Published in final edited form as:
Lifetime Data Anal. 2010 April ; 16(2): 176–195. doi:10.1007/s10985-009-9144-2.

Variable selection in the accelerated failure time model via the bridge method

Jian Huang and
Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242, USA

Department of Biostatistics, University of Iowa, Iowa City, IA 52242, USA

Shuangge Ma
Department of Epidemiology and Public Health, Yale University, New Haven, CT 06520, USA

Jian Huang: jian-huang@uiowa.edu

Abstract

In high throughput genomic studies, an important goal is to identify a small number of genomic markers that are associated with development and progression of diseases. A representative example is microarray prognostic studies, where the goal is to identify genes whose expressions are associated with disease free or overall survival. Because of the high dimensionality of gene expression data, standard survival analysis techniques cannot be directly applied. In addition, among the thousands of genes surveyed, only a subset are disease-associated. Gene selection is needed along with estimation. In this article, we model the relationship between gene expressions and survival using the accelerated failure time (AFT) models. We use the bridge penalization for regularized estimation and gene selection. An efficient iterative computational algorithm is proposed. Tuning parameters are selected using V-fold cross validation. We use a resampling method to evaluate the prediction performance of bridge estimator and the relative stability of identified genes. We show that the proposed bridge estimator is selection consistent under appropriate conditions. Analysis of two lymphoma prognostic studies suggests that the bridge estimator can identify a small number of genes and can have better prediction performance than the Lasso.

Keywords

Bridge penalization; Censored data; High dimensional data; Selection consistency; Stability; Sparse model

1 Introduction

High throughput technologies make it possible to identify genomic markers that are associated with disease development and progression. Gene profiling studies have been extensively conducted using microarrays. Identification of genomic markers from analysis of microarray data may lead to a better understanding of the genomic mechanism beneath disease development and assist future clinical diagnosis and prognosis. Among many disease outcomes measured in microarray studies, censored disease-free or overall survival has attracted much attention. See Alizadeh et al. (2000), Rosenwald et al. (2003), and Dave et al. (2004) for representative examples. Because of the high dimensionality of gene expression data, standard survival analysis techniques cannot be directly used. In addition, among the thousands of genes surveyed, only a subset may be associated with disease. Thus, gene selection is needed along with survival model construction.

When analyzing censored survival data with microarray gene expression measurements, the Cox proportional hazards model and the additive risk model have been adopted (Gui and Li 2005; Ma and Huang 2007). An alternative to those models is the accelerated failure time (AFT) model. Unlike the Cox and additive models, the AFT model is a linear regression model, in which logarithm (or in general a known monotone transformation) of the failure time is directly regressed on gene expressions (Kalbfleisch and Prentice 1980). Compared with the Cox and additive models, the AFT model has an intuitive linear regression interpretation (Wei 1992). In this article, we apply the method of Zhou (1992) and Stute (1993), which uses the Kaplan–Meier weights to account for censoring and has a weighted least squares loss function. The simple form of the loss function makes this estimation approach especially suitable for high dimensional data.

To tackle the high dimensionality problem of gene expression data, various dimension reduction or variable selection techniques have been employed. Previously used dimension reduction techniques include principal component analysis, singular value decomposition, partial least squares, and others. Among the many variable selection techniques developed, penalized selection has attracted extensive attentions. Penalization methods put penalties on the regression coefficients. By properly balancing goodness of fit and model complexity, penalization approaches can lead to parsimonious models with reasonable fit.

The most famous example of penalization methods is the Lasso (Tibshirani 1996), which has been used in gene expression analysis with survival data (Gui and Li 2005; Ma and Huang 2007; Wang et al. 2008). However, it has been shown that the Lasso is in general not variable selection consistent (Leng et al. 2006). Various penalization methods that can have consistent selection have been proposed. Examples include the adaptive Lasso and the SCAD. Another penalty that also enjoys consistent selection is the bridge penalty. Under conventional setup, i.e., when the number of observations is much larger than the number of covariates, the bridge penalty has been investigated. See for example Fu (1998). Huang et al. (2008a) shows that the bridge penalty can have the oracle estimation and selection properties in linear regression models with a divergent number of covariates.

In this article, we consider genomic studies where gene expressions are measured along with censored disease survival. The bridge penalization approach is used for regularized estimation and gene selection. The rest of the article is organized as follows. The AFT model and bridge estimation are introduced in Sect. 2. An efficient computational algorithm is proposed in Sect. 3. Resampling based methods are proposed to evaluate prediction performance and relative stability of selected genes in Sect. 4. Asymptotic selection consistency is established in Sect. 5. Analysis of two lymphoma studies are provided to illustrate the proposed method in Sect. 6. The article concludes with discussions in Sect. 7. Proofs are given in the Appendix.

2 Bridge estimation in the AFT model

Let T_i be logarithm of the failure time and X_i be the p-dimensional gene expressions for the ith subject in a random sample of size n. The AFT model assumes

$$T_i=\alpha+X_i'\beta+\varepsilon_i,\ i=1,\ldots,n, \tag{1}$$

where α is the intercept, $\beta \in \mathrm{IR}^p$ is the regression coefficient, and ε_i is the error term. When T_i is subject to right censoring, we observe (Y_i, δ_i, X_i), where $Y_i = \min\{T_i, C_i\}$, C_i is logarithm of the censoring time, and $\delta_i = 1_{\{T_i \le C_i\}}$ is the censoring indicator.

Estimation in the AFT model with an unspecified error distribution has been studied extensively. The following two approaches have received special attentions. The first is the

Buckley-James estimator which adjusts censored observations using the Kaplan–Meier estimator (Buckley and James 1979); and the second is the rank based estimator motivated by the score function of the partial likelihood function (Ying 1993). Although they both perform well when there are a small number of covariates, with high dimensional gene expression data, both approaches have high computational cost.

