View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Collection Of Biostatistics Research Archive

University of Pennsylvania
UPenn Biostatistics Working Papers

Year 2006 *Paper* 9

Censored Data Regression in High-Dimension and Low-Sample Size Settings For Genomic Applications

Hongzhe Li*

*University of Pennsylvania, hli@cceb.upenn.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/upennbiostat/art9

Copyright ©2006 by the author.

Censored Data Regression in High-Dimension and Low-Sample Size Settings For Genomic Applications

Hongzhe Li

Abstract

New high-throughput technologies are generating various types of high-dimensional genomic and proteomic data and meta-data (e.g., networks and pathways) in order to obtain a systems-level understanding of various complex diseases such as human cancers and cardiovascular diseases. As the amount and complexity of the data increase and as the questions being addressed become more sophisticated, we face the great challenge of how to model such data in order to draw valid statistical and biological conclusions. One important problem in genomic research is to relate these high-throughput genomic data to various clinical outcomes, including possibly censored survival outcomes such as age at disease onset or time to cancer recurrence. We review some recently developed methods for censored data regression in the high-dimension and low-sample size setting, with emphasis on applications to genomic data. These methods include dimension reduction-based methods, regularized estimation methods such as Lasso and threshold gradient descent method, gradient descent boosting methods and nonparametric pathways-based regression models. These methods are demonstrated and compared by analysis of a data set of microarray gene expression profiles of 240 patients with diffuse large B-cell lymphoma together with follow-up survival information. Areas of further research are also presented.

1 Censored Data Regression in High-Dimension and Low-sample Size Settings For Genomic Applications

Hongzhe Li

Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104-6021, USA
Email: hli@cceb.upenn.edu

Summary

New high-throughput technologies are generating various types of high-dimensional genomic and proteomic data and meta-data (e.g., networks and pathways) in order to obtain a systems-level understanding of various complex diseases such as human cancers and cardiovascular diseases. As the amount and complexity of the data increase and as the questions being addressed become more sophisticated, we face the great challenge of how to model such data in order to draw valid statistical and biological conclusions. One important problem in genomic research is to relate these high-throughput genomic data to various clinical outcomes, including possibly censored survival outcomes such as age at disease onset or time to cancer recurrence. We review some recently developed methods for censored data regression in the high-dimension and low-sample size setting, with emphasis on applications to genomic data. These methods include dimension reduction-based methods, regularized estimation methods such as Lasso and threshold gradient descent method, gradient descent boosting methods and nonparametric pathways-based regression models. These methods are demonstrated and compared by analysis of a data set of microarray gene expression profiles of 240 patients with diffuse large B-cell lymphoma together with follow-up survival information. Areas of further research are also presented.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

1.1 INTRODUCTION

High-throughput technologies are generating many types of high-dimensional genomic and proteomics data. Important examples include DNA microarray technology which permits simultaneous measurements of expression levels for thousands of genes, array-based comparative genomic hybridization (aCGH) data which measure the change of DNA copy numbers, array based single nucleotide polymorphism (SNP) data, and mass spectrometry data to measure protein expression levels. Such high-throughout genomic data offer the possibility of a powerful, genome-wide approach to the genetic basis of different types of tumors and can be used for molecular classification of cancers, for studying varying levels of drug responses in the area of pharmacogenomics and for predicting different patients' clinical outcomes. The problem of cancer class prediction using the gene expression data, which can be formulated as predicting binary or multi-category outcomes, has been studied extensively and has demonstrated great promise in recent years (e.g., Golub *et al.* , 1999; Sorlie *et al.* , 2001). There has also been active research of methods development in relating gene expression profiles to other phenotypes, such as quantitative continuous phenotypes or censored survival phenotypes such as time to cancer recurrence or time to death. Due to large variability in time to certain clinical events such as cancer recurrence among cancer patients, and in age of onset of many complex diseases, studying possibly censored survival phenotypes can be more informative than treating the phenotypes as binary or categorical variables.

The goal of linking genomic data to censored survival data is two-fold: to identify genes that are involved in the risk of a clinical event and to build a predictive model for future patients' survival based on both genomic data and patient-specific covariates. These two goals are related but not equivalent, although a good predictive model often implies that the variables used in the model are relatively important or predictive. Due to the problem of censoring, survival analysis models are obviously relevant to this problem. The Cox regression model (Cox, 1972) is the most popular method in regression analysis for censored survival data. Alternately, one can consider the accelerated failure time (AFT) model (Buckley and James, 1979; Wei, 1992) and the additive hazard model (Lin and Ying, 1994). For a given censored data regression model, due to the very high dimensional space of the predictors, i.e., the genes with expression levels measured by microarray experiments, the standard estimation method cannot be applied directly to obtain the parameter estimates. Besides the high-dimensionality, the expression levels of some genes are often highly correlated, which creates the problem of high co-linearity. Finally, we should also expect complex interactions between genes to affect the risk of survival. To deal with these problems, Li and Luan (2003), Li and Gui (2004), Li and Li (2004), Gui and Li (2005ab), Li and Luan (2005) were the first to investigate the use of penalized estimation procedures for the Cox model in the high-dimension and low-sample size settings. These regularized estimation methods were subsequently extended for the AFT models and the additive hazard models by Huang *et al.* (2005) and Ma and Huang (2005ab) by using appropriately defined loss functions.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

The focus of this review is to present some recently developed statistical and computational methods for relating high-throughout genomic data to censored survival outcomes, including both the methods for identifying genes related to such survival outcomes and the methods for building predictive models for future patients survival. The rest of the paper is organized as follows. We first review some commonly used censored data regression models. We then present a class of penalized estimation procedures for various models. We also present ensemble boosting methods for censored data regression models and briefly mention methods based on dimension-reduction and Bayesian variable selection. We present a comparison of some of these methods using a real data set of diffuse large B-cell lymphoma (DLBCL) survival times and gene expression data (Rosenwald *et al.* , 2002). Finally, we give a brief discussion of the methods and present several important problems for future research.

