apply LASSO or adaptive LASSO regularization within each iteration. However, the resulting regularized solution may lose the simple interpretation as a constrained minimizer. It is important to note that while the proposed procedure may be carried out when p increases with the sample size, the asymptotical properties derived in Web Appendices A and B only hold when p is a fixed constant. Using similar arguments as given in Huang, Ma, and Zhang (2008), one may extend the results to the setting when $p = p_n \to \infty$ as $n \to \infty$ but at a slower rate. When p is much bigger than the sample size, e.g., in the context of gene expression data analysis, operationally, the proposed regularization method can be performed with a large number of individual gene expression as covariates in the regression analysis. However, because the theoretical results require that the dimension of predictor is fixed while the sample size $n \to \infty$, we suggest performing an initial screening step, in which relatively few covariates were selected/constructed from the original gene expression measurements, and then conduct the regularized multivariate analysis with the covariates formed in the first step. Note that even after the initial dimension reduction step, the dimension of predictors may still be not small relative to the sample size for performing the standard unregularized estimation as in the breast carcinomas example and this is where the proposed regularization methods are intended to be applied.

The selection of an appropriate penalty parameter is crucial to the performance of regularized estimators. If the primary goal of the regularization is variable selection, i.e., to identify noninformative predictors whose true regression coefficients are zero, one may consider approaches different from optimizing a crossvalidated loss function. Intuitively, the penalty parameter should be set such that the LASSO estimators for most noninformative predictors are zero. One possible ad hoc approach to achieve this is to first augment existing predictors by several randomly generated noise variables that are independent of the survival time and then calculate the entire LASSO regularization path with the augmented predictors. In the end, one may choose the smallest penalty parameter such that all the LASSO regularized regression coefficients of those augmented noise predictors are zero.

8. Supplementary Materials

Web Appendices and Figures referenced in Sections 2.2, 4, and 5 are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

The authors are very grateful to the referees and the editors for insightful comments on the article. The work is partially supported by grant R01 GM079330-01 from the National Institutes of Health.

References

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika* **66,** 429–436.

Chang, H., Nuyten, D., Sneddon, J., Hastie, T., Tibshirani, R., Sorlie, T., Dai, H., He, Y., van't Veer, L., Bartelink, H., van de Rijn, M., Brown, P., and van de Vijver, M. (2005). Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *PNAS* **102,** 3738–3743.

Currie, I. (1996). A note on Buckley-James estimators for censored data. *Biometrika* **83,** 912–915.

Datta, S., Le-Rademacher, J., and Datta, S. (2007). Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and LASSO. *Biometrics* **63,** 259–271.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. T. (2004). Least angle regression. *The Annals of Statistics* **32,** 407–499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96,** 1348–1360.

Fleming, T. and Harrington, D. (2001). *Counting Processes and Survival Analysis*. New York: Wiley.

Friedman, J. and Popescu, B. (2004). *Gradient Directed Regularization*. Technical Report, Stanford University, Stanford, California.

Gui, J. and Li, H. (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size setting, with applications to microarray gene expression data. *Bioinformatics* **21,** 3001–3008.

Hastie, T., Rosset, S., Tibshirani, R., and Zhu, J. (2004). The entire regularization path for the support vector machine. *Journal of Machine Learning Research* **5,** 1391–1415.

Hoerl, A. and Kennard, R. (1970). Ridge regression: Biased estimation for non-orthogonal problem. *Technometrics* **12,** 55–67.

Huang, J. and Harrington, D. (2005). Iterative partial least squares with right-censored data analysis: A comparison to other dimension reduction techniques. *Biometrics* **61,** 17–24.

Huang, J., Ma, S., and Xie, H. (2006). Regularized estimation in the accelerated failure time model with high dimensional covariates. *Biometrics* **62,** 813–820.

Huang, J., Ma, S., and Zhang, C. (2008). Adaptive LASSO for sparse high-dimensional regression models. *Statistica Sinica*, in press.

Jin, Z., Lin, D. Y., and Wei, L. J. (2003). Rank-based inference for the accelerated failure time model. *Biometrika* **90,** 341–353.

Jolliffe, I. (1986). *Principal Component Analysis*. New York: Springer-Verlag.

Kim, Y. and Kim, J. (2004). Gradient LASSO for feature selection. In *Proceedings of the Twenty-First International Conference on Machine Learning*, 473–480. New York: ACM.

Knight, K. and Fu, W. (2000). Asymptotics for LASSO-type estimators. *Annals of Statistics* **28,** 1356–1378.

Koul, H., Susarla, V., and Van Ryzin, J. (1981). Regression analysis with randomly right-censored data. *Annals of Statistics* **9,** 1276–1288.

Leng, C. and Ma, S. (2007). Path consistent model selection in additive risk model via LASSO. *Statistics in Medicine* **26,** 3753–3770.

Li, H. (2008). Censored data regression in high-dimension and low sample size settings for genomic applications. In *Statistical Advances in Biomedical Sciences: State of Art and Future Directions*, A.

Biswas, S. Datta, J. Fine, and M. Segal (eds), 1st edition. Hoboken, New Jersey: Wiley.

Lu, W. and Zhang, H. (2007). Variable selection for proportional odds model. *Statistics in Medicine* **26,** 3771–3781.

Martens, H. and Naes, T. (1989). *Multivariate Calibration*. Chichester: John Wiley & Sons Ltd.

Park, S. (1981). Collinearity and optimal restrictions on regression parameters for estimating responses. *Technometrics* **23,** 289–295.

Perou, C., Sorlie, T., Eisen, M., van de Rijn, M., Jeffrey, S., Rees, C., Pollack, J., Ross, D., Johnsen, H., Akslen, L., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S., Lonning, P., Borresen-Dale, A., Brown, P., and Botstein, D. (2000). Modecular portaits of human breast tumours. *Nature* **6797,** 747–752.

Ritov, Y. (1990). Estimation in a linear regression model with censored data. *Annals of Statistics* **18,** 303–328.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Association, Series B* **58,** 267–288.

Tibshirani, R. (1997). The LASSO method for variable selection in the Cox model. *Statistics in Medicine* **16,** 385–395.

Tsiatis, A. (1996). Estimating regression parameters using linear rank tests for censored data. *Annals of Statistics* **18,** 305–328.

van de Vijver, M., He, Y., van't Veer, L., Dai, H., Hart, A., Voskuil, D., Schreiber, G., Peterse, J., Roberts, C., Marton, M., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E., Friend, S., and Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine* **347,** 1999–2009.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer.

Wang, H., Li, G. and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection via the LAD-LASSO. *Journal of Business and Economics Statistics* **25,** 347–355.

Wang, S., Nan, B., Zhu, J., and Beer, D. (2008). Doubly penalized Buckley-James method for survival data with high dimensional covariates. *Biometrics* **64,** 132–140.

Wei, L. J. (1992). The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine* **11,** 1871–1879.

Zhang, H. and Lu, W. (2007). Adaptive LASSO for Cox's proportional hazards model. *Biometrika* **94,** 691–703.

Zheng, Y., Cai, T., and Feng, Z. (2006). Application of the time-dependent ROC curves for prognostic accuracy with multiple biomarkers. *Biometrics* **62,** 279–287.

Zhu, J., Hastie, T., Rosset, S., and Tibshirani, R. (2003). 1-norm support vector machines. *Neural Information Processing Systems* **16,** 49–56.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association* **101,** 1418–1429.

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

Received September 2007. *Revised March* 2008.
Accepted April 2008.