A computationally more feasible approach is the weighted least squares (LS) approach (Zhou 1992; Stute 1993). Let $\hat{F}_n$ be the Kaplan–Meier estimator of F, the distribution function of T. It can be computed as $\widehat{F}_n(y)=\sum_{i=1}^{n} w_i 1\{Y_{(i)} \le y\}$. Here w_is are the jumps in the Kaplan–Meier estimator computed as $w_1=\frac{\delta_{(1)}}{n}$ and $w_i=\frac{\delta_{(i)}}{n-i+1}\prod_{j=1}^{i-1}\left(\frac{n-j}{n-j+1}\right)^{\delta_{(j)}}$, $i = 2, \ldots, n$. w_is have also been referred to as the Kaplan–Meier weights in Stute (1993). Here $Y_{(1)} \le \cdots \le Y_{(n)}$ are the order statistics of Y_i's, $\delta_{(1)}, \ldots, \delta_{(n)}$ are the associated censoring indicators, and $X_{(1)}, \ldots, X_{(n)}$ are the associated covariates. The weighted LS loss function is

$$\frac{1}{2}\sum_{i=1}^{n} w_i (Y_{(i)} - \alpha - X_{(i)}'\beta)^2.$$

We center $X_{(i)}$ and $Y_{(i)}$ with their w_i-weighted means, respectively. Let $\overline{X}_w=\sum_{i=1}^{n} w_i X_{(i)}/\sum_{i=1}^{n} w_i$ and $\overline{Y}_w=\sum_{i=1}^{n} w_i Y_{(i)}/\sum_{i=1}^{n} w_i$. Denote $X^*_{(i)}=(nw_i)^{1/2}(X_{(i)} - \overline{X}_w)$ and $Y^*_{(i)}=(nw_i)^{1/2}(Y_{(i)} - \overline{Y}_w)$. We can then rewrite the weighted LS loss function as

$$Q_n(\beta)=\frac{1}{2}\sum_{i=1}^{n}(Y^*_{(i)} - X^{*'}_{(i)}\beta)^2. \quad (2)$$

The bridge penalized objective function is

$$L_n(\beta)=Q_n(\beta)+\lambda\sum_{j=1}^{p}|\beta_j|^{\gamma}, \quad (3)$$

where λ is a data dependent tuning parameter and γ > 0 is the bridge index. The value $\hat{\beta}_n$ that minimizes (3) is called the bridge estimator (Frank and Friedman 1993;Fu 1998).

The bridge estimator includes two important special cases. When γ = 2, it is the familiar ridge estimator, which does not have a "built-in" variable selection mechanism. When γ = 1, it is the Lasso estimator. In this article, we focus on the case with γ < 1.

3 Computation

3.1 Computational algorithm

Direct minimization of $L_n(\beta)$ is difficult, since the bridge penalty is not convex. An approximation approach is proposed in Huang et al. (2008a). As an alternative, we consider the following approach, which is more efficient and does not need any approximation.

For 0 < γ < 1, define

$$S_n(\beta,\theta)=Q_n(\beta)+\sum_{j=1}^{p}\theta_j^{1-1/\gamma}|\beta_j|+\tau_n\sum_{j=1}^{p}\theta_j, \quad (4)$$

where τ_n is a penalty parameter.

Proposition 1 *If* $\lambda=\tau_n^{1-\gamma}\gamma^{-\gamma}(1-\gamma)^{\gamma-1}$, *then* $\hat{\beta}_n$ *minimizes* $L_n(\beta)$ *if and only if* $(\hat{\beta}_n, \hat{\theta})$ *minimizes* $S_n(\beta, \theta)$ *subject to* $\hat{\theta}_j \geq 0$ *for* $j = 1, \ldots, p$.

This proposition can be proved as in Huang et al. (2009), in which a similar result is shown for the group bridge estimator in linear regression without censoring. Based on Proposition 1, we propose the following iterative algorithm for computing the bridge estimate in the AFT models.

1. Compute an initial estimate $\beta^{(0)}$. Specifically, we propose using the Lasso estimate, i.e, the minimizer of Eq. 3 with $\gamma = 1$.

   For $s = 1, 2, \ldots$

2. Compute $\theta_j^{(s)}=\left(\frac{1-\gamma}{\tau_n\gamma}\right)^{\gamma}|\beta_j^{(s-1)}|^{\gamma}$, $j = 1, \ldots, p$.

3. Compute $\beta^{(s)}=\arg\min_{\beta}\{Q_n(\beta)+\sum_{j=1}^{p}(\theta_j^{(s)})^{1-1/\gamma}|\beta_j|\}$.

4. Repeat Steps 2–3 until convergence.

The proposed algorithm always converges, since at each step the non-negative objective function (4) decreases. In our numerical studies, convergence is usually achieved within ten iterations. In Step 1, we choose the Lasso estimate as the initial value with the penalty parameter in the Lasso criterion determined by *V*-fold cross validation. Theorem 1 in Sect. 5 establishes that the Lasso tends to select all the important genes plus a few false positives. Thus, using the Lasso as the starting value will not miss any important genes. The main computational cost comes from Step 3, which computes a weighted Lasso estimate and can be achieved with many existing algorithms as such the LARS (Efron et al. 2004) or the boosting (Ma and Huang 2007). In this article, we adopt the boosting, since its computational cost is relatively insensitive to the number of genes.

3.2 Tuning parameter selection

We use *V*-fold cross validation to determine the tuning parameter λ. For a pre-defined integer *V*, partition the data randomly into *V* non-overlapping subsets with equal sizes. For a given λ, we define $CV\ score=\sum_{\upsilon=1}^{V}Q^{(\upsilon)}(\widehat{\beta}^{(-\upsilon)})$, where $\hat{\beta}^{(-\upsilon)}$ is the bridge estimator of β based on the data without the υ^{th} subset and $Q^{(\upsilon)}$ is the function defined in (2) evaluated on the υ^{th} subset. Optimal tuning is defined as the minimizer of the CV score. In this article, we set $V = 5$.

4 Evaluation

With gene expression data, $p \gg n$. Most of the conventional evaluation techniques are valid only under the $p \ll n$ scenario and cannot be applied here. In this study, we are most interested in two aspects: (1) prediction performance. That is, whether the identified genes and corresponding AFT models can make proper predictions for subjects not used in the model estimation; and (2) stability of identified genes. Early studies have shown that gene signatures identified from analysis of gene expression data may suffer from low reproducibility. That is,

genes identified using different data sets may differ significantly. Ideally, evaluation should be based on independent data, which is usually not available. As an alternative, we propose the following resampling based approaches.

4.1 Evaluation of prediction

We propose prediction evaluation based on random partitions as follows.

1. Partition the data randomly into a training set of size n_1 and a testing set of size n_2 with $n_1 + n_2 = n$. We use $n_1 = 2/3n$.
2. Compute the bridge estimate using the training set data. Cross validation is needed to select the optimal tuning for the training set.
3. Use the training set estimate to make predictions for subjects in the testing set. Specifically, we first compute the risk scores $X'\hat{\beta}$. We then dichotomize the risk scores at the median and create two risk groups (referred to as high and low risk groups respectively). Compare the survival functions for the two risk groups, and compute the logrank statistic.
4. To avoid over fitting caused by a "lucky" partition, we repeat Steps 1–3 $B = 500$ times. Each time a new partition is made and the value of the logrank statistic is computed.

We partition the dataset in Step 1. To generate a fair evaluation, we recompute tuning parameters for each individual partitions in Step 2. In Step 3, we adopt the logrank statistic as the evaluation measurement. A larger logrank statistic suggests that the high and low risk groups are better separated and the proposed approach is more effective. We create two risk groups, mainly because of the small sample sizes. By repeating the partitioning process many times, we can obtain a Monte Carlo estimation of the *distribution* of the logrank statistics (as opposed to a single logrank statistic in several early studies). We call it the *observed predictive distribution* (OPD) of the logrank statistic.