1.2 CENSORED DATA REGRESSION MODELS

Suppose that we have a sample size of n from which to estimate the relationship between the survival time T and the gene expression levels $X = \{X_1, \cdots, X_p\}$ of p genes. In addition, let Z be the vector of other patient-specific covariates. Due to censoring, for $i = 1, \cdots, n$, the ith datum in the sample is denoted by $(t_i, \delta_i, x_{i1}, x_{i2}, \cdots, x_{ip}, z_i)$, where δ_i is the censoring indicator and t_i is the survival time if $\delta_i = 1$ or censoring time if $\delta_i = 0$, and $x_i = \{x_{i1}, x_{i2}, \cdots, x_{ip}\}^{'}$ is the vector of the gene expression level of p genes for the ith sample. In this section, we briefly review the three most commonly used censored data regression models, including the Cox proportional hazards model, the accelerated failure time model and the additive hazard model.

1.2.1 The Cox proportional hazards model

The Cox proportional hazards model is the most commonly used censored data regression model in survival analysis. The model assumes the following hazard function for cancer recurrence or death at time t,

$$\begin{aligned} \lambda(t|X, Z) &= \lambda_0(t)\exp(F(X, Z)) \\ &= \lambda_0(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \gamma^{'} Z) \\ &= \lambda_0(t)\exp(\beta^{'} X + \gamma^{'} Z), \end{aligned} \tag{1.1}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function, $F(X, Z)$ is the function which links the (X, Z) to the hazard function. If this function is assumed to be linear, the $\beta = \{\beta_1, \cdots, \beta_p\}$ is the vector of the regression coefficients related to the p genomic data, and $X = \{X_1, \cdots, X_p\}$ is the vector of gene expression levels with the corresponding sample values of $x_i = \{x_{i1}, \cdots, x_{ip}\}$ for the ith sample. Finally, γ is the risk ratio parameter associated with covariate vector Z.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Based on the available sample data, the Cox's partial likelihood (Cox, 1972) can be written as

$$L(\beta,\gamma)=\prod_{r\in D}\frac{\exp(\beta^{'}x_r+\gamma^{'}z_r)}{\sum_{j\in R_r}\exp(\beta^{'}x_j+\gamma^{'}z_j)},$$

where D is the set of indices of the events (e.g., deaths) and R_r denotes the set of indices of the individuals at risk at time t_r-0. Note that when $p>n$, there is no unique β to maximize this partial likelihood function and therefore some regularization is required (see Sections 1.3.1, 1.3.2 and 1.3.3). Even when $p\leq n$, some regularization may still be required in order to reduce the variances of the estimates and to improve the prediction performance.

1.2.2 Accelerated failure time model

Let T be the random variable of time to event. For the ith individual, let t_i be the respective random variable. Let c_i be the censoring times, assumed to be $i.i.d$ and follows a survival function $G(t)=Pr(c_i>t)$. The linear AFT model assumes

$$g(T)=\alpha+\beta^{'}X+\gamma^{'}Z+\epsilon, \tag{1.2}$$

where g is some pre-specified monotone function (e.g., log function), ϵ is (heteroscedastic) unobservable error, assumed to be independent with zero means and bounded variances across n individuals. Due to censoring, for $i=1,\cdots,n$, the ith datum in the sample is denoted by $(y_i,\delta_i,x_{i1},x_{i2},\cdots,x_{ip})$, where δ_i is the censoring indicator, y_i is g transformation of the survival time if $\delta_i=1$ or g transformed of the censoring time if $\delta_i=0$, i.e.,

$$y_i=\min(g(t_i),g(c_i)),\quad \text{and } \delta_i=I[t_i\leq c_i],\ \ i=1,\cdots,n.$$

Wei (1980) discussed some advantages of using such AFT models over the popular Cox regression model, including easy interpretation of the model parameters and better fits for some data sets. One approach for estimating the parameter β is the Buckley and James (BJ) (1979) procedure. In Section 1.3.4, we present simple modification of the BJ procedure to deal with the problem of large p. Alternatively, one can estimate β by minimizing the inverse probability of censoring weighted (IPCW) loss function introduced in Robins and Rotnitsky (1992). Based on this loss function, one can develop regularized estimation procedures for β and extend the random forests and boosting procedure to censored survival data (see Section 1.4.2). For simplicity, we only consider model (1.2) without covariate Z.

1.2.3 Additive hazard regression models

The additive risk model as described in Lin and Ying (1994) assumes the following conditional hazard at time t,

$$\lambda(t|X,Z(.))=\lambda_0(t)+\beta^{'}X+\gamma^{'}Z(t), \tag{1.3}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

given a p-dimensional vector of genomic data X and patient-specific covariate $Z(.)$, which can be time-dependent. Here β, γ and $\lambda_0(t)$ denote the unknown regression parameter and the unknown baseline hazard function. In the following discussion, we simply assume that there is no covariate Z in the model (1.3). Denote $\{N_i(t) = I(t_i \leq t, \delta_i = 1); t \geq 0\}$ and $\{Y_i(t) = I(t_i \leq t); t \geq 0\}$ as the observed event process and the at-risk process. Lin and Ying (1994) proposed the following estimation equation for β,

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\infty} X\{dN_i(t) - Y_i(t)d\Lambda_0(\beta, t) - Y_i(t)\beta^{'} X dt\} = 0,$$

where

$$\Lambda_0 = \sum_{i=1} \int_0^t \frac{\{dN_i(u) - Y_i(u)\beta^{'} X_i du\}}{\sum_{i=1}^{n} Y_i(u)}$$

is the estimate of the baseline hazard function. As noted by Lin and Ying (1994) and Ma and Huang (2005), the resulting estimation of β is obtained by solving the equation

$$\left[\sum_{i=1}^{n} \int_0^{\infty} Y_i(t)\{X - \bar{X}(t)\}^{\oplus 2} dt\right] \beta = \left[\sum_{i=1}^{n} \int_0^{\infty} \{X - \bar{X}(t)\} dN_i(t)\right], \quad (1.4)$$

where $\bar{X}(t) = \sum_{i=1}^{n} Y_i(t)X_i / \sum_{i=1}^{n} Y_i(t)$. As noted by Ma and Huang (2005), the estimate of β by this equation is equivalent to minimizing a loss function of β (see Section 1.3.6). Based on this loss function, the Lasso or threshold gradient descent procedure can be developed for estimating the β in the additive hazard model (1.10).

