Biostatistics (2009), **1**, 1, *pp.* 1–8
Printed in Great Britain

Rank-based estimation in the ℓ_1-regularized partly linear model for censored outcomes with application to integrated analyses of clinical predictors and gene expression data

By BRENT A. JOHNSON

Department of Biostatistics, Emory University, Atlanta, GA 30322, USA
e-mail: bajohn3@emory.edu

Summary

We consider estimation and variable selection in the partial linear model for censored data. The partial linear model for censored data is a direct extension of the accelerated failure time model, the latter of which is a very important alternative model to the proportional hazards model. We extend rank-based lasso-type estimators to a model which may contain nonlinear effects. Variable selection in such partial linear model has direct application to high-dimensional survival analyses which attempt to adjust for clinical predictors. In the microarray setting, previous methods can adjust for other clinical predictors by assuming that clinical and gene expression data enter the model linearly in the same fashion. Here, we select important variables after adjusting for prognostic clinical variables but the clinical effects are assumed nonlinear. Our estimator is based on stratification and can be extended naturally to account for multiple nonlinear effects. We illustrate the utility of our method through simulation studies and application to the Wisconsin prognostic breast cancer data set.

Some key words: Lasso; Logrank; Penalized least squares; Survival analysis

1. Introduction

This note is concerned with estimation and computation in the ℓ_1-regularized partial linear model for censored data. To fix ideas, we write the statistical model

$$\log T_i = \phi(\mathbf{Z}_i) + \mathbf{X}_i'\boldsymbol{\beta} + \varepsilon_i,\ \ (i = 1, \ldots, n), \tag{1}$$

where T_i is a failure time variable, $\phi(\mathbf{Z}_i)$ is a unknown function of predictors $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})'$, $\mathbf{X}_i$ is a d-vector of fixed predictors, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)'$ is a d-vector of regression coefficients, and $(\varepsilon_1, \ldots, \varepsilon_n)$ are independent and identically distributed errors with distribution function F. The goal is to estimate the regression coefficients $\boldsymbol{\beta}$ while setting some estimates equal to zero using the observed data $\{(Y_i, \delta_i, \mathbf{X}_i, \mathbf{Z}_i),\ i = 1, \ldots, n\}$, where $Y_i = \min(T_i, C_i)$, $\delta_i = I(T_i \leq C_i)$, C_i is a random censoring variable and $I(\cdot)$ denotes the indicator function. The assumptions one adopts can make this estimation problem challenging theoretically and numerically.

In survival analysis, the statistical model (1) without the nonlinear term $\phi(\cdot)$ is called the semiparametric accelerated failure time (AFT) model (cf. Kalbfleisch and Prentice, 2002). One family of estimators for regression coefficients $\boldsymbol{\beta}$ in the AFT model are called the weighted logrank estimators (See Section 2) and derived through inverting linear rank tests (Prentice, 1978; Tsiatis, 1990). Our methods extend the class of weighted logrank estimators and, in the sequel, we adopt the adjective "rank-based" to conform with related methods in the literature (cf. Jin and others, 2003). By now, several authors have studied variable selection in the AFT model (cf. Datta and others, 2007; Huang and others, 2007; Johnson, 2008; Cai and others, 2009). Among the many available methods, only Johnson (2008) and Cai and others (2009) propose procedures based on weighted logrank estimators. In this paper, we propose rank-based variable selection in the partly linear model (1) by extending a stratified Gehan-type estimator (Chen and others, 2005). The stratified estimator is advantageous in that it allows for consistent estimation of regression coefficients without nonparametric smoothing via splines or kernels.