When $n \gg p$, the logrank statistic is asymptotically χ^2 distributed. With gene expression data and $n \ll p$, it is not clear how effective the χ^2 approximation is. To tackle this problem, we propose the following permutation based approach to generate the reference distribution for the OPD. We first randomly permute the event times together with the censoring indicators. We then follow the same procedure as for the OPD and obtain a Monte Carlo estimation of the distribution of the logrank statistic under permutation. We call it *permutation predictive distribution* (PPD) of the logrank statistic. With permutation, the event times and gene expressions are expected to be independent. The distribution of logrank statistics so computed can serve as the reference distribution for the OPD.

Calculations of the OPD and PPD are parallel: the OPD is calculated from the *observed* data, whereas the PPD is calculated from the *permuted* data. Well separated OPD and PPD may indicate that the propose approach can identify genes and models with satisfactory prediction performance, whereas substantially overlapped distributions suggest that either the proposed approach is not effective or the gene expressions do not have good discriminant power.

4.2 Evaluation of stability

The prediction evaluation described above assesses overall performance of the proposed approach and selected genes/models. In what follows, we evaluate the relative stability of each identified gene. The rationale behind the proposed approach is that, if a gene is more "important" or more "stable", it should be identified "more often" in analysis of multiple data sets. Since multiple independent data sets not available, we resort to random sampling again.

We first randomly sample $n_1 = 2/3n$ subjects. We then use the bridge approach to identify genes in the sampled subset. We repeat this procedure $B = 500$ times. For the jth gene, we count c_j, the number of times it is identified. The proportion $o_j = c_j/B$ gives a measure of the relative importance and stability of the jth gene, and will be referred to as the *observed occurrence index* (OOI). Following the same rationale as in the above section, we also permute the data and recompute the occurrence index, which will be referred to as the *permutation occurrence index* (POI) (since permutated data is used). The occurrence indexes are simply byproducts of the prediction evaluation and incur no additional computational cost.

5 Asymptotic properties

In this section, we investigate asymptotic properties of the proposed bridge approach with $p \gg n$. We are especially interested in the gene selection consistency property, because once genes are consistently selected, standard approaches can lead to consistent estimates. We note that for fixed p, the asymptotic results can be obtained easily using standard approaches. Since the case with fixed p is not relevant to our data applications, we will not consider it here.

We note that, with the proposed iterative algorithm, the Lasso estimate is used as the starting value. Genes not selected by the Lasso will not be selected in the final model. Thus, it is crucial to first establish properties of the Lasso estimate under the present data/model setup. Careful inspection of the proposed computational algorithm suggests that, once the initial estimate is obtained, in each step, an adaptive Lasso estimate is computed. Thus, we are able to use similar methods as in Huang et al. (2008b), which studies properties of the adaptive Lasso in high dimensional linear regression models, to establish properties of the bridge estimate.

We consider the rescaled $X^*_{(i)}$ and $Y^*_{(i)}$ defined in Sect. 2. For simplicity of notations, we use $X_{(i)}$ and $Y_{(i)}$ to denote $X^*_{(i)}$ and $Y^*_{(i)}$ hereafter. Let $Y = (Y_{(1)}, \ldots, Y_{(n)})'$. Let X be the $n \times p$ covariate matrix consisting of row vectors $X'_{(1)}, \ldots, X'_{(n)}$. Let $X_1, \ldots, X_p$ be the p columns of X. Let $W = \mathrm{diag}(nw_1, \ldots, nw_n)$ be the diagonal matrix of the Kaplan–Meier weights. For $A \subseteq \{1, \ldots, p\}$, let $X_A = (X_j, j \in A)$ be the matrix with columns X_j's for $j \in A$. Denote $\Sigma_A = X'_A W X_A / n$. Denote the cardinality of A by $|A|$.

Let $\beta_0 = (\beta_{01}, \ldots, \beta_{0p})'$ be the true value of the regression coefficients. Let $A_1 = \{j : \beta_{0j} \neq 0\}$ be the set of nonzero coefficients and let $q = |A_1|$. We make the following assumptions.

(A1) The number of nonzero coefficients q is finite.

(A2) (a) The observations (Y_i, X_i, δ_i), $1 \leq i \leq n$ are independent and identically distributed; (b) The errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed with mean 0 and finite variance σ^2. Furthermore, they are subgaussian, in the sense that there exist $K_1, K_2 > 0$ such that the tail probabilities of ε_i satisfy $P(|\varepsilon_i| > x) \leq K_2 \exp(-K_1 x^2)$ for all $x \geq 0$ and all i.

(A3) (a) The errors $(\varepsilon_1, \ldots, \varepsilon_n)$ are independent of the Kaplan–Meier weights $(w_1, \ldots, w_n)$; (b) The covariates are bounded. That is, there is a constant $M > 0$ such that $|X_{ij}| \leq M$, $1 \leq i \leq n$, $1 \leq j \leq p$.

(A4) The covariate matrix satisfies the sparse Riesz condition (SRC) with rank q^*: there exist constants $0 < c_* < c^* < \infty$, such that for $q^* = (3 + 4C)q$ and $C = c^*/c_*$, with probability converging to 1, $c_* \leq \dfrac{\nu' \Sigma_A \nu}{\|\nu\|^2} \leq c^*$, $\forall A$ with $|A| = q^*$ and $\nu \in \mathrm{IR}^{q^*}$ where $\|\cdot\|$ is the ℓ_2 norm.

By (A1), the model is sparse in the sense that although the total number of covariates may be large, the number of covariates with nonzero coefficients is still small. The tail probability assumption in (A2) has been made with high-dimensional linear regression models. See for example Zhang and Huang (2008). With assumption (A3), it can be shown that the subguassian tail property still holds under censoring. The SRC condition (A4) has been formulated in study of the Lasso with linear regressions without censoring (Zhang and Huang 2008). This condition implies that all the eigenvalues of any $d \times d$ submatrix of $X'WX/n$ with $d \leq q^*$ lie between c^* and c^*. It ensures that any model with dimension no greater than q^* is identifiable.

We first consider the Lasso estimator defined as $\tilde{\beta}=\arg\min\{Q_n(\beta)+\lambda\sum_{j=1}^{p}|\beta_j|\}$. With $\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)'$, let $\tilde{A}_1 = \{j, \tilde{\beta}_j \neq 0\}$ be the set of nonzero Lasso estimated coefficients.