1.3 REGULARIZED ESTIMATION FOR CENSORED DATA REGRESSION MODELS

In this section, we review several regularized estimation procedures for estimating the censored data regression models reviewed in previous sections. Most of these procedures are based on extensions of the procedures developed for linear regression and classification, with appropriate definitions of the loss functions.

1.3.1 L_2 penalized estimation of the Cox model using Kernel

Since the dimension of x_i vector is usually much larger than the sample size n, standard methods such as the Cox partial likelihood for estimating the unspecified function f is unfeasible. In addition, to deal with the problem of collinearity, the most popular approach is to use the penalized partial likelihood, including the L_2 penalized estimation, which is often called the ridge regression. Li and Luan (2003) investigated the L_2 penalized estimation of the Cox model in the high-dimensional low-sample size settings and applied their method to relate the gene expression profile

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

to survival data. To avoid the inversion of large matrix, they used the kernel tricks to reduce the computation to involve only the inversion of a matrix of the size of the sample size. Consider the model (1.1) with no covariate Z, a regularized formulation of the Cox regression is considered as a variational problem in reproducing kernel Hilbert space H,

$$min_{f\in H} R_{reg}(f) = \frac{1}{n}\sum_{i=1}^{n} V(t_i, \delta_i, f(x_i)) + \xi||f||_H^2,$$

where $V(t_i, \delta_i, f(x_i))$ is the loss function which is a function of f depending on only the values of $f(x)$ at the data points, $\{f(x_i)\}_{i=1}^n$. For the general Cox model (1.1), we propose to use the negative log partial likelihood as the loss function and reformulate the problem as finding function $f(x)$ such that

$$R_{reg} = -\frac{1}{n}\sum_{i=1}^{n} \delta_i[f(x_i) - \log\{\sum_{j\in R_i} \exp(f(x_j)\}] + \xi||f||_H^2 \tag{1.5}$$

is minimized, where $R_i = \{j = 1, \cdots, n, x_j \geq x_i\}$ is the set of individuals who were at risk at time x_i.

The solution to this problem was given by Kimeldorf and Wahba (1971), and is known as the representer theorem. By this theorem, the optimal $f(x)$ has the form:

$$f(x) = b + \sum_{i=1}^{n} a_i K(x, x_i) \tag{1.6}$$

where K is a positive definite reproducing kernel, which gives the inner product in the transform space. Since b can be absorbed into the baseline hazard function in model 1.1, we can omit b in the following discussion. For the simplest case of inner product kernel with $K(x_i, x_j) =< x_i, x_j >$, the function $f(x)$ can be expressed as a linear function of x_is. In the case when the data are not linearly separatable, one can choose a more general kernel such as the polynomial kernels with $K(x_i, x_j) = (< x_i, x_j > +1)^d$ or the Gaussian kernels with $K(x_i, x_j) = \exp(||x_i - x_j||/\sigma_d^2)$, where d and σ_d^2 are the kernel parameters. From the representer formula (1.6), it can be shown that minimizing equation (1.5) is equivalent to the finite dimensional form:

$$R_a = -\delta^{'}(K_a a) + \delta^{'}\log\{\sum_{j\in R_i} \exp(K_a a)\}] + \xi a^{'} K_q a, \tag{1.7}$$

where $a^{'} = (a_1, \cdots, a_n\}$, the regressor matrix $K_a = [K(x_i, x_j)]_{n\times n}$, and the regularization matrix $K_q = K_a$. Here the matrix K_a is called the kernel matrix. One can use the Newton-Raphson method to minimize the loss function over a, which is n-dimension.

This procedure can be simply modified to include other covariates Z. For example, we can estimate γ in model (1.1) by maximizing a profile partial likelihood.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

1.3.2 L_1 penalized estimation of the Cox model using least angle regression

One limitation of the L_2 penalized estimation of the Cox model is that it uses all the genes in the prediction and does not provide a way of selecting relevant genes for prediction. However, from a biological point of view, one should expect that only a small subset of the genes is relevant to predicting the phenotypes. Including all the genes in the predictive model introduces noises and is expected to lead to poor predictive performance. Due to the high-dimensionality, the standard variable selection methods such as stepwise and backward selection cannot be applied. Lasso was proposed by Tibshirani (1996) for variable selection for linear models and was further extended for variable selection for the Cox proportional hazard models (Tibshirani, 1997). Consider the model (1.1) with no covariate Z, let $l(\beta) = \log L(\beta)$ to be the log of the partial likelihood function, then the Lasso estimate of β (Tibshirani, 1996, 1997) can be expressed as

$$\hat{\beta}(s) = \text{argmax } l(\beta), \text{subject to } \sum_{j=1}^{p} |\beta_j| \leq s,$$

where s is a tuning parameter determining how many covariates with coefficients are zero.

Tibshirani (1997) proposed the following iterative procedure to reformulate this optimization problem with constraint as a Lasso problem for linear regression models. Specifically, let $\eta = \beta' X$, $\mu = \partial l/\partial \eta$, $A = -\partial^2 l/\partial \eta \eta^T$ and $z = \eta + A^- \mu$. With this reparameterization, a one-term Taylor series expansion for $l(\beta)$ has the form of

$$(z - \eta)^T A (z - \eta).$$

Although there are multiple choices of A^-, it is easy to show that if $rank(A) = n-1$, for any A^- that satisfies $AA^-A = A$ and $z = \eta + A^-\mu$, $(z-\eta)^T A(z-\eta)$ is invariant to the choice of the generalized inverse of A. The iterative procedure of Tibshirani (1997) involves the following four steps,

1. Fix s and initialize $\hat{\beta} = 0$.
2. Compute η, μ, A and z based on the current value of $\hat{\beta}$.
3. Minimize $(z - \beta' X)^T A (z - \beta' X)$ subject to $\sum |\beta_j| \leq s$.
4. Repeat steps 2 and 3 until $\hat{\beta}$ does not change.