These methods were developed for a microarray application at Emory's Winship Cancer Institute relating gene expression data and time to prostate cancer recurrence, which may be right-censored. Most modern model selection techniques perform variable selection on an arbitrary set of predictors. It is the user's prerogative to control the input predictors, including some collection of gene expression, clinical predictors or both. Unfortunately, the user's choices may not reflect well what the scientist really desires. If one selects variables on either gene expression or clinical variables independently, the final model does not accurately reflect the complex correlations among the clinical and gene expression data. If we include clinical predictors along-side gene expression with no account of the variable type, then the potential problems are two-fold. First, the final model may exclude a clinical variable which we know to be scientifically relevant. Second, clinical predictors and gene expression data are treated as equals in the eyes of the statistical learner. For users familiar with the underlying (optimization) techniques of specific model selection methods, it is possible to circumvent the former problem by forcing clinical variables in the model and, hence, only regularize the gene expression data. This method of forcing active coefficients within regularized estimation does not address the latter problem, however, and is possibly beyond the expertise of the average user. Estimation and variable selection in the partial linear model has the potential to address both scientific issues simultaneously.

The main substantive contribution of the paper is the idea of jointly modeling clinical and genetic predictors through the partly linear model and simultaneously performing model selection on genetic components. The end product of this procedure is a sparse model which includes scientifically-relevant clinical covariates and data-relevant genetic components. The methodological contributions of the paper are two-fold. First, we propose a new rank-based variable selection procedure in the partly linear model for censored data where no similar method exists. The second methodologic contribution is entirely computational. We propose a new algorithm for regularized estimation in the AFT model which extends naturally to the partly linear model for censored data. Existing computational strategies for regularized rank-based estimation in the AFT model include local quadratic approximation atop simulated annealing (Johnson, 2008) and a path-based algorithm (Cai and others, 2009). The algorithm by Cai and others (2009) produces exact lasso coefficient estimates while Johnson's (2008) method does not. Our new algorithm produces precise lasso coefficient estimates through an intriguing extension of least absolute deviation regression. Finally, the new procedure is propagated easily as the algorithm can be adapted to the `quantreg` package in `R`.

2. Methods

2.1. Background

A classic definition of the Gehan estimator (Prentice, 1978; Tsiatis, 1990) is defined as the solution to the system of estimating equations, $0 = \mathbf{U}_G(\boldsymbol{\beta})$, where

$$\mathbf{U}_{\mathrm{G}}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\,(\mathbf{X}_i - \mathbf{X}_j)\,I\{e_i(\boldsymbol{\beta}) \le e_j(\boldsymbol{\beta})\},$$

and $e_i(\boldsymbol{\beta}) = \log Y_i - \mathbf{X}_i'\boldsymbol{\beta}$. Evidently, $\mathbf{U}_{\mathrm{G}}(\boldsymbol{\beta})$ is the d-dimensional gradient of the convex loss function, $n\mathcal{L}_{\mathrm{G}}(\boldsymbol{\beta})$, where

$$\mathcal{L}_{\mathrm{G}}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\{e_i(\boldsymbol{\beta}) - e_j(\boldsymbol{\beta})\}^-,$$

$c^- = \max(-c, 0)$. Jin and others (2003) approximated $\mathcal{L}_{\mathrm{G}}(\boldsymbol{\beta})$ by $\mathcal{L}_{\mathrm{M}}(\boldsymbol{\beta})$, where

$$\mathcal{L}_{\mathrm{M}}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i|e_i(\boldsymbol{\beta}) - e_j(\boldsymbol{\beta})| + \left|M - \boldsymbol{\beta}'\sum_{k=1}^{n}\sum_{l=1}^{n}\delta_k(\mathbf{X}_l - \mathbf{X}_k)\right|,$$

and M is a large constant. Because the loss $\mathcal{L}_{\mathrm{M}}(\boldsymbol{\beta})$ is written as the sum of absolute deviations, the minimizer may be found using least absolute deviation (lad) regression (e.g. `quantreg` in `R`).