Theorem 1 *Suppose that (A1)–(A4) hold and* $\lambda \geq O(1)\sqrt{n \log p}$. *Then*

- **i.** *With probability converging to 1,* $|\tilde{A}_1| \leq (2 + 4C)q$.
- **ii.** *If* $\lambda/n \to 0$ *and* $(\log p)/n \to 0$, *then with probability converging to 1, all the covariates with nonzero coefficients are selected.*
- **iii.** $\|\tilde{\beta} - \beta_0\|_2^2 \leq \frac{16\lambda^2 q}{n^2 c_*^2} + O_p\left(\frac{|\tilde{A}_1| \log p}{n c_*^2}\right)$. *In particular, if* $\lambda = O(\sqrt{n \log p})$, *then* $\|\tilde{\beta} - \beta_0\|_2^2 = O_p(\log p/n)$.

This theorem suggests that, with high probability, the number of covariates selected by the Lasso is a finite multiply of the number of covariates with nonzero coefficients. Moreover, all the covariates with nonzero coefficients are selected with probability converging to one. This justifies using the Lasso as the initial estimator in the algorithm proposed in Sect. 3.1. In addition, the Lasso estimator is estimation consistent.

Starting from the initial Lasso estimator $\tilde{\beta}$, we denote $\hat{\beta}$ as the estimate after one iteration (in the algorithm described in Sect. 3.1). Simple algebra shows that the value of $\theta_j^{(1)}$ computed in Step 2 of the proposed algorithm is $\theta_j^{(1)} = (\lambda/2)|\tilde{\beta}_j|^{-1/2}$. Thus Step 3 of the proposed algorithm is

$$\widehat{\beta}=\arg\min\left\{Q_n(\beta)+\frac{\lambda}{2}\sum_{j=1}^{p}|\tilde{\beta}_j|^{-1/2}|\beta_j|\right\}.$$

$\hat{\beta}$ computed above takes the form of an adaptive Lasso estimator. Of note, here, the penalty parameter is the same as the λ used in the Lasso estimator.

For any vector $x = (x_1, x_2, \ldots)$, denote its sign vector by $\mathrm{sgn}(x) = (\mathrm{sgn}(x_1), \mathrm{sgn}(x_2), \ldots)$ where $\mathrm{sgn}(x_i) = 1, 0, -1$ if $x_i > 0, = 0, < 0$, respectively.

Theorem 2 *Suppose that (A1)–(A4) are satisfied,* $(\log p)/n \to 0$, *and* $\lambda = O(\sqrt{n \log p})$. *Then*

$$P(sgn(\widehat{\beta}) = sgn(\beta_0)) \to 1.$$

The above theorem shows that the one-step estimator is sign consistent. Thus, the one-step estimator is selection consistent, in the sense that it can correctly distinguish covariates with zero and nonzero coefficients with probability converging to 1. Following similar arguments,

we can prove that any finite-step estimator (computed from the algorithm described in Sect. 3.1) is sign consistent and hence selection consistent. We note that, although the one-step estimator is selection consistent, our numerical studies suggest that iterating until convergence tends to improve finite sample performance.

We note that in Theorem 2, we allow $\log p = o(n)$ or $p = \exp(o(n))$. Thus the dimension of covariates can be larger than the sample size, which accommodates gene expression data.

6 Data analysis

6.1 Mantle cell lymphoma data

A study using microarray expression analysis of mantle cell lymphoma (MCL) is reported in Rosenwald et al. (2003). The primary goal of this study is to identify genes that have good predictive power of patients' survival risk. Among 101 untreated patients with no history of previous lymphoma, 92 were classified as having MCL based on established morphologic and immunophenotypic criteria. Survival times of 64 patients were available and the other 28 patients were censored. The median survival time was 2.8 years (range 0.02 to 14.05 years). Lymphochip DNA microarrays were used to quantify mRNA expression in the lymphoma samples from the 92 patients. The gene expression data that contains expression values of 8810 cDNA elements is available at http://llmpp.nih.gov/MCL/.

We model survival with the AFT model, and use the proposed bridge approach for gene selection. Although there is no limitation on the number of genes that can be used in the proposed approach, we pre-process the data as follows to exclude noises and gain further stability: (1) Un-supervised screening: compute the interquartile ranges of all gene expressions. Remove genes with interquartile ranges smaller than their first quartile. 6608 genes pass this screening; (2) Supervised screening: compute correlation coefficients of the uncensored survival times with gene expressions. Select 500 genes with the largest absolute values of the correlation coefficients. We then standardize these 500 gene expressions to have zero mean and unit variance. We note that the supervised screening utilizes the survival information. In the random sampling based evaluation, to guarantee a fair evaluation, the supervised screening needs to be conducted for each sampled data.

We employ the proposed approach and select the optimal tuning with 5-fold cross validation. Genes selected with the bridge approach and their corresponding estimates are shown in Table 1. For comparison, we also provide the Lasso estimate. 40 and 34 genes are identified using the Lasso and bridge approaches, respectively. Because of the special setup of the computational algorithm, genes identified using the bridge are a subset of those identified using the Lasso.

We evaluate prediction performance using the approach described in Sect. 4.1. For comparison, we also evaluate the Lasso approach using the same evaluation technique. We show in Fig. 1 (upper panels) the density estimates of OPD and PPD. We can see that (1) the bridge yields well separated OPD and PPD (Wilcoxon test p-value < 0.001), which suggests satisfactory prediction performance. The 90% percentile of the PPD is 3.32, and 62% of the logrank statistics from the OPD are larger than that value; (2) the PPD is close to the χ^2 distribution, but there is still very small discrepancy. The 95% and 90% percentiles of the PPD are 4.11 and 3.32, respectively, which are slightly larger than their counterparts (3.84 and 2.71) from the χ^2; (3) the prediction performance of Lasso is also satisfactory, but inferior compared to that of the bridge. The mean and median of the Lasso OPD are 4.617 and 3.616, which are smaller than their bridge counterparts (5.319 and 4.404).

We evaluate stability of identified genes using the approach described in Sect. 4.2. Results are shown in the lower panels of Fig. 1. For a better view, we only plot the 500 genes that pass the screening in the whole dataset. We can see that (1) most of the identified genes have relatively large OOI; (2) there are a few genes that are not identified, but have moderate OOI. It is still not clear why those genes are not identified. Such a question is worth investigating in future studies; and (3) with permutated data, the POI for all genes are small, with no genes having significantly larger occurrence indexes than others.

6.2 Diffuse large B-cell lymphoma data

The DLBCL (diffuse large B-cell lymphoma) data was first analyzed in Rosenwald et al. (2002). This dataset consists of 240 patients with DLBCL, including 138 patient deaths during the followup. Expression profiles of 7399 genes are obtained. Missing values are imputed using a K-nearest neighbors approach. We carry out supervised selection and select 500 genes with the largest absolute values of marginal correlation coefficients with the uncensored event times to gain further stability. Gene expressions are then normalized to have zero mean and unit variance. We note that, in the random sampling based evaluation, the supervised screening is conducted for each sampled data.