Tibshrani (1997) proposed to use the quadratic programming for solving Step 3. However, in the high-dimension and low-sample size setting, i.e., in the case when $p >> n$, the quadratic programming algorithm cannot be directly applied. Gui and Li (2005) proposed a simple modification of the LARS algorithm of Efron *et al.* (2004) for Step 3. Specifically, Gui and Li (2005) apply the Choleski decomposition to obtain $T = A^{1/2}$ such that $T'T = A$, then Step 3 of the iterative procedure can be rewritten as

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Step 3: minimize $(y - \beta' \hat{X})^T (y - \beta' \hat{X})$ subject to $\sum |\beta_j| \leq s$,

where $y = Tz$ and $\hat{X} = TX$. This can be efficiently solved by using the LARS-Lasso procedure as presented in Efron *et al.* (2004).

Segal (2006) proposed to use LARS to minimize a residual-based loss function for the Cox model, in which the IRWLS iterations are not required. Park and Hastie (2006) proposed a generalization of the LARS algorithm for the Cox model using the predictor-corrector algorithm of convex optimization. Finally, for a given tuning parameter s, one can estimate γ in model (1.1) by maximizing a profile partial likelihood in γ. For a given γ and s, the LARS procedure can be used for estimating the β, denoted as $\beta(\gamma, s)$. Then we can maximize the partial likelihood over γ.

1.3.3 Threshold gradient descent procedure for the Cox model

Treating the negative log partial likelihood function ($-l(\beta)$) as the loss function, Gui and Li (2005) presented a threshold gradient descent (TGD) regularization procedure for estimating the β in the Cox model following the key idea presented in Friedman and Popescu (2004). The main idea of the TGD is that during the gradient descent minimization, a thresholding is imposed to the absolute values of the gradients. Specifically, for any threshold value $0 \leq \tau \leq 1$, the threshold gradient descent algorithm for Cox model involves the following five steps,

1. $\beta(0) = 0$, $\nu = 0$.
2. Calculate η, μ, $g(\nu) = \partial l / \partial \beta$ for the current β.
3. $f_j(\nu) = I[|g_j(\nu)| \geq \tau \cdot max_{0 \leq k \leq n} |g_k(\nu)|]$
4. Update $\beta(\nu + \triangle\nu) = \beta(\nu) + \triangle\nu \cdot g(\nu) \cdot f(\nu)$, $\nu = \nu + \triangle\nu$.
5. Repeat steps 2-4 until β converge.

This procedure involves two tuning parameters τ and ν, both of which control the sparsity of the estimates of β. Compared to the Lasso estimate of the Cox model, this TGD procedure is computationally fast and does not involve matrix inversion. Simulations and applications to real data sets indicated that when $\tau = 1$, the TGD procedure performs very similarly to the Lasso procedure. Note that this procedure is quite general and can be applied to essentially any convex loss function. Finally, if covariate Z is included in model (1.1), one can estimate γ by maximizing a profile partial likelihood or by iteratively updating γ and β during the TGD iterations.

1.3.4 Regularized Buckley-James Estimations for the AFT model

Buckley and James (1979) used the transformation ϕ on the observed responses y_i, where $\phi(y_i) = \delta_i y_i + (1 - \delta_i) E(t_i | t_i \leq y_i)$ and proposed to simultaneously update $\phi(y_i)$ and β at each step and proceed iteratively:

1. Select an initial estimate β_0, and let $\tilde{t}_i = \beta_0 x_i$.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

2. Compute the residuals $e_i = y_i - \tilde{t_i}$ and estimate transformation

$$\begin{aligned}\hat{\phi}(y_i) &= \delta_i y_i + (1-\delta_i)\left[\tilde{t}_i - \{\hat{S}_e(e_i)\}^{-1}\int_{e_i}^{\infty} sd\{\hat{S}_e(s)\}\right] \\ &= \delta_i y_i + (1-\delta_i)\left[\tilde{t}_i + \frac{\sum_k^u \nu_k \tilde{t}_k I(e_i < e_k)}{\hat{S}_e(e_i)^{-1}}\right]\end{aligned}$$

where $\hat{S}_e(s)$ is the Kaplan-Meier estimator of the survival function based on $\{\epsilon_i, \delta_i\}_{i=1}^n$, ν_k is the probability mass assigned to uncensored residual e_k, and $\sum^u$ denotes summation over uncensored values only.

3. Apply least-squares estimation to $\{\hat{\phi}(y_i), x_i\}$ and update β.

4. Stop if β converges or oscillates. Otherwise, go to step 2.

When $p > n$, one cannot implement the least-square estimation in Step 3 of the BJ procedure. However, one can perform LARS-Lasso or the threshold gradient descent procedure for Step 3 to obtain a regularized estimation of β. Alternatively, one can perform L_2 penalization or partial least squares (PLS) methods for estimating the β in Step 3 (Huang and Harrington, 2004), which provides a PLS procedure for linear models with censoring on the responses.

1.3.5 Regularization based on inverse probability of censoring weighted loss function for the AFT models

If there is no censoring in the data, the most commonly used method for estimating the model (1.3) is by minimizing a quadratic loss function

$$l(\beta) = \sum_{i=1}^{n}(y_i - \beta' x_i)^2,$$

over β. However, such a loss function cannot be evaluated at the censored observations. One solution to this problem is to use the inverse probability of censoring weighted (IPCW) loss function introduced in Robins and Rotnitsky (1992). Robins and Rotnitsky (1992) showed that for any loss function $l(T, F(x))$, one has

$$E(l(T, F(x))\Delta G(T|x)) = E(l(T, F(x))),$$

where the T is the random variable of time to event, $F(x)$ is an estimator, $G(T|x)$ is the survival function of the censoring variable, which may be dependent on x. This suggests the use of the following loss function to estimate the AFT model (1.2),

$$l_{ipcw}(\beta) = \sum_{i=1}^{n}\left[(y_i - \beta x_i)^2 \frac{\delta_i}{G(y_i)}\right], \tag{1.8}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

where $G(t)$ is the survival function of the censoring variable. This loss function can be regarded as the weighted squared loss function with weight $w_i = \delta_i/G(y_i)$ for the ith individual. In practice, $G(t)$ is of course unknown and needs to be estimated by the Kaplan-Meier estimator, $\hat{G(t)}$, from the observation (t_i, δ_i, x_i). Notice for the purpose of estimating $G(t)$, a $\delta_i = 0$ means a complete observation and $\delta_i = 1$ means a censored observation.