The Gehan estimator with lasso (Tibshirani, 1996) penalty is defined $\widehat{\boldsymbol{\beta}}_{\mathrm{G}(1)} = \min_{\boldsymbol{\beta}}\{\mathcal{L}_{\mathrm{G}}(\boldsymbol{\beta}) + \lambda_n\sum_{j=1}^{d}|\beta_j|\}$. Both Johnson (2008) and Cai and others (2009) note that $\widehat{\boldsymbol{\beta}}_{\mathrm{G}(1)}$ is the solution to a linear programming problem. However, using the approximation by Jin and others (2003), the lasso-type estimator is equivalently written as $\min_{\boldsymbol{\beta}}\{\mathcal{L}_{\mathrm{M}}(\boldsymbol{\beta}) + \lambda_n\sum_{j=1}^{d}|\beta_j|\}$. The significance of the approximation is that the resulting constrained optimization may be carried out through simple data augmentation. In an unpublished 2008 Emory University Technical Report, B.A. Johnson showed that if the Gehan estimate is the solution to the lad regression of $\mathbf{V}$ on $\mathbf{W}$, then the regularized Gehan estimate is simply the solution to the lad regression of $\mathbf{V}^*$ on $\mathbf{W}^*$, where $\mathbf{V}^* = (\mathbf{V}', \mathbf{0}_d')'$, $\mathbf{W}^* = (\mathbf{W}', \lambda\mathbf{I}_d)'$, where $\mathbf{0}_d$ is a d-dimensional vector of zeros and $\mathbf{I}_d$ is an identity matrix of size d. The data augmentation technique for regularized lad estimates with uncensored data was first proposed by Wang and others (2007).

2.2. The stratified Gehan estimator

Chen and others (2005) recently proposed a rank-based estimator in the partly linear model for censored data. Their estimator extends the Gehan estimator by stratifying over levels of $\mathbf{Z}$ and arguing that such procedure leads to a consistent estimator of the regression coefficients in (1). Compared with the majority of estimators in partly linear model with uncensored data, the estimator by Chen and others (2005) is different in that it does not require nonparametric smoothing.

Intuitively, the estimator by Chen and others (2005) is defined by stratifying the sample into K_n strata $\{S_1, \ldots, S_{K_n}\}$ according to user-defined levels of $\mathbf{Z}$ and minimizing a new stratified loss function. Let $\mathbb{I}_k$ denote the indices of subjects belonging to strata S_k. Their argument is that for subjects belonging to the same strata S_k, we have $\phi(\mathbf{Z}_i) = c_k + R_n$, for all $i \in \mathbb{I}_k$, where the constant c_k varies by strata, $k = 1, \ldots, K_n$ and R_n is an asymptotically negligible remainder term. Chen and others (2005) propose to minimize the loss function

$$\mathcal{L}_{\mathrm{S}}(\boldsymbol{\beta}) = \sum_{k=1}^{K_n} \sum_{i,j \in \mathbb{I}_k} \delta_i |e_i(\boldsymbol{\beta}) - e_j(\boldsymbol{\beta})| + \left| M - \boldsymbol{\beta}' \sum_{k=1}^{K_n} \sum_{l,m \in \mathbb{I}_k} \delta_m (\mathbf{X}_l - \mathbf{X}_m) \right|,$$

for a large number M, which is a direct generalization of the approximate Gehan loss $\mathcal{L}_M(\boldsymbol{\beta})$ above. Naturally, the ℓ_1-regularized stratified estimator is defined

$$\widehat{\boldsymbol{\beta}}_{\mathrm{S}(1)} = \min_{\boldsymbol{\beta}} \left\{ \mathcal{L}_{\mathrm{S}}(\boldsymbol{\beta}) + \lambda_n \sum_{j=1}^{d} |\beta_j| \right\}.$$

When $K_n = 1$, the stratified estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{S}(1)}$ reduces to the Gehan estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{G}(1)}$.