With the proposed approach and optimal tuning selected using 5-fold cross validation, 44 genes are identified. As a comparison, the Lasso identifies 46 genes. The UNI-QID, gene names, and bridge and Lasso estimated coefficients are shown in Table 2.

We evaluate prediction performance and show the results in Fig. 2. Similar conclusions as those in Sect. 6.1 can be drawn. With the bridge, the OPD has mean and median 4.416 and 3.674, respectively, which are larger than their Lasso counterparts (3.53 and 2.59). 61% of the logrank statistics from the OPD are greater than the 90% percentile of the PPD. The Wilcox test suggests that the OPD and PPD are well separated (p-value < 0.001). Evaluation of stability using the occurrence index is presented in Fig. 2 (lower panels). The observations are similar to those summarized in Sect. 6.1.

6.3 Remark

Analyses of the MCL and DLBCL data suggest that the bridge approach is capable of identifying a smaller number of genes than the Lasso. With gene expression data, a smaller number of identified genes means more focused hypothesis for future confirmation studies, and is thus preferred. In addition, prediction performance of the bridge is better than that of the Lasso. We note that, although prediction is not based on completely independent data, by properly using resampling and comparing the bridge and Lasso on the same basis, the prediction comparison is expected to be valid.

7 Discussions

Genomic studies with high dimensional markers measured along with censored survival outcomes are becoming more and more common. In this article, we model the relationship between gene expressions and censored survival with AFT models. AFT models have been commonly adopted and provide useful alternatives to the Cox and additive hazards models. Of note, since it is still not clear how to compare different models under the “large p, small n” setting, we do not pursue any model comparison. More methodological studies are needed before such a comparison can be conducted.

We propose using the bridge penalty for gene selection. Our numerical studies suggest that the bridge has better performance than the Lasso in terms of variable selection in AFT models. There are other penalties, for example the adaptive Lasso and SCAD, that can be used in the

present setup. Since it is beyond the scope of this paper to compare our proposed method with all the existing ones, we only pursue comparison with the Lasso, which has been commonly used as benchmark.

Acknowledgments

This work is partially supported by CA120988 from the National Cancer Institute and DMS 0805670 from the National Science Foundation. We thank the editors and reviewers for their helpful and constructive comments on an earlier version of the paper.

Appendix: Proofs

Let $\tau = (\tau_1, \ldots, \tau_n)'$ where $\tau_i = w_i\varepsilon_{(i)} \equiv w_i(Y_{(i)} - X'_{(i)}\beta_0)$.

Lemma 1 *Suppose that conditions (A2) and (A3) hold. Let* $\xi_j = \sum_{i=1}^{n} X_{ij}\tau_i$, $1 \le j \le p$. *Let* $\xi_n = \max_{1\le j\le p} |\xi_j|$. *Then*

$$\mathrm{E}(\xi_n) \le C_1 \sqrt{\log(p)}\left(\sqrt{2C_2 n \log(p)} + 4\log(2p) + C_2 n\right)^{1/2},$$

where C_1 *and* C_2 *are two positive constants. In particular, when* $\log(p)/n \to 0$,

$$\mathrm{E}(\xi_n) = O(1)\sqrt{n \log p}.$$

Proof of Lemma 1 Let $s_{nj}^2 = \sum_{i=1}^{n} X_{ij}^2$. Conditional on X_{ij}'s, assumptions (A2) and (A3) imply that ξ_j's are subgaussian. Let $s_n^2 = \max_{1\le j\le p} s_{nj}^2$. By (A2) and the maximal inequality for subgaussian random variables (Van der Vaart and Wellner 1996, Lemmas 2.2.1 and 2.2.2),

$$\mathrm{E}\left(\max_{1\le j\le p}\left|\xi_j\right| \middle| X_{ij}, 1 \le i \le n, 1 \le j \le p\right) \le C_1 s_n \sqrt{\log(p)},$$

for a constant $C_1 > 0$. Therefore,

$$\mathrm{E}\left(\max_{1\le j\le p}|\xi_j|\right) \le C_1\sqrt{\log(p)}\,\mathrm{E}(s_n). \tag{5}$$

Since

$$\sum_{i=1}^{n}\mathrm{E}[X_{ij}^2 - \mathrm{E}X_{ij}^2]^2 \le 4C_2 n, \tag{6}$$

and

$$\max_{1\le j\le p}\sum_{i=1}^{n}\mathrm{E}X_{ij}^2 \le C_2 n, \tag{7}$$

by Lemma 4.2 of Van de Geer (2008), (6) implies

$$\mathrm{E}\left(\max_{1\le j\le p}\left|\sum_{i=1}^{n}\{X_{ij}^2-\mathrm{E}X_{ij}^2\}\right|\right)\le \sqrt{2C_2n\log(p)}+4\log(2p).$$

Therefore, by (7) and the triangle inequality,

$$\mathrm{E}s_n^2\le \sqrt{2C_2n\log(p)}+4\log(2p)+C_2n.$$

Now since $\mathrm{E}s_n\le(\mathrm{E}s_n^2)^{1/2}$, we have

$$\mathrm{E}s_n\le\left(\sqrt{2C_2n\log(p)}+4\log(2p)+C_2n\right)^{1/2}. \tag{8}$$

The lemma follows from (5) and (8).

In the proofs below, let $Y^*=W^{1/2}Y$ and $X^*=W^{1/2}X$. Then

$$Q_n(\beta)=\frac{1}{2}\sum_{i=1}^{n}(Y_i^*-X_i^{*'}\beta)^2=\frac{1}{2}\|Y^*-X^*\beta\|^2,$$

where $\|\cdot\|$ is the ℓ_2 norm.

Proof of Theorem 1, part (i) Part (i) follows from the proof of Theorem 1 of Zhang and Huang (2008). The only difference is that here we use the subgaussian assumption to control certain tail probabilities, instead of the normality condition assumed in Zhang and Huang (2008). Since subgaussian random variables have the same tail behavior as normal random variables, the argument of Zhang and Huang goes through.

Proof of Theorem 1, part (ii) Part (ii) follows from part (iii) and the assumption that the number of nonzero coefficients is fixed. Thus the absolute values of the nonzero coefficients are bounded away from 0 by a positive constant independent of n.