Alternatively, we can use the robust Huber (1964) loss function as

$$l^H_{ipcw}(\beta) = \sum_{i=1}^{n} l^H_i(y_i, x_i; \beta)\frac{\delta_i}{G(y_i)} \tag{1.9}$$

where $l^H_i(.)$ is the Huber loss function for the ith observation defined as

$$l^H_i(y_i, x_i, \beta) = \begin{cases} (y_i - \beta x_i)^2/2 & |y_i - \beta x_i| < \tau \\ \tau(|y_i - \beta x_i| - \tau/2) & |y_i - \beta x_i| < \tau \end{cases},$$

where τ is the transition point, and its value is often taken to be αth quartile of the current absolute residuals $\tau(\beta) = \text{quantile}_\alpha\{|y_i - \beta x_i|\}_{i\in D}$. Here $1-\alpha$ is a specified fraction of the observations that are treated as outliers, subject to absolute loss.

Based on the loss function defined by equation (1.8), Huang *et al.* (2005) developed L_1 penalized estimation or lasso by using the LARS, i.e., min $l_{ipcw}(\beta)$ subject to $\sum_{i=1}^{p}|\beta_i| < s$, and a threshold gradient descent procedure. For the Huber version of the loss function (1.9), one can similarly perform a gradient boosting procedure or the threshold gradient descent procedure (Friedman 2001; Friedman and Popescu, 2004). Finally, based on loss functions defined in equations (1.8) or (1.9), one can easily develop principal components or partial least square components analysis for the AFT models.

1.3.6 Penalized estimation for the additive hazard models

In the estimation equation (1.4) for β in the additive hazard model (1.3), we denote $H^i = \int_0^\infty Y_i(t)\{X_i - \bar{X}(t)\}^{\oplus 2}dt$, and $R^i = \int_0^\infty \{X_i - \bar{X}(t)\}dN_i(t)$, and $H^i_{s,l}$ as the (s,l) element of H^i and the sth components of R^i and β as R^i_s and β_s, then the equation (1.4) is equivalent to the following p equations:

$$\left(\sum_{i=1}^{n} H^i_{s,1}\right)\beta_1 + \cdots + \left(\sum_{i=1}^{n} H^i_{s,p}\right)\beta_p = \sum_{i=1}^{n} R^i_s, \;\; i = 1, \cdots, p.$$

Ma and Huang (2005ab) further note that the estimate defined by this equation is the same as minimizing the following loss function $L(\beta)$,

$$\beta = argmin_\beta \left\{ l(\beta) = \sum_{s=1}^{p}\left\{\left(\sum_{i=1}^{n} H^i_{s,1}\right)\beta_1 + \cdots + \left(\sum_{i=1}^{n} H^i_{s,p}\right)\beta_p - \sum_{i=1}^{n} R^i_s\right\}^2\right\}.$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Based on this loss function, Ma and Huang (2005ab) proposed to use the Lasso or the TGD procedure to obtain regularized estimates of β.

1.4 SURVIVAL ENSEMBLE METHODS

In recent years, ensemble methods such as random forests (Breiman, 1994) and boosting procedures (Freund, 1995; Freund and Schapire, 1997; Friedman, 2001; Bühlmann, 2003; Bühlmann and Yu, 2003) have gained much popularity in classification and linear regression analysis because of their superior predictive performances. In addition, Bühlmann (2003) demonstrated the applicability of the boosting procedure in the high-dimensional settings. In this section, we first review two extensions of the gradient boosting procedure to the Cox model and the AFT model.

1.4.1 The smoothing spline based boosting algorithm for the nonparametric additive Cox model

Li and Luan (2005) proposed to use the boosting procedure for estimating the function $F(X)$ in model (1.1) nonparametrically. Boosting essentially is an iterative procedure to update function estimators successively. Friedman (2001) developed a novel general framework, called "Gradient Boosting Machine," to obtain additive expansions adapted to any fitting criterion. The framework is quite general and works for various models. For linear regression with no censoring, Bühlmann and Yu (2003) show that the L_2 boosting achieves the optimal rate of convergence.

Following Friedman (2001) and Bühlmann and Yu (2003), Li and Luan (2005) proposed a component-wise boosting procedure using cubic smoothing splines as base learner. At the kth boosting step, they obtain the estimate of the function, $F^{(k)}(X)$, which is a nonparametric additive function of each component of X, some of which are identically zero. It should be noted that when the iteration k increases by 1, one more term is added to the fitted procedure, however, this term may have already been in the model. Due to the dependence of this new term on the previous terms, the complexity of the fitted model is not increased by a constant amount. The final model provides an estimate of possible nonlinear effects of gene expression levels on the risk of an event.

1.4.2 Random forests and gradient boosting procedure for the AFT models

Hothorn *et al.* (2006) presented a random forests algorithm and a gradient boosting algorithm for the construction of prognostic and diagnostic AFT models by using the IPCW loss function (1.8). Using the IPC weights, Hothorn *et al.* proposed to modify the original random forests procedure of Breiman (1994) in two ways: in the bootstrap (or bagging) step, the case samples are weighted by their IPC weights to obtain the case counts, and in the base learner step, the tree is built using the learner sample with case counts obtained from the bootstrap step. Similarly, using this IPCW loss function, the generic gradient descent boosting procedure of Friedman (2001) can be

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

directly applied to develop a boosting procedure for a linear model with censoring. The base learner can be regression tree, univariate splines or component-wise least squares. One benefit of using the component-wise least squares as base learner is that there is a closed form definition of AIC score, which can be used for selecting the boosting step.

1.5 NONPARAMETRIC PATHWAY BASED REGRESSION MODELS

For many complex diseases, especially for cancers, there are many types of meta-data available that are related to biological pathways. Currently, information derived from meta-data such as known biological knowledge has been used primarily to select promising candidates for genetic characterization and for studying gene-gene and gene-environment interactions. Such information has hardly been utilized in the modeling step for identifying such interactions or for identifying genes or pathways that are related to the phenotypes.