2.3. Operating characteristics

Due to space limitations, we briefly outline the basic large sample properties for our lasso-type extension of the stratified Gehan estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{S}(1)}$. To place the concepts in proper context, it will be easier to work with general convex loss function. Without loss of generality, define a convex loss function $\mathcal{L}_{\bullet}(\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ belongs to a compact parameter space, $\boldsymbol{\beta}_0$ is the true value for $\boldsymbol{\beta}$, and we assume that $\lim n^{-1}\mathcal{L}_{\bullet}(\boldsymbol{\beta})$ converges strongly to a finite limit, uniformly in $\boldsymbol{\beta}$. Define the lasso-type estimator $\widehat{\boldsymbol{\beta}}_{\bullet(1)} = \min_{\boldsymbol{\beta}}\{n^{-1}\mathcal{L}_{\bullet}(\boldsymbol{\beta}) + \lambda_n \sum_j |\beta_j|\}$, the gradient vector $\mathbf{U}(\boldsymbol{\beta}) = (\partial/\partial\boldsymbol{\beta})\mathcal{L}_{\bullet}(\boldsymbol{\beta})$, the slope matrix $\mathbf{A}$ such that $n^{1/2}\mathbf{U}(\boldsymbol{\beta}) - n^{1/2}\mathbf{U}(\boldsymbol{\beta}_0) \cong (\boldsymbol{\beta} - \boldsymbol{\beta}_0)'\mathbf{A}$ and assume that $n^{1/2}\mathbf{U}(\boldsymbol{\beta}_0) \rightarrow_d N(0, \mathbf{B})$. Then, as $n^{1/2}\lambda_n \rightarrow \lambda_0 \geq 0$, one can show that, under suitable regularity conditions, $n^{1/2}(\widehat{\boldsymbol{\beta}}_{\bullet(1)} - \boldsymbol{\beta}_0) \rightarrow_d \operatorname{argmin}\{\boldsymbol{\Omega}(\mathbf{u})\}$, where

$$\boldsymbol{\Omega}(\mathbf{u}) = \mathbf{u}'\mathbf{w} + \mathbf{u}'\mathbf{A}\mathbf{u} + \lambda_0 \sum_{j=1}^{d} \left[u_j \mathrm{sgn}(\beta_{0j}) I(\beta_j \neq 0) + |u_j| I(\beta_{0j} = 0) \right], \qquad (2)$$

and $\mathbf{w}$ is a normal random vector with covariance $\mathbf{B}$. The large sample properties of Tibshirani's (1996) lasso, including the expression in (2), are due to Knight and Fu (2000). The local asymptotic properties may be extended to specific loss functions as special cases. Recently, Huang and others (2006) extended the (2) to inverse probability weighted estimators while Cai and others (2009) considered the extension of (2) to the Gehan estimator, i.e. $\mathcal{L}_{\bullet}(\boldsymbol{\beta}) = \mathcal{L}_{\mathrm{G}}(\boldsymbol{\beta})$, through a novel application of U-processes.

Finally, substitute $\mathcal{L}_{\bullet}(\boldsymbol{\beta}) = \mathcal{L}_{\mathrm{S}}(\boldsymbol{\beta})$ and let $\widehat{\boldsymbol{\beta}}_{\mathrm{S}} = \min \mathcal{L}_{\mathrm{S}}(\boldsymbol{\beta})$. Chen and others (2005) have shown that, under regularity conditions, $n^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathrm{S}} - \boldsymbol{\beta}_0) \rightarrow_d N(0, \mathbf{A}_{\mathrm{S}}^{-1}\mathbf{B}_{\mathrm{S}}\mathbf{A}_{\mathrm{S}}^{-1})$, where $\mathbf{A}_{\mathrm{S}}$ and $\mathbf{B}_{\mathrm{S}}$ are defined in Chen and others (2005, Appendix). By coupling the conditions in Cai and others (2009) along with the conditions in Chen and others (2005), we expect that $n^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathrm{S}(1)} - \boldsymbol{\beta}_0)$ converges in distribution to $\operatorname{argmin}\{\boldsymbol{\Omega}_{\mathrm{S}}(\mathbf{u})\}$, where $\boldsymbol{\Omega}_{\mathrm{S}}(\mathbf{u})$ is defined exactly as in (2) but with $\mathbf{A}_{\mathrm{S}}$ and $\mathbf{B}_{\mathrm{S}}$ replacing $\mathbf{A}$ and $\mathbf{B}$, respectively. Although the statement here is not rigorous, it can be made so under appropriate technical conditions.

3. Application to breast cancer recurrence

Street and others (1995) have studied classification models for breast cancer tumor types and regression models for breast cancer recurrence. Our primary interest lies in the latter regression model for censored data. We adopt the partly linear model for censored data in (1) where T is time (in months) to breast cancer recurrence, $\mathbf{Z} = (Z_1, Z_2)'$ is tumor size (Tsize) and number of lymph nodes (Lnode), respectively, and $\mathbf{X} = (X_1, \ldots, X_{30})'$ is a 30-dimensional feature vector. The feature vector $\mathbf{X}$ is taken from from a digitized image of a fine needle aspirate of a breast mass and describe characteristics of the cell nuclei present in the image. The data consist of three summary statistics (mean, SE, worst) for each of ten features: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. These data are freely available on the UCI repository of machine learning databases (Blake and Merz, 1998). Although a total of 198 samples were collected, only 47 (23.7% of 198) samples were taken from women who experienced breast cancer recurrence. A Kaplan-Meier curve of failure time shows that support of the failure time distribution is modest compared to the support of the follow-up times.

This same data set (i.e. the Wisconsin prognostic breast cancer (WPBC)) was analyzed previously by Bühlmann and Hothorn (2007) using inverse-probability weighted boosting from which they concluded the following ten variables were important: the mean radius, texture, perimeter, smoothness, and symmetry; the standard error (SE) of texture, smoothness, concavepoints, and symmetry; and "worst" concavepoints. We analyzed the WPBC data using our regularized rank-based estimators. Our analyses assume nonlinear effects in one or both of the clinical variables, tumor size or number of lymph nodes. The levels of tumor size were always determined by quantiles while number of lymph nodes was coded by hand. With two levels, the latter strata are defined by zero or greater than zero lymph nodes. For three levels, we split the "greater than zero" group into those with one lymph node and greater than one lymph node. Finally, with four levels, we have a "zero" level, a "one" level, "one to four" lymph nodes, and "greater than four" lymph nodes level. We consider univariate stratification for tumor size and lymph nodes separately in Table 1 while Table 2 considers two-way stratified estimators. For comparison purposes, Table 2 also includes the Gehan lasso (i.e. $K_n = 1$) with and without tumor size and number of lymph nodes. We tuned the regularization parameter through five-fold cross-validation.

Table 1. *Coefficient estimates for rank-based partial linear model stratified on tumor size or number of lymph nodes only. Table entries are multiplied by 1000.*

		Tumor size only					Lymph node only		
Term	$K_n =$	2	3	4	5	9	2	3	4
mean_symmetry		129	164	311	237	283	168	208	224
mean_fractaldim		14	0	198	0	0	73	0	0
SE_perimeter		0	0	-27	0	0	0	0	45
SE_compactness		0	0	-76	0	0	0	0	0
worst_perimeter		-469	-521	-494	-476	-426	-470	-440	-594
worst_smoothness		0	0	-223	0	0	0	0	-25
worst_concavity		0	0	-18	0	0	0	0	-50

In Table 1, we immediately notice that mean symmetry and worst perimeter are strongly associated with breast cancer relapse, a result that appeared consistently across different numbers of strata. Interestingly, the stratified estimator with $K_n = 4$ using only tumor size chose a very complex model compared to the other models. We attribute this to be an artifact of the error in cross-validation. Compared to the ten variables selected by Bühlmann and Hothorn (2007), only mean symmetry is chosen in both procedures. However, we note that worst perimeter is correlated with many of the predictors in the Bühlmann and Hothorn (2007) model — for example, worst perimeter is highly correlated with mean radius ($r = 0{\cdot}92$), mean perimeter ($r = 0{\cdot}93$), and modestly correlated with worst concave points ($r = 0{\cdot}50$). Hence, some model differences may be explained by multicollinearity.

Analytic results for bivariate stratification over tumor size and numbers of lymph nodes are presented in Table 2. The conclusions from results in Table 2 are similar to those reported in Table 1. Now, however, mean fractal dimension is also mildly related to breast cancer recurrence in addition to the two variables from Table 1. Of the seven independent variables in Table 1, only mean symmetry agrees with any of ten variables in Bühlmann and Hothorn (2007). Finally, we found it was difficult to cross-validate the stratified estimator as the number of levels increased and this consideration dictated why the five- and nine-level analyses presented in Table 1 could not be extended in Table 2. This difficulty reflects well-known finite sample limitations of stratified estimators.

Table 2. *Coefficient estimates for regularized Gehan and rank-based partial linear model through bivariate stratification. Table entries are multiplied by 1000.*

	Gehan		2-way stratification (K_n Tumor size, K_n Lymph nodes)							
Term	w/o Z_i	w/Z_i	(2,2)	(2,3)	(2,4)	(3,2)	(3,3)	(3,4)	(4,2)	(4,3)
Tsize		-69								
Lnode		-240								
mean_symmetry	180	172	6	341	175	31	300	466	31	345
mean_fractaldim	8	30	86	50	0	27	213	182	81	16
SE_texture	0	0	0	0	0	0	0	58	0	0
SE_perimeter	0	0	0	0	0	0	0	-114	0	0
SE_compactness	0	0	0	0	0	0	0	-29	0	0
SE_symmetry	0	0	0	0	0	0	-61	-96	0	0
worst_radius	0	-72	0	0	0	0	0	0	0	0
worst_texture	0	0	0	0	0	0	0	34	0	0
worst_perimeter	-603	-426	-291	-390	-433	-296	-393	-449	-272	-373
worst_smoothness	0	0	0	-84	-72	0	-327	-489	0	-80
worst_concavity	0	0	0	0	-8	0	0	-42	0	0

4. Simulation Studies

We conducted numerous simulation studies to assess the cost for ignoring the nonlinear effect in (1) and fitting an ordinary AFT model instead. Due to space limitations, the simulation details have been moved to online supplementary material and only our conclusions are summarized below (See http://www.biostatistics.oxfordjournals.org). First, when the true function ϕ is linear, the model precision from ordinary Gehan lasso beats the stratified estimator, which agrees with intuition. At the same time, it is interesting to note that stratified estimator gradually achieves similar operating characteristics as the unstratified estimator as the sample size increases and number of strata K_n increases. Nevertheless, the stratified estimator is far too cumbersome if the unknown function ϕ is indeed linear in Z_i. Now, the big improvements in the stratified estimator are seen when the unknown function ϕ is nonlinear. For example, when the sample size $n = 75$ and $\sigma = 1{\cdot}5$, the partial model error (PME) is 7.49 and 4.73 for the unpenalized and regularized Gehan, respectively. We compare this to the PME of the unpenalized and regularized stratified estimator, 1.03 and 0.70, respectively. Hence, there is an average seven-fold increase in PME if we fit the Gehan lasso when the true underlying model is a nonlinear (i.e. quadratic) function of Z_i.

5. Remarks

This paper describes rank-based estimation and variable selection in the ℓ_1-regularized partial linear model for censored data. The proposed regularized estimator extends the stratified rank-based estimator by Chen and others (2005). Computationally, we offer a novel strategy for computing regularized Gehan estimates and extend this strategy to stratified estimator. Theoretical properties of the regularized Gehan estimator have been established elsewhere (Johnson, 2008; Johnson and others, 2008; Cai and others, 2009) and we expect that similar properties apply to the stratified estimator under suitable regularity conditions. While we have only focused on lasso estimation, the stratified estimator can accomodate other penalty functions (cf. Johnson and others, 2008) with no additional difficulty. Compared with the computational methods proposed by Johnson (2008) and Cai and others (2009), the methods in this paper have the advantage that they may be easily implemented in standard software.

A premise of this paper is that many applications of variable selection on gene expression are naive in that they do not adequately adjust for important clinical variables. In this paper, we suggest using the partly linear model where clinical predictors enter nonlinearly and gene expression variables enter linearly. If clinical predictors also enter the statistical model linearly, then model (1) reduces to an ordinary AFT model. In simulation studies available as supplementary material (http://www.biostatistics.oxfordjournals.org), we show that there can be a potentially large price to pay in terms of model precision when the true underlying model is partly linear but we fit a linear model instead. Because many clinical predictors are already known to be related to cancer recurrence, building recurrence models through model (1) by selecting genetic features after adjusting for nonlinear clinical effects makes better scientific sense and potentially reduces model error at the same time.

Acknowledgements

The author gratefully acknowledges helpful editorial comments by Scott Zeger, Peter Diggle, and an anonymous AE that significantly improved the manuscript presentation. The research of the first author was supported in part by grants from the National Institutes of Allergies and Infectious Diseases (R03AI068484) and Emory's Center for AIDS Research.

References

Blake, C. L. and Merz, C. J. (1997) UCI repository of machine learning databases. Available at `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

BÜHLMANN, P. and HOTHORN, T. (2007) Boosting algorithms: regularization, prediction, and model fitting (with Discussion). *Statistical Science* **4** 477–505.

CAI, T., HUANG, J. and TIAN, L. (2009) Regularized estimation for the accelerated failure time model. *Biometrics* (In press).

CHEN, K., SHEN, J. and YING, Z. (2005) Rank estimation in partial linear model with censored data. *Statistica Sinica* **15** 767–779.

DATTA, S., LE-RADEMACHER, J. and DATTA, S. (2007) Predicting survival from microarray data by accelerated failure time modeling using partial least squares and lasso. *Biometrics* **63** 259–271.

GEHAN, E. A. (1965) A generalized Wilcoxon test for comparing arbitrarily single-censored samples. *Biometrika* **52** 203–23.

HUANG, J., MA, S. and XIE, H. (2006) Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics* **62** 813–820.

JIN, Z., LIN, D. Y., WEI, L. J. and YING, Z. (2003) Rank-based inference for the accelerated failure time model. *Biometrika* **90** 341–353.

JOHNSON, B. A. (2008) Variable selection in semiparametric linear regression with censored data. *J. R. Statist. Soc. Ser. B* **70** 351–370.

JOHNSON, B. A., LIN, D. Y. and ZENG D. (2007) Penalized estimating functions and variable selection in semiparametric regression models. *J. Amer. Statist. Assoc.* **103** 672–680.

KALBFLEISCH, J. D. and PRENTICE, R. L. (2002) *The Statistical Analysis of Failure Time Data.* John Wiley: New York.

KNIGHT, K. and FU, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.*, **28**, 1356–1378.

PRENTICE, R. L. (1978) Linear rank tests with right-censored data. *Biometrika* **65** 167–179.

STREET, W. N., MANGASARIAN, O. L. and Wolberg, W. H. (1995) An inductive learning approach to prognostic prediction. In *Proceedings of the Twelfth International Conference on Machine Learning.* Morgan Kaufmann, San Francisco, CA.

TIBSHIRANI, R. J. (1996) Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288.

TSIATIS, A. A. (1990) Estimating regression parameters using linear rank tests for censored data. *Ann. Statist.* **18** 354–372.

WANG, H., LI, G. and JIANG, G. Robust regression shrinkage and consistent variable selection through the lad-lasso. *J. Business Econom. Statist.* **11** 1–6.

[*Received*]