Proof of Theorem 1, part (iii) By the definition of $\tilde{\beta}$,

$$\|Y^*-X^*\tilde{\beta}\|_2^2+2\lambda\sum_{j=1}^{p_n}|\tilde{\beta}_j|\le\|Y^*-X^*\beta_0\|^2+2\lambda\sum_{j=1}^{p_n}|\beta_{0j}|.$$

Thus

$$\|Y^*-X^*\tilde{\beta}\|_2^2+2\lambda\sum_{j\in A_1}|\tilde{\beta}_j|\le\|Y^*-X^*\beta_0\|^2+2\lambda\sum_{j\in A_1}|\beta_{0j}|.$$

This implies

$$\|Y^* - X^*\tilde{\beta}\|_2^2 - \|Y^* - X^*\beta_0\|^2 \le 2\lambda \sum_{j\in A_1} |\tilde{\beta}_j - \beta_{0j}|.$$

That is,

$$\|X^*(\tilde{\beta} - \beta_0)\|^2 - 2\tau' X^*(\tilde{\beta} - \beta_0) \le 2\lambda \sum_{j\in A_1} |\tilde{\beta}_j - \beta_{0j}|. \tag{9}$$

Let $B = A_1 \cup A_2 = \{j : \beta_{0j} \neq 0 \text{ or } \tilde{\beta}_j \neq 0\}$. Note that $|B| \le q_*$ with probability converging to 1 by part (i), where q^* is given in (A4). Denote $X_B^*=(X_j^*, j \in B)$, $\tilde{\beta}_B = (\tilde{\beta}_j, j \in B)$, and $\beta_{0B} = (\beta_{0j}, j \in B)$. Denote

$$\eta_B = X_B^*(\tilde{\beta}_B - \beta_{0B}).$$

Since $A_1 \subset B$,

$$\sum_{j\in A_1} |\tilde{\beta}_j - \beta_{0j}| \le \sqrt{|A_1|}\|\tilde{\beta}_{A_1} - \beta_{0A_1}\| \le \sqrt{|A_1|}\|\tilde{\beta}_B - \beta_{0B}\|. \tag{10}$$

By (9) and (10),

$$\|\eta_B\|^2 - 2\tau'\eta_B \le 2\lambda\sqrt{|A_1|}\|\tilde{\beta}_B - \beta_{0B}\|. \tag{11}$$

Let τ_B^* be the projection of τ to the span of X_B^*, i.e., $\tau_B^* = X_B^*(X_B^{*'}X_B^*)^{-1}X_B^{*'}\tau$. We have

$$\tau'\eta_B = \tau' X_B^*(\tilde{\beta}_{nB} - \beta_{0B}) = \left\{(X_B^{*'}X_B^*)^{-1/2}X_B^{*'}\tau\right\}' \left\{(X_B^{*'}X_B^*)^{1/2}(\tilde{\beta}_B - \beta_{0B})\right\}.$$

Therefore, by the Cauchy–Schwarz inequality,

$$2|\tau'\eta_B| \le 2\|\tau_B^*\| \cdot \|\eta_B\| \le 2\|\tau_B^*\|^2 + \frac{1}{2}\|\eta_B\|^2. \tag{12}$$

Combining (11) and (12),

$$\|\eta_B\|^2 \le 4\|\tau_B^*\|^2 + 4\lambda\sqrt{|A_1|} \cdot \|\tilde{\beta}_B - \beta_{0B}\| \tag{13}$$

By the SRC condition (A4), $\|\eta_B\|^2 \ge nc_*\|\tilde{\beta}_B - \beta_{0B}\|^2$. Thus (13) implies

$$nc_*\|\tilde{\beta}_B - \beta_{0B}\|^2 \le 4\|\tau_B^*\|^2 + \frac{(4\lambda\sqrt{|A_1|})^2}{2nc_*} + \frac{1}{2}nc_*\|\tilde{\beta}_B - \beta_{0B}\|^2.$$

It follows that

$$\|\tilde{\beta}_B - \beta_{0B}\|^2 \le \frac{8\|\tau_B^*\|^2}{nc_*} + \frac{16\lambda^2|A_1|}{n^2c_*^2}. \tag{14}$$

Now

$$\|\tau_B^*\|^2 = \|(X_B^{*'}X_B^*)^{-1/2}X_B^{*'}\tau\|^2 \le \frac{1}{nc_*}\|X_B^{*'}\tau\|^2 \le \frac{1}{nc_*}\max_{A:|A|\le q^*}\|X_A^{*'}\tau\|^2.$$

We have

$$\max_{A:|A|\le q^*}\|X_A^{*'}\tau\|^2 = \max_{A:|A|\le q^*}\sum_{j\in A}|X_j^{*'}\tau|^2 \le q^*\max_{1\le j\le p}|X_j^{*'}\tau|^2.$$

By Lemma 1,

$$\max_{1\le j\le p}|X_j^{*'}\tau|^2 = n\max_{1\le j\le p}|n^{-1/2}X_j^{*'}\tau|^2 = O_p(n\log p).$$

Therefore,

$$\|\tau_B^*\|^2 = O_p\left(\frac{q^*\log p}{c_*}\right). \tag{15}$$

The result follows from (14) and (15).

Proof of Theorem 2 The proof follows from the argument of Huang et al. (2008b). So we only provide the basic idea below. Let $a_j = |\tilde{\beta}_j|^{-1/2}/2$, $1 \le j \le p$. By the Karush–Kuhn–Tucker conditions, $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$ is the unique solution of the adaptive Lasso if

$$\begin{cases} X_j^{*'}(Y^* - X^*\widehat{\beta}) = \lambda_n a_j \operatorname{sgn}(\widehat{\beta}_j), & \widehat{\beta}_{nj} \ne 0 \\ |X_j^{*'}(Y^* - X^*\widehat{\beta})| \le \lambda_n a_j, & \widehat{\beta}_j = 0 \end{cases} \tag{16}$$

and the vectors $\{X_j^*, \widehat{\beta}_j \ne 0\}$ are linearly independent. Recall $A_1 = \{j : \beta_{0j} \ne 0\}$. Let $\tilde{\mathbf{s}}_{n1} = (a_j \operatorname{sgn}(\beta_{0j}), j \in A_1)'$ and $X_{A_1}^* = (X_j^*, j \in A_1)$, $\beta_{0A_1} = (\beta_j, j \in A_1)'$. So $X_{A_1}^*$ is a $n \times q$ matrix.