Wei and Li (2006) proposed a pathways-based boosting procedure for estimating a nonparametric pathway based regression models. Suppose that we have K pathways whose activities may be related to the phenotype of interest. Assume that there are p_k genes involved in the kth pathway. We allow that some genes belong to multiple pathways and let p be the total number of genes involved in the K pathways and therefore $p \leq \sum_{k=1}^{K} p_k$. Suppose that we have n independent individuals and let $y_i = (t_i, \delta_i)$, where t_i is time to event or censoring and δ_i is an event indicator. Let $x_{ij}^{(k)}$ be the genomic measurement of the jth gene in the kth pathway for the ith patient, $x_i^{(k)} = \{x_{i1}^{(k)}, \cdots, x_{ip_k}^{(k)}\}$ be the vector of the genomic measures of the genes in the kth pathway for the ith patient, and let $x_i = (x_i^{(1)}, \cdots, x_i^{(K)})$ be the vector of the genomic measurements of all the p genes. Here the genomic measurements can be SNP data or gene expression data. Our goal is to relate the phenotype data Y to $X = \{X^{(1)}, \cdots, X^{(K)}\}$ in order to identify the pathways that are related to the phenotype and to identify genes and their interactions that determine the pathway activities.

Here we assume that the phenotype is related to the total activity level across multiple pathways through an additive pathway activity function,

$$F(X) = \sum_{k=1}^{K} F_k(X^{(k)}), \tag{1.10}$$

where $F_k(X^{(k)})$ can be interpreted as the activity level associated with the kth pathway as determined by the genomic measurements of the p_k genes in this pathway. We assume that conditioning on the genes of the pathways, the pathway activities across the K pathways are additive. For the censored survival phenotype, we can assume that the hazard function at time t given the observed genomic data X is

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

modeled as

$$\lambda(t|X, Z) = \lambda_0(t) \exp(F(X) + \gamma Z), \tag{1.11}$$

where $\lambda_0(t)$ is the baseline hazard function, $F(X)$ is the pathway activity function as defined in (1.10), Z is a covariate vector and γ is the corresponding risk ratio parameter. The main motivation of these models is that we aim to model complex interactions between genes within pathways nonparametrically, rather than assume particular parametric forms for functions $F_k(X^{(k)})$. We use the term "nonparametric pathway-based regression" (NPR) to particularly emphasize this point, i.e., the genetic and pathways effects are modeled nonparametrically. It is obvious that without any constraints on the functions $F_k(X^{(k)})$, model (1.11) is not identifiable.

Wei and Li (2006) proposed a general pathway-based gradient descent boosting procedure to identify such NPR models with the particular form of (1.11). The key idea of our proposed extension of the boosting procedure of Friedman (2001) is that instead of performing gradient boosting over all the p genes, we perform gradient descent boosting over genes in each of the K pathways separately. We first consider the case when no other covariates are included in model (1.11). Let $L(y_i, F(x_i))$ be a loss function for the ith observation, which can be defined as negative of the partial likelihood based on model (1.11). During each of the boosting iterations, one pathway is picked that gives the best fit of the negative gradients using the base learner. This effectively utilizes the known pathway information and reduces the dimensionality from considering all the genes to only considering those genes in a given pathway. Then the functions are updated by adding the tree corresponding to the k^*th pathway selected. In order to model interactions between genes in a given pathway, Wei and Li (2006) proposed to use a J-terminal node regression tree (Breiman *et al.* , 1984) as the base learning procedure. The boosting procedure with regression trees as base procedures inherits the favorable characteristics of trees such as robustness and flexibility in modeling interactions (Breiman *et al.* , 1984). In addition, trees tend to be quite robust against the addition of irrelevant input variables and therefore serve as internal feature selection (Friedman, 2001; Breiman *et al.* , 1984). J controls the size of the tree, which is often chosen to be small.

1.6 DIMENSION-REDUCTION-BASED METHODS AND BAYESIAN VARIABLE SELECTION METHODS

There have also been some attempts to generalize the dimension-reduction procedures to censored survival data. Li and Gui (2004) and Park *et al.* (2003) generalized the partial least squares (PLS) method to the Cox model taking into account censoring. Li and Li (2004) extended the sliced inverse procedure to censored data. Similarly, one can also develop PLS procedures for the AFT models (Huang and Harrington, 2004) and the additive hazard models. One limitation of such extensions is that these procedures do not provide a rigorous way of selecting genes in the model. Given that we expect that only a small set of genes might be related to survival endpoints,

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

these procedures may introduce too much noise to the estimation, and therefore may have relatively low predictive performance. Bair *et al.* (2005) proposed a supervised principal components analysis, where genes are selected by univariate Cox regression analysis and the selected genes are used to defined several principal components. The number of genes and the number of components used in the final model are selected by cross-validation. While this is a step forward than the use of principal components analysis on all the genes, selecting genes by univariate analysis may not capture possible joint effects of genes.

Bayesian variable and model selection procedures for linear regression models and for clustering analysis (George, 2000; George and McCulloch, 1993; Tadesse *et al.*, 2005; West, 2003) in the high-dimensional settings can also be extended for censored data regression models (Tadesse, personal communication, 2006). However, we have seen publications of such extensions in the literature.

1.7 APPLICATION TO A REAL DATA SET AND COMPARISONS

We demonstrate the utility of some of these procedures using a published data set of DLBCL by Rosenwald *et al.* (2002). This data set includes a total of 240 patients with DLBCL, including 138 patient deaths during the follow-ups with median death time of 2.8 years. Rosenwald *et al.* divided the 240 patients into a training set of 160 patients and a validation set or test set of 80 patients and built a multivariate Cox model. The gene expression measurements of 7,399 genes are available for analysis.

For the Cox model, we applied several methods to build a predictive model using the training data set and we used zero as a cutoff point of the risk scores and divided the test patients into two groups based on whether they have positive or negative risk scores. Using the L_1 penalized estimation, the two groups of patients show very significant differences (p-value=0.0004) in overall survival between the high-risk group and low-risk group. We observe that the two risk groups defined by the LARS-Cox estimated model showed more significant differences in risk of death than the groups defined by the other three models: p-value of 0.0004 versus 0.003, 0.003 and 0.034 for the partial Cox regression method of Li and Gui (2004), the L_2 penalized method of Li and Luan (2003) and the supervised principal components analysis method of Bair and Tibshirani (2004), respectively Finally, the AUCs based on the risk scores estimated by the LARS-Cox procedure are also higher than those from the other three procedures.