Define

$$\widehat{\beta}_{A_1} = \left(X_{A_1}^{*'}X_{A_1}^*\right)^{-1}(X_{A_1}^{*'}Y^* - \lambda_n\tilde{\mathbf{s}}_{n1}) = \beta_{0A_1} + C_{11}^{-1}(X_{A_1}^{*'}\tau - \lambda_n\tilde{\mathbf{s}}_{n1})/n, \tag{17}$$

where $C_{11} = X_{A_1}^{*'}X_{A_1}^*/n$. If $\operatorname{sgn}(\hat{\beta}_{A_1}) = \operatorname{sgn}(\beta_{0A_1})$, then the equation in (16) holds for $\tilde{\beta} = (\widehat{\beta}'_{A1}, \mathbf{0}')'$. Thus, since $X^*\widehat{\beta} = X_{A_1}^*\widehat{\beta}_{A_1}$ for this $\hat{\beta}$,

$$\operatorname{sgn}(\widehat{\beta})=\operatorname{sgn}(\beta_0) \quad \text{if} \begin{cases} \operatorname{sgn}(\widehat{\beta}_{A_1})=\operatorname{sgn}(\beta_{0A_1}) \\ \left|X_j^{*'}\left(Y^*-X_{A_1}^*\widehat{\beta}_{A_1}\right)\right| \le \lambda_n a_j, \forall j \notin A_1. \end{cases} \tag{18}$$

Let $H_n=I_n - X_{A_1}^* C_{11}^{-1} X_{n1}^{*'}/n$. It follows from (17) that
$Y^* - X_{A_1}^*\widehat{\beta}_{A_1} = \tau - X_{A_1}^*(\widehat{\beta}_{A_1} - \beta_{0A_1}) = H_n\tau + X_{A_1}^* C_{11}^{-1}\tilde{\mathbf{s}}_{n1}\lambda_n/n$, so that by (18),

$$\operatorname{sgn}(\widehat{\beta})=\operatorname{sgn}(\beta_0) \quad \text{if} \begin{cases} \operatorname{sgn}(\beta_{0j})(\beta_{0j}-\widehat{\beta}_j) \le |\beta_{0j}|, & \forall j \in A_1 \\ \left|X_j^{*'}\left(H_n\tau + X_{A_1}^* C_{11}^{-1}\tilde{\mathbf{s}}_{n1}\lambda_n/n\right)\right| < \lambda_n a_j, & \forall j \notin A_1. \end{cases} \tag{19}$$

Thus, by (19) and (17),

$$\begin{aligned} P\{\operatorname{sgn}(\widehat{\beta}) \ne \operatorname{sgn}(\beta_0)\} \le\ & P\left\{|\mathbf{e}_j' C_{11}^{-1} X_{A_1}^{*'}\tau|/n \ge |\beta_{0j}|/2 \text{ for some } j \in A_1\right\} \\ & + P\left\{|\mathbf{e}_j' C_{11}^{-1}\tilde{\mathbf{s}}_{n1}|\lambda_n/n \ge |\beta_{0j}|/2 \text{ for some } j \in A_1\right\} \\ & + P\left\{|X_j^{*'} H_n\tau| \ge \lambda_n a_j/2 \text{ for some } j \notin A_1\right\} \\ & + P\left\{|X_j^{*'} X_{A_1}^* C_{11}^{-1}\tilde{\mathbf{s}}_{n1}|/n \ge a_j/2 \text{ for some } j \notin A_1\right\} \\ =\ & P\{B_{n1}\} \\ & + P\{B_{n2}\} \\ & + P\{B_{n3}\} \\ & + P\{B_{n4}\}, \quad \text{say,} \end{aligned}$$

where $\mathbf{e}_j$ is the unit vector in the direction of the j-th coordinate. Therefore, to prove the theorem, it suffices to show that each probability in the last line converges to zero. The same argument as in Huang et al. (2008b) can be used here and is omitted. This completes the outline of the proof.

References

Alizadeh AA, Eisen MB, Davis RE, Ma C, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403:503–511. [PubMed: 10676951]

Buckley J, James I. Linear regression with censored data. Biometrika 1979;66:429–436.

Dave SS, Wright G, Tan B, et al. Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. New Engl J Med 2004;351:2159–2169. [PubMed: 15548776]

Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. Ann Stat 2004;32:407–499.

Frank IE, Friedman JH. A statistical view of some chemometrics regression tools (with discussion). Technometrics 1993;35:109–148.

Fu WJ. Penalized regressions: the bridge versus the Lasso. J Comput Graph Stat 1998;7:397–416.

Gui J, Li H. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics 2005;21:3001–3008. [PubMed: 15814556]

Huang J, Ma SG, Xie HL. Regularized estimation in the accelerated failure time model with high-dimensional covariates. Biometrics 2006;62:813–820. [PubMed: 16984324]

Huang J, Horowitz JL, Ma S. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Ann Stat 2008a;36:587–613.

Huang J, Ma SG, Xie HL, Zhang C-H. A group bridge approach for variable selection. Biometrika 2009;96:339–355. [PubMed: 20037673]

Huang J, Ma S, Zhang C. Adaptive Lasso for high-dimensional regression models. Stat Sinica 2008b; 18:1603–1618.

Kalbfleisch, JD.; Prentice, RL. The statistical analysis of failure time data. New York: John Wiley; 1980.

Leng C, Lin Y, Wahba G. A note on the LASSO and related procedures in model selection. Stat Sinica 2006;16:1273–1284.

Ma S, Huang J. Additive risk survival model with microarray data. BMC Bioinform 2007;8:192.

Rosenwald A, Wright G, Chan WC, Conners JM, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large B cell lymphoma. New Engl J Med 2002;346:1937–1947. [PubMed: 12075054]

Rosenwald A, Wright G, Wiestner A, Chan WC, et al. The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. Cancer Cell 2003;3:185–197. [PubMed: 12620412]

Stute W. Consistent estimation under random censorship when covariables are available. J Multivar Anal 1993;45:89–103.

Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc B 1996;58:267–288.

van de Geer S. High-dimensional generalized linear models and the Lasso. Ann Stat 2008;36:614–645.

Van der Vaart, AW.; Wellner, JA. Weak convergence and empirical processes: with applications to statistics. New York: Springer; 1996.

Wang S, Nan B, Zhu J, Beer DG. Doubly penalized Buckley-James method for survival data with high-dimensional covariates. Biometrics 2008;6:132–140. [PubMed: 17680828]

Wei LJ. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. Stat Med 1992;11:1871–1879. [PubMed: 1480879]

Ying ZL. A large sample study of rank estimation for censored regression data. Ann Stat 1993;21:76–99.

Zhang C, Huang J. The sparsity and bias of the Lasso selection in high-dimensional linear regression. Ann Stat 2008;36:1567–1594.

Zhou M. M-estimation in censored linear models. Biometrika 1992;79:837–841.

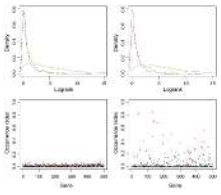

Fig 1.

MCL data. *Left-upper panel*: Lasso estimation. Green solid line: OPD; Red dash-dotted line: PPD; Black dotted line: density function of 500 χ^2 distributed random variables; *Right-upper penal*: Bridge estimation. Green solid line: OPD; Red dash-dotted line: PPD; Black dotted line: density function of 500 χ^2 distributed random variables; *Left-lower panel*: permuted occurrence index. Red "+" points correspond to genes identified using the bridge; *Right-lower panel*: observed occurrence index. Red "+" points correspond to genes identified using the bridge

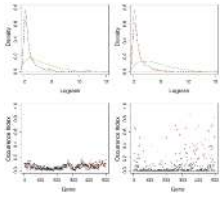

Fig 2.

DLBCL data. *Left-upper panel*: Lasso estimation. Green solid line: OPD; Red dash-dotted line: PPD; Black dotted line: density function of 500 χ^2 distributed random variables; *Right-upper penal*: Bridge estimation. Green solid line: OPD; Red dash-dotted line: PPD; Black dotted line: density function of 500 χ^2 distributed random variables; *Left-lower panel*: permuted occurrence index. Red "+" points correspond to genes identified using the bridge; *Right-lower panel*: observed occurrence index. Red "+" points correspond to genes identified using the bridge

Table 1

Mantle cell lymphoma data

UNIQID	Gene name	Lasso	Bridge
16541	Coagulation factor V (proaccelerin, labile factor)	0.180	0.133
16561	Aurora kinase B	−0.040	−0.040
16822	Chemokine (C-C motif) ligand 3	−0.032	−0.044
17174	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 11	−0.048	−0.069
17434	RAD51 homolog (RecA homolog, E. coli) (S. cerevisiae)	−0.023	−0.041
23972	Zinc finger protein, multitype 2	0.012	0.041
24376	Serine (or cysteine) proteinase inhibitor, clade A, member 9	0.167	0.126
24379	Immunoglobulin superfamily receptor translocation associated 1	0.082	0.102
24488	Eukaryotic translation elongation factor 1 alpha 1	0.176	0.106
24845	Interleukin 2 receptor, beta	0.135	0.109
24972		0.038	0.042
26192	Aldolase B, fructose-bisphosphate	0.124	0.061
26474	Chemokine (C-C motif) ligand 3	−0.027	−0.063
26475	Chemokine (C-C motif) ligand 3	−0.054	−0.054
27116	RAB interacting factor −0.007392391	−0.046	−0.061
27659	Hypothetical protein MGC61571	0.027	0.080
27678		0.020	0.036
27838	Lymphocyte antigen 64 homolog, radioprotective 105kDa	−0.069	
27969	CDC-like kinase 3	−0.055	
28075	Transforming, acidic coiled-coil containing protein 1	−0.019	−0.038
28638	BCL2-related protein A1	0.050	0.073
28645	EPH receptor B4	−0.029	
28990	Cell division cycle 2, G1 to S and G2 to M	−0.035	−0.021
29347	Split hand/foot malformation (ectrodactyly) type 1	0.017	
29357	Polymerase (DNA directed), epsilon 2 (p59 subunit)	−0.018	−0.031
29897	Asp (abnormal spindle)-like, microcephaly associated	−0.029	−0.027
30110	Natural killer-tumor recognition sequence	0.056	
30144	Natural killer-tumor recognition sequence	0.016	0.052
30284	PR domain containing 15	−0.031	−0.053
31049	Polymerase (DNA directed), theta	−0.053	−0.043
31081	Asp (abnormal spindle)-like, microcephaly associated	−0.036	−0.059
31101	Similar to CG1399-PB	−0.028	−0.082
32023	Membrane-spanning 4-domains, subfamily A, member 1	0.023	0.071
32187	AF15q14 protein	−0.024	−0.022
32830	Hypothetical protein LOC284019	−0.080	−0.108
32935	Zinc finger protein 148 (pHZ-52)	0.018	0.041
32979		0.036	0.073
33781	Chromosome 6 open reading frame 83	0.018	0.053
33851	GRIP and coiled-coil domain containing 2	0.044	

UNIQID	Gene name	Lasso	Bridge
33880	HP1-BP74	−0.062	−0.055

Genes identified using the Lasso and Bridge: UNIQID, gene names, and estimates

Table 2

DLBCL data

UNIQID	Gene name	Lasso	Bridge
16117	diacylglycerol kinase, delta (130 kD)	−0.155	−0.112
32446		−0.110	−0.101
26329		−0.030	−0.047
25933		0.031	0.074
29585	Epoxide hydrolase 2, cytoplasmic	0.081	0.067
28388	Bloom syndrome	0.044	0.127
17414	Occludin	−0.033	−0.074
30348	chromatin assembly factor 1, subunit A (p150)	−0.059	−0.071
17722	RAD23 homolog A (S. cerevisiae)	−0.088	−0.102
28837	G1 to S phase transition 1	−0.035	−0.064
34827	pM5 protein	−0.120	−0.094
24231	calnexin	−0.032	−0.059
33026	Sialyltransferase 7D (N-acetyl galactosaminide alpha-2,6-sialyltransferase)	−0.166	−0.148
29944	Solute carrier family 21 (organic anion transporter), member 12	−0.066	−0.100
28737	PAS domain containing serine/threonine kinase	0.081	0.114
34729	Forkhead box O1A (rhabdomyosarcoma)	0.233	0.167
34042	CD19 antigen	0.246	0.148
27704	Early B-cell factor	0.072	0.099
30355	Early B-cell factor	0.050	0.090
27681	G protein-coupled receptor 18	0.222	0.163
27341	Sarcoma amplified sequence	0.054	0.042
26231		−0.032	
26185		−0.019	−0.072
24400	Monoglyceride lipase	−0.068	−0.079
16636	Glucose regulated protein, 58 kD	−0.057	−0.056
28641	Osteoblast specific factor 2 (fasciclin I-like)	0.081	0.100
26081	Growth arrest-specific 1	0.044	0.063
26020	Melanoma cell adhesion molecule	0.075	0.095
19363	Lymphotoxin beta (TNF superfamily, member 3)	0.040	0.089
27509	Matrix metalloproteinase 9 (gelatinase B, 92 kD gelatinase)	0.040	0.101
24433		0.083	0.146
28415	PTK7 protein tyrosine kinase 7	−0.034	−0.076
16179	Lymphocyte-specific protein tyrosine kinase	0.040	0.090
17140	Protein tyrosine phosphatase, non-receptor type 2	0.060	
31728	Hypothetical protein FLJ00024	−0.057	−0.058
28681	CD58 antigen, (lymphocyte function-associated antigen 3)	0.034	0.070
17292		−0.052	−0.090
33912	Nuclear factor of kappa light polypeptide gene enhancer	−0.034	−0.088
29117	—Frizzled homolog 1 (Drosophila)	0.145	0.100
17182	Caspase 10, apoptosis-related cysteine protease	−0.035	−0.079

UNIQID	Gene name	Lasso	Bridge
16701	myosin, light polypeptide 2, regulatory, cardiac, slow	0.051	0.087
17391	tec protein tyrosine kinase	−0.121	−0.103
30130	Homo sapiens cDNA FLJ12727 fis, clone NT2RP2000027	−0.168	−0.121
23922		−0.034	−0.089
32836	ESTs	−0.168	−0.158
24612	immunoglobulin superfamily receptor translocation associated 1	0.050	0.093

Genes identified using the Lasso and Bridge: UNIQID, gene names, and estimates