Due to computational difficulty with the penalized estimation procedures for the AFT model and the additive hazards model, 1656 genes out of 7399 with large correlation coefficients (with the uncensored event times) were chosen for the AFT and the additive hazard model analysis. The results are summarized as follows: for the AFT model, the modified Lasso identified 37 genes and resulted in a test set p-value of 0.05, the TGD procedure identified 91 genes with a test set p-value of 0.776; for the additive hazard model, the modified Lasso selected 7 genes with a test set p-value of 0.331, and the TGD procedure identified 10 genes with a test set p-value of 0.13. These results indicate that at least for this particular data set, the

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

AFT or the additive hazard models did not provide as good predictive results as the Cox regression models.

Finally, analysis by Li and Luan (2005) using the splines-based boosting procedure indicated that some genes indeed show strong nonlinear effects on the risk of death from lymphoma.

1.8 DISCUSSION AND FUTURE RESEARCH TOPICS

It is clinically relevant and very important to predict a patient's time to cancer relapse or time to death due to cancer after treatment using gene expression profiles of the cancerous cells prior to the treatment. Powerful statistical methods for such prediction allow microarray gene expression data to be used most efficiently. Due to high-dimensionality of the genomic data, standard estimation and test methods for various censored data regression models cannot be applied directly to analyze such data. In this paper, we have reviewed the latest developed regularized estimation procedures for several classes of the most commonly used censored survival data regression models, including the Cox proportional hazards model, the accelerated failure time model and the additive hazard models. The methods reviewed include penalization estimation, threshold gradient-based regularization and gradient boosting procedures. These methods have been evaluated by simulation studies and have shown to be effective in identifying relevant genes and in building predictive models (Gui and Li, 2005; Li and Luan 2005; Ma and Huang, 2005ab).

Among the methods reviewed, most of the penalized estimation procedures are developed for censored data regression models with simple linear functional forms (i.e., $\beta^{'} X$). The kernel-based L_2 penalization (Li and Luan, 2003) and extensions of the boosting procedure or random forests to censored data regression (Li and Luan 2005; Tothorn *et al.*, 2006) allow for nonlinear effects and potential gene-gene interaction effects on the risk of event. In general, based on published results in the papers we reviewed, we observed that the methods with variable selection often perform better in prediction than the dimension-reduction-based procedures. In addition, we should expect that the ensemble methods such as boosting and random forests perform better in prediction than other methods, especially in the high-dimension and low sample size settings. However, we should not expect one model or method to always perform better than the others. One useful avenue of research is to comprehensively compare these methods by simulations and application to many different data sets. Besides empirical results, theoretical results are also required in order to gain insights into the methods and to provide theoretical basis for the methods proposed.

While the emphasis of this review is on the methods for identifying genes that are related to censored survival outcome and building predictive models for future patients' survival using gene expression, there are several other interesting topics related to censored data regression in the high-dimension and low sample size settings that deserve further research. We present in the following some of the problems and

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

possible solutions and some possible extensions of the methods presented in this paper.

1.8.1 Test of treatment effect adjusting for high-dimensional genomics data

Consider the clinical trial setting where a treatment effect is evaluated with time to clinical event as an endpoint. In standard analysis of the data obtained from the clinical trials, the treatment effect is often tested using the Cox model adjusting for other low-dimensional covariates. It is becoming common practice that high-dimensional genomic data are often collected for such clinical trials. How to adjust for the genomic heterogeneity when testing for the treatment effect deserves further research. For example, in model (1.1) where Z is the treatment indicator in randomized clinical trials, the null hypothesis is

$$\gamma = 0,$$

where the effect of genomic data β is treated as a high-dimensional nuisance parameter. A valid test for such a null hypothesis is required. A related problem is to identify a subset of patients who respond to treatment differently.

1.8.2 Development of flexible models for gene-gene and gene-environment interactions

Most of the models and methods reviewed in this paper assume a simple linear functional form to relate genomic data to the phenotypes. However, most phenotypes are expected to be affected by the interplay of different genes and environments and therefore simple linear functional form cannot capture the complexity of the genomic effects on phenotypes. Ensemble methods using trees offer one way of modeling potential interactions between the variables. However, new methods are required for identifying and assessing such complex interactions. This is especially challenging in the high-dimension and low sample size settings. For example, the patient rule induction method (PRIM) (Friedman and Fisher, 1999) provides an alternative to the tree method, which may capture the genomic interactions better.

1.8.3 Methods for other types of genomic data

The methods presented in this paper are mainly developed for microarray gene expression data, where the data structures are relatively simple. Since genes usually function in coordinated modules, it is often observed that the expression levels of some genes are highly correlated. Methods that can account for such clusters of genes in the models are expected to predict well. In addition, special features of other types of genomic data such as aCGH data, mass-spectrometry data and genome-wide SNP/haplotype data need to be accounted for when building predictive models. For example, if one wants to build a predictive model using aCHG data, one needs to account for local dependency of the measurements. Similarly, if one wants to identify SNPs that are related to censored survival phenotypes, one has to account for linkage

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

disequilibrium for the SNPs. Tibshirani *et al.* (2005) recently proposed a fused-Lasso procedure, which provides a way of accounting for such local dependency.

1.8.4 Development of pathway- and network-based regression models for censored survival phenotypes

Since genes and proteins almost never work alone, they interact with each other and with other molecules in highly structured but incredibly complex ways. Understanding this interplay of human genome and environmental influences is crucial to developing a systems understanding of human health and disease. An important avenue for future research is to develop methods that can incorporate known biological knowledge such as pathways and networks into statistical modeling in order to limit the search space for gene-gene and gene-environment interactions. Wei and Li (2006) presented an attempt to incorporate known biological pathways/networks information into the censored data regression model in order to reduce the dimensionality of the problem. However, how to best identify genes and pathways that are related to censored survival phenotypes clearly deserves future research.

1.8.5 Final remarks

High-throughput genomic and proteomic data provide a unique opportunity for dissecting genes and pathways that are related to risk of complex diseases or the responses to treatments. Due to variation of disease onset or time to clinical event, studying censored survival data can gain additional power in identifying genes and pathways involved. As user-friendly software packages implementing these methods become available, we should expect to see more applications of these methods in identifying genes and pathways involved in complex diseases. We should also expect more new method developments in this important area.

ACKNOWLEDGMENTS

This research was supported by NIH grant R01-ES09911 and a Pennsylvania Department of Health grant. I thank my students and postdocs (Dr. Yihui Luan, Dr. Gui Jiang and Zhi Wei) for implementing some of the ideas presented in this article, Dr. Shuange Ma for providing the analysis results using the AFT and the additive hazards model for the lymphoma data set, and Mr. Edmund Weisberg, MS for his editorial help.

REFERENCES

Bair E and Tibshirani R (2004): Semi-supervised methods for predicting patient survival from gene expression papers. *PLoS Biology*, 2, 5011-5022.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Breiman L (1994): Random Forests. *Machine Learning*, 45 (1), 5-32.

Breiman L, Friedman JH, Olshen RA and Stone CJ (1984): *Classification and Regression Trees*. Monterey: Wadsworth and Brooks/Cole.

Bühlmann P (2003): Boosting methods: why they can be useful for high-dimensional data. In *Proceedings of the 3rd International Workshop on Distributed Computing (DSC 2003).*

Bühlmann P and Yu B (2003): Boosting with the L_2-Loss: Regression and Classification. *Journal of American Statistical Association*, 98, 324-339.

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

Cox DR (1972): Regression models and life-tables. *Journal of the Royal Statistical Society. Series B*, 34,187-220.

Efron B, Johnston I, Hastie T and Tibshirani R (2004) Least angle regression. *Annals of Statistics*, 32, 407499

Freund Y (1995): Boosting a weak learning algorithm by majority. *Information and Computation*, 121, 256-285.

Freund Y and Schapire R (1997): A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55, 119-139.

Friedman J (2001): Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29, 1189-1232.

Friedman J and Fisher N (1999): Bump hunting in high dimensional data. *Statistics and Computing*, 9, 123-143.

Friedman J, Hastie T and Tibshirani R (2000): Additive logistic regression: a statistical view of boosting (with discussion). *The Annals of Statistics*, 28, 337-407.

Friedman JH and Popescu BE (2004): Gradient directed regularization. Technical Report, Stanford University.

George EI (2000): The Variable Selection Problem. *Journal of the American Statistical Association*, 95, 1304-1308.

George EI and McCulloch RE (1993): Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association*, 88, 881-889

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield, CD and Lander ES (1999): Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286,531-537.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Gui J and Li H (2005): Threshold Gradient Descent Method for Censored Data Regression, with Applications in Pharmacogenomics. *Pacific Symposium on Biocomputing*, 10,272-283.

Gui J and Li H (2005): Penalized Cox Regression Analysis in the High-Dimensional and Low-sample Size Settings, with Applications to Microarray Gene Expression Data. *Bioinformatics*, 21, 3001-3008.

Heagerty PJ, Lumley T and Pepe M (2000): Time dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 56, 337-344.

Huang J and Harrington D (2002): Penalized partial likelihood regression for right-censored data with bootstrap selection of the penalty parameter. *Biometrics*, 58, 781-791.

Huang J, Ma S and Xie H (2005): Regularized Estimation in the Accelerated Failure Time Model with High Dimensional Covariates. Technical report, University of Iowa. 2005.

Huber P (1964): Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 53, 73-101.

Hothorn T, Buhlmann P, Dudoit S, Molinaro AM, and van der Laan MJ (2006): Survival Ensembles. *Biostatistics*, in press.

Kimeldorf GS and Wahba G (1971): A correspondence between Bayesan estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics*, 2, 495-502.

Li H and Gui J (2004): Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics*, 20(suppl 1), i208-i215.

Li H and Luan Y (2003): Kernel Cox regression models for linking gene expression profiles to censored survival data. *Pacific Symposium of Biocomputing*, 8, 65-76.

Li H and Luan Y (2005): Boosting Proportional Hazards Models Using Smoothing Splines, with Applications to High-Dimensional Microarray Data. Bioinformatics, 21, 2403-2409.

Li L and Li H (2004): Dimension Reduction Methods for Microarrays with Application to Censored Survival Data. Bioinformatics, 20, 3406-3412.

Lin DY and Ying Z (1994): Semiparametric Analysis of the Additive Risk Model. *Biometrika*, 81, 61-71.

Ma S and Huang J (2005a): Threshold gradient descent regularization in the additive risk model with high-dimensional covariates. Tech report #346, University of Iowa.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Ma S and Huang J (2005b): Lasso methods for additive risk models with high-dimensional covariates. Tech report #347, University of Iowa.

Park PJ, Tian L and Kohane IS (2002): Linking expression data with patient survival times using partial least squares. *Bioinformatics*, 18, S120-127.

Park MY and Hastie T (2006): An L1 regularization-path algorithm for generalized linear models. Technical report, Stanford University.

Robins J and Rotnitsky A (1992): *Recovery of information and adjustment for dependent censoring using surrogate markers*. Chapter Aids epidemiology, Methodological issues. Bikhauser.

Rosenwald A, Wright G, Chan W, Connors JM, Campo E, Fisher R, Gascoyne RD, Muller-Hermelink K, Smeland EB and Staut LM (2002): The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-Cell lymphoma. *The New England Journal of Medicine*, 346, 1937-1947.

Segal MR (2006): Microarray Gene Expression Data with Linked Survival Phenotypes: Diffuse Large-B-Cell Lymphoma Revisited. *Biostatistics*, in press.

Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC and Brown, PO, Botstein D, Eystein Lłnning P, Błrresen-Dale, A (2001): Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98, 10869-10874.

Tadesse MG, Sha N and Vannucci M (2005): Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100,602-617.

Tibshirani R (1995): Regression shrinkage and selection via the Lasso. *Journal of Royal Statistical Society B*, 58, 267-288.

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

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

Wei LJ (1992): The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine*, 11, 1871-1879.

Wei Z and Li H (2006): Nonparametric Pathway-Based Regression Models for Analysis of Genomic Data. UPenn Biostatistics Working Papers. UPenn Biostatistics Working Paper Series. Working Paper 6. http://www.biostatsresearch.com/upennbiostat/papers/art6.

West M (2003): Bayesian factor regression models in the "large p, small n" paradigm. Bayesian Statistics, 7, 723-732.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive