Penalized Smoothed Partial Rank Estimator for the Nonparametric Transformation Survival Model with High-dimensional Covariates

Wei Dai*[1] and Yi Li†[2]

[1]Harvard University
[2]University of Michigan

May, 2013

Abstract

Microarray technology has the potential to lead to a better understanding of biological processes and diseases such as cancer. When failure time outcomes are also available, one might be interested in relating gene expression profiles to the survival outcome such as time to cancer recurrence or time to death. This is statistically challenging because the number of covariates greatly exceeds the number of observations. While the majority of work has focused on regularized Cox regression model and accelerated failure time model, they may be restrictive in practice. We relax the model assumption and and consider a nonparametric transformation model that makes no parametric assumption on either the transformation function or the error distribution. We propose a more flexible estimator, called penalized smoothed partial rank estimator, by regularizing the partial rank estimator with SCAD penalty. We also develop an efficient algorithm to obtain the whole solution path. Extensive simulations demonstrate the advantages of the proposal and the new method has been applied to a real genomic study.

KEYWORDS: nonparametric transformation model; survival analysis; variable selection; SCAD; multiple myeloma.

*wed942@mail.harvard.edu
†yili@umich.edu

1 Introduction

Microarray technology enables researchers to measure the expression levels for tens of thousands of genes in a tissue sample simultaneously, which has the potential to revolutionize our understanding of biological processes and diseases such as cancer. Traditionally, gene expression profiles have been used for predicting phenotypes such as cancer classification or responses to certain treatment. This problem can be formulated as predicting binary or multi-category outcomes and has been studied extensively in recent years (See Ma and Huang (2008) for a comprehensive review). When, in addition to gene expression data, (possibly censored) failure times are available, one may naturally want to link gene expression profiles to the survival outcomes such as time to cancer recurrence or time to death and to better understand the prognosis of cancer patients. Our work is motivated by such needs in multiple myeloma.

Multiple myeloma is the world's second-most common hematological cancer, characterized by excessive numbers of abnormal plasma cells in the bone marrow and overproduction of intact monoclonal immunoglobulin. Despite recent progress in treatment, there are still wide clinical and pathophysiologic heterogeneities. Patients experience survival periods of a few months to more than 10 years. Identifying important biomarkers associated with survival can help investigators better understand the molecular etiology of this disease and discover new therapeutic targets. Also building a risk score system with good predictive power could enable physicians to identify high risk patients and contribute to personalized medicine. In a clinical study with 170 multiple myeloma patients and tens of thousands of gene expression level measurements for each patient, classical methods as discussed below fail because the number of genes greatly exceeds the number of patients.

A common approach that yields a sparse model is penalized estimation, which modifies the minimization of a usual empirical risk function denoted by $\hat{R}_n(\theta)$ by adding a penalty $\lambda P(\theta)$ and thus minimize

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

where θ is the unknown parameter associated with outcome and λ is the tuning parameter that controls the degree of regularization. Tibshirani (1997) and Gui and Li (2005) developed regularized Cox regression methods by adding an L_1 LASSO penalty to the partial likelihood. Fan and Li (2002) extended the smoothly clipped absolute deviation (SCAD) penalty to Cox model. Meanwhile, regularization has also been applied to estimators under the AFT model, see Huang et al. (2006), Cai et al. (2009) and Li et al. (2010).

However, these models may be restrictive in practice. For example, the proportional hazards assumption for Cox regression model or the exponential risk form may not hold for certain applications. One feasible remedy is the nonparametric transformation model (Khan and Tamer, 2007), which postulates that a monotone increasing transformation of the failure time depends on the covariates through a linear model and makes no parametric assumptions on either the transformation function or the error distribution. This model is general

and includes the aforementioned Cox and AFT models as special cases.

Khan and Tamer (2007) proposed the partial rank (PR) estimator for the nonparametric transformation survival model. Their approach allows the censoring time to depend on the covariates as long as it is conditionally independent of the survival time given the covariates, and there is not restriction on the support of the censoring time. However, this estimator is based on maximization of a discontinuous function, making it impossible to compute when there are multiple covariates. Song and Ma (2007) improved the PR estimator by smoothing the discontinuous objection function. Their smoothed partial rank (SPR) estimator can be obtained using Newton-Raphson algorithm. However, the SPR estimator cannot handle the high-dimensional data encountered in microarray studies.

In this work, we build upon the SPR estimator and propose a penalized smoothed partial rank (PSPR) estimator for variable selection and estimation in the nonparametric transformation survival model with high-dimensional covariates. The proposed estimator maximizes a smoothed version of the SCAD-penalized objective function. The optimization is challenging because the objective function is nonconcave and the penalty is nonconcave and non-differentiable at zero. We propose an efficient solution path algorithm for computing the PSPR estimator.

The rest of this paper is organized as follows. In Section 2, we propose the PSPR estimator. In Section 3, we introduce the coordinate accent algorithm and discuss implementation issues. Simulation results are shown in Section 4. We illustrate a real application of the proposed method to build a risk score for multiple myeloma patients in Section 5 and some final remarks are provided in Section 6.

2 Methods

2.1 Model Definition

Let T denote the survival time, C denote the censoring time, and $\mathbf{Z}$ be a length p vector of covariates. Under right censoring, the observed survival data are $U = min(T, C)$ and $\delta = I(T \leq C)$. Assume that the survival time depends on the covariates through the nonparametric transformation model,

$$g(T) = \beta' \mathbf{Z} + e \tag{1}$$

where $g(\cdot)$ is an unspecified monotone increasing function, e is the error term with an unknown distribution independent of $\mathbf{Z}$, β is a length p vector of regression coefficients. This model is flexible and includes several popular models as special cases. For example, the Cox regression model is a special case of (1) with e following the standard extreme value distribution; the AFT model is a special case of (1) with $g(\cdot) = \log(\cdot)$.

The nonparametric transformation model is scale invariant and β is identifiable up to a scale constant. Without loss of generality, we assume the the first component of β is 1, i.e. the first covariate is the 'anchor variable'. In a clinical setting, the anchor is often chosen to be the treatment assignment. We suggest a simple way of selecting the anchor variable in Section 3.2.

2.2 Smoothed Partial Rank Estimator

Suppose the observed data $(U_i, \delta_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$, are independent and identically distributed as $(U, \delta, \mathbf{Z})$. The partial rank estimator (Khan and Tamer, 2007) maximizes the following objective function

$$\widetilde{R}_n(\beta) = \frac{1}{n(n-1)} \sum_{i \neq k} \delta_i I(U_i \leq U_k) I(\beta' \mathbf{Z}_i < \beta' \mathbf{Z}_k) \tag{2}$$

with the constraint that the first element of β equals 1.

However, the objective function $\widetilde{R}_n(\beta)$ is a weighted sum of indicator functions and hence discontinuous. Maximization is very difficult with multiple covariates, not to mention high-dimensional inputs. Song and Ma (2007) proposed to use the scaled sigmoid function

$$s_n(u) = \frac{1}{1 + \exp(-u/\sigma_n)} \tag{3}$$

to approximate the indicator function $I(u > 0)$. σ_n is a sequence of strictly positive and decreasing numbers satisfying $\lim_{n \to \infty} \sigma_n = 0$. Therefore, $\widetilde{R}_n(\beta)$ can be approximated by

$$R_n(\beta) = \frac{1}{n(n-1)} \sum_{i \neq k} \delta_i I(U_i \leq U_k)\, s_n(\beta'(\mathbf{Z}_k - \mathbf{Z}_i)) \tag{4}$$

The SPR estimator is obtained by maximizing (4) assuming the first variable is the anchor variable. Since $R_n(\beta)$ is a smoothing function of β, the computation of the SPR estimator can be accomplished through standard Newton-Raphson algorithm.

2.3 Penalized Smoothed Partial Rank Estimator

From a biological point of view, it is important to identify a small subset of genes that may be involved in the biological process determining the survival outcome. Extensive work has been done on simultaneous variable selection and estimation through penalization. For example, Tibshirani (1996) proposed the least absolute shrinkage and selection operator (LASSO) which minimizes the least squares with L_1 penalty defined as $P_\lambda(|\beta|) = \lambda|\beta|$. Frank and Friedman (1993) considered the L_q penalty, $P_\lambda(|\beta|) = \lambda|\beta|^q, (q > 0)$, which yields the bridge regression. Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) penalty, defined by

$$P'_\lambda(\beta) = \lambda \Big\{ I(\beta \leq \lambda) + \frac{(a\lambda - \beta)_+}{(a-1)\lambda} I(\beta > \lambda) \Big\},\ \beta \geq 0 \text{ for some } a > 2 \tag{5}$$

The SCAD has been shown to possess an oracle property with proper choice of regularization parameter. Namely, the true regression coefficients that are zero are automatically estimated as zero, and the remaining coefficients are estimated as well as if the correct submodel were known in advance.

Due to the aforementioned nice theoretical properties, we consider adding the SCAD penalty to (4) which leads to the following penalized smoothed partial rank (PSPR) estimator

$$\widehat{\beta} = \arg\max_{\beta} \Big\{ R_n(\beta) - \sum_{j=1}^{p} P_\lambda(|\beta_j|) \Big\} \tag{6}$$

where $P_\lambda(\cdot)$ is the SCAD penalty function with regularization parameter λ and p is the number of covariates. For identifiability, we take the first variable to be the anchor variable.

3 Computation and Implementation

3.1 Coordinate-Minorization-Ascent Algorithm

The optimization problem in (6) is challenging because $R_n(\beta)$ is nonconcave and the penalty is nonconcave and non-differentiable at zero. Recently, Zou and Li (2008) developed a new iterative algorithm based on local linear approximation (LLA) for maximizing the nonconcave penalized likelihood. However, LLA is not appropriate for nonconcave objective functions. To overcome the computation difficulties, we propose a new coordinate-minorization-ascent algorithm similar to that in Tseng (1988), Lange et al. (2000), Hunter and Lange (2004) and Xue et al. (2010).

To proceed, we rewrite the optimization problem as

$$\max_{\beta} \Big\{ R_n(\beta) - \sum_{j=1}^{p} P_\lambda(|\beta_j|) \Big\}$$

where

$$R_n(\beta) = \frac{1}{n(n-1)} \sum_{i \neq k} \frac{w_{ki}}{1 + e^{-\beta'(\mathbf{X}_k - \mathbf{X}_i)}}$$

where $w_{ki} = \delta_i I(U_i \leq U_k)$ and $\mathbf{X}_i = \mathbf{Z}_i/\sigma_n = (z_{i1}/\sigma_n, \ldots, z_{ip}/\sigma_n)'$ is the scaled covariate vector for the i-th observation.

Let $\widetilde{\beta}$ be the current estimate. The coordinate-ascent algorithm sequentially updates $\widetilde{\beta}_j$ by solving the following univariate optimization problem

$$\widetilde{\beta}_j \Leftarrow \arg\max_{\beta_j} \Big\{ R_n(\beta_j; \beta_{j'} = \widetilde{\beta}_{j'}, j' \neq j) - P_\lambda(|\beta_j|) \Big\} \tag{7}$$

However, the maximizer of (7) does not have a closed-form solution. To speed up computation, we propose to use the MM idea to derive a closed-form update that increase rather

than directly maximizing the objective function in (7).

Using a second order Taylor expansion and Theorem 1 from Zou and Li (2008), one can find a minorization function of the objective function in (7). Following the MM algorithm we show in Appendix that one can simply update

$$\widetilde{\beta}_j^{new} = S(\widetilde{\beta}_j + \frac{\widetilde{a}_j}{\widetilde{b}_j}, \frac{P'_\lambda(|\widetilde{\beta}_j|)}{\widetilde{b}_j}) \tag{8}$$

where $S(\cdot, \cdot)$ is the soft-thresholding operator (Tibshirani, 1996)

$$S(r,t) = sign(r)(|r| - t)_+$$

$$\widetilde{a}_j = \frac{1}{n(n-1)} \sum_{i \neq k} \frac{w_{ki}(x_{kj} - x_{ij}) e^{-\beta'(\mathbf{X}_k - \mathbf{X}_i)}}{(1 + e^{-\beta'(\mathbf{X}_k - \mathbf{X}_i)})^2}$$

$$\widetilde{b}_j = \frac{1}{6\sqrt{3}\, n(n-1)} \sum_{i \neq k} w_{ki}(x_{kj} - x_{ij})^2$$

To recap, the proposed algorithm is summarized as follows.

Step 1 Initialization of $\widetilde{\beta}$.

Step 2 Cyclically update $\widetilde{\beta}_j$ $(2 \leq j \leq p)$ via soft-thresholding $\widetilde{\beta}_j \Leftarrow S(\widetilde{\beta}_j + \frac{\widetilde{a}_j}{\widetilde{b}_j}, \frac{P'_\lambda(|\widetilde{\beta}_j|)}{\widetilde{b}_j})$

Step 3 Repeat Step 2 till convergence.

For any given λ, one may obtain a penalized estimator for β. Larger values of λ generate more sparse models. One is usually interested in models for more than one value of λ. Toward this end, we compute the solutions for a path of λ values. We begin with λ sufficiently large to set the estimated $\beta = 0$, and decrease λ until we arrive near the unregularized solution. Notice in the updating formula (8) that if $P'_\lambda(0) = \lambda \leq \widetilde{a}_j$, then $\widetilde{\beta}_j^{new} = 0$. Thus, we set the first λ to be

$$\lambda_{max} = \max_{j \neq 1} \left| \frac{1}{n(n-1)} \sum_{i \neq k} \frac{w_{ki}(x_{kj} - x_{ij}) e^{-(x_{k1} - x_{i1})}}{(1 + e^{-(x_{k1} - x_{j1})})^2} \right|$$

Note that λ_{max} is the smallest λ value that shrinks all covariate (except the anchor variable) estimates to zero.

This algorithm can also accommodate other types of penalty as long as the penalty function has finite derivative at zero. For example, to compute the LASSO-penalized estimator, one simply replace $P'_\lambda(|\widetilde{\beta}|)$ with λ.

3.2 Selection of the anchor variable

We identify the anchor variable as follows. Compute the c-statistics of the p variables on at a time,

$$c_j = \max \left\{ \frac{\sum_{i \neq k} \delta_i I(U_i \leq U_k) I(z_{ij} < z_{kj})}{\sum_{i \neq k} \delta_i I(U_i \leq U_k)}, \frac{\sum_{i \neq k} \delta_i I(U_i \leq U_k) I(z_{ij} > z_{kj})}{\sum_{i \neq k} \delta_i I(U_i \leq U_k)} \right\} \quad (9)$$

which is a scaled version of the partial rank objective function (2) for the model with only one variable. It measures the marginal association between each variable and the survival outcome under the nonparametric transformation model. The variable with the largest c-statistics is chosen as the anchor variable. For the anchor variable, if the c-statistics is attained at the first component of (9), $\beta_{(1)} = 1$, otherwise $\beta_{(1)} = -1$.

3.3 Selection of Tuning Parameters

The performance of the final model depends on the tuning parameter (λ, a) for the SCAD penalty and the scaling constant σ_n for the smoothing function. We follow Fan and Li (2001) and let $a = 3.7$. This value has been shown to work well in practice. Similar to Lin et al. (2011), we can approximate $R_n(\beta)$ with a quadratic form and regard the original problem as a kind of penalized least squares. Thus, we propose to select the optimal λ by maximizing

$$AIC_\lambda = \log(R_n(\hat{\beta})) - \frac{df_\lambda}{n}$$

where the degree of freedom df_λ can be approximated by the number of nonzero coefficient estimates. Although Wang et al. (2007) showed that the BIC type selector can identify the true model consistently, we find it places too heavy penalty for new addition of variables in our practice. The AIC criterion works well in our simulation studies.

The accuracy of the sigmoid approximation depends on the tuning parameter σ_n. Apparently a smaller σ_n may lead to better approximation thus more precise estimator. However, numerical studies also show that for extremely small σ_n, the maximization procedure may be unstable. A rule of thumb for choosing σ_n is to guarantee a majority of $|\beta'(\mathbf{X}_k - \mathbf{X}_i)/\sigma_n| > 5$ (Gammerman, 1996). We propose the following approach for choosing σ_n. Initialize $\sigma_n^0 = 1$ and construct the PSPR estimate $\hat{\beta}^0$ using the afore mentioned tuning scheme. Then let σ_n be the largest constant such that 95% of $|\hat{\beta}^{0\prime}(\mathbf{X}_k - \mathbf{X}_i)/\sigma_n|$ is greater than 5. Our simulation studies show that it works well.

4 Simulation Studies

In this section, we conduct simulation studies to assess the finite-sample performance of the PSPR estimator and compare it with l_1 penalized Cox proportional hazards model and regularized AFT model. Estimation and tuning for l_1 penalized Cox model can be implemented in R package *glmnet*. The optimal regularizing parameter is chosen using 10-fold cross validation. For regularized AFT model, we used the algorithm developed by Cai, Huang and

Tian (2008) and select the optimal tuning parameter following their suggestion, namely 5-fold cross validation. The run time for AFT model becomes inhibitive when sample size or number of covariates is moderate, which is the reason why we only include it in part of the comparison.

First, we investigate the performance of the proposed PSPR estimator when $p < n$. We simulate 100 datasets with $n = 100, 200$ observations, respectively, from the following model: $\log(T|\mathbf{Z}) = \beta_0'\mathbf{Z} + e$ where e follows the standard extreme value distribution. We assume that the first 5 out of a total of 50 variables are related to T through the above model and the nonzero coefficients are $2.1, 1.8, 1.9, 2.2$ and 2.7 (generated from an uniform$(1, 3)$ distribution). The covariates $\mathbf{Z} = (Z_1, \ldots, Z_{50})$ is generated from a multivariate normal with mean zero and variance covariace matrix $\Sigma = (\sigma_{jk})_{50\times 50} = (\rho^{|j-k|})$. We vary ρ to be $0, 0.5, 0.9$, respectively, to mimic the scenarios in which the correlation is weak, moderate and high.The censoring variable C was generated from uniform$[0, \delta]$, where δ was chosen to achieve about 40% of censoring.

Model performance is evaluated from the following aspects.

1. False Negative Rate (FNR): the average proportion of true nonzero variables being set to zero

2. False Positive Rate (FPR): the average proportion of true zero variables being set to nonzero

3. Model size: the average number of nonzero variables in the selected model

4. Correct model rate: the percentage of the method identifying the true model (no false positives or false negatives)

5. Wrong variable rate: the proportion of true zero variables among those being selected as nonzero

6. Bias: average of $(\hat{\beta} - \beta_0)'(\hat{\beta} - \beta_0)$. To remove the different scaling factors introduced by the selection of the anchor variable, we rescale $\hat{\beta}$ and β_0 such that they all have l_2 norm equals to 1.

7. AUC: average of $\dfrac{\sum_{i\neq k} \delta_i I(U_i \leq U_k) I(\hat{\beta}'\mathbf{Z}_i < \hat{\beta}'\mathbf{Z}_k)}{\sum_{i\neq k} \delta_i I(U_i \leq U_k)}$. If one plug in β_0 instead of $\hat{\beta}$, this provides an oracle reference as "the best one can do".

As shown in Table 1, when the correlation is not very high, the proposed PSPR method can identify the correct model with very high probability (over 92%) and estimate the coefficients with high accuracy. The FNR and FPR are almost 0 and the bias is very low. The AUC is over 0.93 and the difference in AUC between the PSPR model and the true model is smaller than 0.005. As sample size increases, the performance gets even better with 100% correct model rate, smaller bias and higher AUC. When we move to the difficult high correlation setting, the proposed approach can still achieve AUC as high as 0.958,

only 0.01 lower than the true model. Note that the FPR is still close to 0 which comes with the price of a relatively high FNR. In other words, it tends to miss some nonzero variables, which is expected to happen in order to keep FPR low when the correlation is very high among variables. The good news is that as sample size increases from 100 to 200, the FNR drops to 9% and PSPR method identifies the correct model nearly half of the times.

We also assess the model performance when the signal is weak (Table 2). The simulation settings are the same as before except that the nonzero coefficients are $1.05, 0.9, 0.95, 1.1$ and 1.35. The performance is still impressive with correct model rate about half and high AUC over 0.85 in low and moderate correlations. The correct model rate improves to about 90% as the sample size increases from 100 to 200.

Next we consider the scenarios when $p > n$. We adopt the same simulation settings as before except that there are 950 extra zero variables, bringing 995 zero variables in total. The results are presented in Table 3. The performance regarding identifying the correct model, estimating the coefficients and prediction is very good when the correlation is not high. When the sample size reaches 200, the proposed method can identify the correct model almost all the time. The performance in the presence of very high correlations is also acceptable.

Finally, we consider the cases when the "true" model is not a Cox or AFT model. We simulated 100 datasets with 100 observations, respectively, from the following models.

1. $3\Phi^{-1}(T|\mathbf{Z}) - 1 = \beta_0'\mathbf{Z} + e$ where e follows an uniform$(-2, 2)$ distribution.
2. $4\Phi^{-1}(T|\mathbf{Z}) = \beta_0'\mathbf{Z} + e$ where e is a random variable with probability 0.5 to equal 1 or -1.
3. $2\Phi^{-1}(T|\mathbf{Z}) = \beta_0'\mathbf{Z} + e$ where e follows a standard Laplace distribution with density function $f(x) = \frac{1}{2}e^{-|x|}$.

where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal cumulative distribution function and β_0 is set to be $1.05, 0.9, 0.95, 1.1, 1, 35$ plus 45 zero components. The results are shown in Table 4. The PSPR estimator is fairly robust as the performance is not affected by the change of underlying model assumptions. Cox and AFT models tend to pick a much larger model characterized by high false positive rates. This effect is more prominent when the correlations are not high.

Correctly identifying the anchor variable is crucial in our proposed approach. Here we also evaluate how our method performs regarding selecting an anchor variable whose true coefficient is nonzero. As we can see from Table $5 - 8$, regardless of sample size, correlation and signal strength, a variable with true nonzero coefficient is selected almost every time. Furthermore, the variable with the largest coefficient value is selected with highest probability when the correlation is low. This trend strengthens as the sample size increases. However, as the correlation becomes stronger, this effect is diluted as expected.

5 Application to a Multiple Myeloma Study

We study overall survival for 170 multiple myeloma patients enrolled in a clinical trial. The median followup was 58 months. During the study, a total of 76 deaths were observed, that is 55% of observations were censored. Subject gene expression levels were measured using Affymetrix Human Exon 1.0 ST Array before treatment. We randomly subset the 170 patients into a training set of 120 and a validation set of 50. We build the risk score using only the data from the training set.

Expression values were measured for 18708 genes and $\log_2$-transformed. Since the expression values were from the newer microarray technology whose noise level is low, we do not filter out any gene due to low signal to noise ratio. We identify RUNX1 as the anchor variable using the approach in Section 3.2. We keep the top 1000 genes with highest c-statistics and apply the proposed procedure to the 120 patients in the training set and tune the model as Section 3.3 describes. The final model select 10 genes besides RUNX1, which are presented in Table 11. The Cox model selects 18 genes, among which 4 are common with those selected by PSPR. The linear combination of the selected gene expression values and their parameter estimate can be used as a risk score to classify future patients.

To evaluate the predictive performance of the proposed risk score, we calculate the risk scores for the 50 patients in the validation set. The c-statistics for PSPR score and Cox score are 0.66 and 0.65, respectively, which suggests good predictive power. There is an apparent linear trend between the PSPR and Cox risk score (Figure 5) with Spearman correlation of 0.76 which is highly significant. In practice, clinicians are interested in identifying 10-20% of the patients with shorter survival for intensive treatment. Here we classify a patient to be of high risk if the predicted risk score exceeds the 15th percentile in the training set. The Kaplan-Meier curves for the high and low risk groups are shown in Figure 1. The separation is highly significant with hazard ratio of 3.72 and p-value 0.0034. When half of patients are classified as high risk, one can still see a significant effect of the grouping according to the predicted risk score (Figure 2) with hazard ratio 3.15 and p-value 0.00754. The Cox score does not have enough power to separate half high risk patients from the rest (Figure 4).

6 Conclusion

Our proposed PSPR estimator relaxes the model assumptions for popular methods used for high-dimensional survival analysis. Compared to other estimation and variable selection methods, ours poses minimal restrictions. Therefore, our estimator is more flexible and robust and should be preferred when the model assumptions cannot be justified. Based on the smoothed approximation and coordinate-accent and MM algorithms, we develop an efficient algorithm for the challenging optimization problem. Both simulation studies and real data application demonstrate the potential of this new method.

References

[1] Cai, T., Huang, J., and Tian, L. (2008). Regularized estimation for the accelerated failure time model. *Biometrics* **65,** 394–404.

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

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

[4] Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model an dfrailty model. *Annals of Statistics* **30,** 74–99.

[5] Frank, I.E. and Friedman, J.H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35,** 109–135.

[6] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularized paths for generalized linear models via coordinate descent. *Journal of statistical Software* **33,** 1–22.

[7] Gammerman, A. (1996). Computational Learning and Probabilistic Reasoning. New York: Wiley.

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

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

[10] Hunter, D. and Lange, K. (2004). A tutorial on MM algorithms. *American Statistician* **58,** 30-37.

[11] Hunter, D. and Li, R. (2005). Variable selection usgin mm algorithms. *Annals of Statistics* **33,** 1617–1642.

[12] Khan, S. and Tamer, E. (2007). Partial rank estimation of duration models with general forms of censoring. *Journal of Econometrics* **136,** 251–280.

[13] Lange, K., Hunter, D., and Yang, I. (2000). Optimization transfer using surrogate objective functions (with discussion). *Journal of Computational and Graphical Statistics* **9,** 1-59.

[14] Li, H. and Gui, J. (2004). Partial cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics* **20,** i208–i215.

[15] Li, Y., Dicker, L., and Zhao, S. (2010) A new class of Dantzig selectors for censored linear regression models. Harvard University Biostatistics Working paper Series.

[16] Lin, H., Zhou, L., Peng, H., and Zhou, X. (2011). Selection and combination of biomarkers using ROC method for disease classification and prediction. *The Canadian Journal of Statistics* **39,** 324-343.

[17] Ma, S. and Huang, J. (2008). Penalized feature selection and classification in bioinformatics. *Brief Bioinform* **9,** 392–403.

[18] Massy, W. (1965). Principal components regression in exploratory statistical research. *Journal of the American Statistical Association* **60,** 234-236.

[19] Nguyen, D. and Rocke, D. (2002). Partial least squares proportional harzard regression for application to DNA microarrays. *Bioinformatics* **18,** 1625–1632.

[20] Park, P., Tian, L., and Kohane, I. (2002). Linking expression data with patient survival times using partial least squares. *Bioinformatics* **18,** S120–S127.

[21] Song, X. and Ma, S. (2007). A semiparmetric approach for the nonparametric transformation survival model with multiple covariates. *Biostatistics* **8,** 197–211.

[22] Storey, J. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100,** 9440–9445.

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

[24] Tseng, P. (1988). Coordinate ascent for maximizing nondifferentiable concave functions. Technical report LIDS-P-1840, Laboratory for informaition and decision systems, Massachusetts Institute of Technology.

[25] Wang, H., Li, R., and Tsai, C. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94,** 553-568.

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

[27] Xue, L., Zou, H., and Cai, T. (2010). Non-concave penalized composite likelihood estimation of sparse ising models. Technical Report.

[28] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihhod models. *Annals of Statistics* **36,** 1509–1533.

Appendix

A.1 Derivation of the update formula

The proposed algorithm is a hybrid of the coordinate-ascent algorithm (Tseng, 1988) and the minorization-maximization (MM) algorithm (Lange et al., 2000; Hunter and Lange, 2004). Let $\widetilde{\beta}$ be the current estimate. The coordinate-ascent algorithm sequentially updates $\widetilde{\beta}_j$ by solving the following univariate optimization problem

$$\widetilde{\beta}_j \Leftarrow \arg\max_{\beta_j} \Big\{ R_n(\beta_j; \beta_{j'} = \widetilde{\beta}_{j'}, j' \neq j) - P_\lambda(|\beta_j|) \Big\} \tag{10}$$

However, the maximizer of (10) does not have a closed-form solution. Using some numerical optimization technique to find the exact maximizer is not ideal because it may need slow down the algorithm to reach convergence. To exploit the power of coordinate-ascent method, one can use the MM idea to derive a closed-form update that increase rather than maximize the objective function in (10).

Recall that

$$R_n(\beta) = \frac{1}{n(n-1)} \sum_{i \neq k} \frac{w_{ki}}{1 + e^{-\beta'(\mathbf{X}_k - \mathbf{X}_i)}}$$

$$\frac{\partial R_n(\beta)}{\partial \beta_j} = \frac{1}{n(n-1)} \sum_{i \neq k} \frac{w_{ki}(x_{kj} - x_{ij}) e^{-\beta'(\mathbf{X}_k - \mathbf{X}_i)}}{(1 + e^{-\beta'(\mathbf{X}_k - \mathbf{X}_i)})^2}$$

$$\begin{aligned} \frac{\partial^2 R_n(\beta)}{\partial \beta_j^2} &= \frac{1}{n(n-1)} \sum_{i \neq k} \frac{w_{ki}(x_{kj} - x_{ij})^2 e^{-\beta'(\mathbf{X}_k - \mathbf{X}_i)} (e^{-\beta'(\mathbf{X}_k - \mathbf{X}_i)} - 1)}{(1 + e^{-\beta'(\mathbf{X}_k - \mathbf{X}_i)})^3} \\ &\geq -\frac{1}{6\sqrt{3}\, n(n-1)} \sum_{i \neq k} w_{ki}(x_{kj} - x_{ij})^2 \end{aligned}$$

where $x_{ij} = z_{ij}/\sigma_n$, i.e., the j-th scaled covariate for the i-th observation.

Then by Taylor's expansion,

$$\begin{aligned} & R_n(\beta_j; \beta_{j'} = \widetilde{\beta}_{j'}, j' \neq j) \\ = \; & R_n(\widetilde{\beta}) + \frac{\partial R_n(\beta)}{\partial \beta_j}\Big|_{\beta = \widetilde{\beta}} (\beta_j - \widetilde{\beta}_j) + \frac{1}{2} \frac{\partial^2 R_n(\beta)}{\partial \beta_j^2}\Big|_{\beta = \alpha(\widetilde{\beta})} (\beta_j - \widetilde{\beta}_j)^2 \\ \geq \; & R_n(\widetilde{\beta}) + \widetilde{a}_j(\beta_j - \widetilde{\beta}_j) - \frac{1}{2} \widetilde{b}_j (\beta_j - \widetilde{\beta}_j)^2 \qquad (11) \\ \equiv \; & Q(\beta_j) \end{aligned}$$

where

$$\widetilde{a}_j = \frac{\partial R_n(\beta)}{\partial \beta_j}\Big|_{\beta = \widetilde{\beta}} = \frac{1}{n(n-1)} \sum_{i \neq k} \frac{w_{ki}(x_{kj} - x_{ij}) e^{-\beta'(\mathbf{X}_k - \mathbf{X}_i)}}{(1 + e^{-\beta'(\mathbf{X}_k - \mathbf{X}_i)})^2}$$

$$\widetilde{b}_j = \frac{1}{6\sqrt{3}\, n(n-1)} \sum_{i \neq k} w_{ki}(x_{kj} - x_{ij})^2$$

Since $P_\lambda(\cdot)$ is concave on $[0, +\infty)$,

$$P_\lambda(|\beta_j|) \leq P_\lambda(|\widetilde{\beta}_j|) + P'_\lambda(|\widetilde{\beta}_j|)\,(|\beta_j| - |\widetilde{\beta}_j|) \equiv L(|\beta_j|) \tag{12}$$

(11) and (12) together tells us that $Q(\beta_j) - L(|\beta_j|)$ is a minorization function of the objective function in (10). Following the MM algorithm one can update $\widetilde{\beta}_j$ by

$$\widetilde{\beta}_j \Leftarrow \arg\max_{\beta_j} \Big\{ Q(\beta_j) - L(|\beta_j|) \Big\}$$

After some algebra, one can get the simple update

$$\widetilde{\beta}_j^{new} = S(\widetilde{\beta}_j + \frac{\widetilde{a}_j}{\widetilde{b}_j}, \frac{P'_\lambda(|\widetilde{\beta}_j|)}{\widetilde{b}_j})$$

where $S(\cdot, \cdot)$ is the soft-thresholding operator (Tibshirani, 1996)

$$S(r, t) = sign(r)(|r| - t)_+$$

ρ	Sample Size	Method	FNR %	FPR %	Model Size	Correct Model %	Wrong Variable %	Bias	AUC
0		Truth							0.934 (0.002)
	100	PSPR	0	0	5.0	99	0.2	0.007	0.930 (0.003)
		Cox	0	32.8	19.8	0	73.5	3.980	0.924 (0.005)
		AFT	0	30.8	18.9	0	71.2	0.027	0.920 (0.008)
	200	PSPR	0	0	5.0	100	0	0.004	0.932 (0.002)
		Cox	0	39.2	22.6	0	76.8	3.992	0.929 (0.002)
0.5		Truth							0.955 (0.001)
	100	PSPR	0.8	0.1	5.0	92	0.9	0.019	0.952 (0.006)
		Cox	0	27.9	17.6	0	69.4	3.977	0.950 (0.005)
		AFT	0	23.6	15.6	0	61.6	0.024	0.950 (0.004)
	200	PSPR	0	0	5.0	100	0	0.005	0.955 (0.001)
		Cox	0	34.8	20.7	0	74.9	3.991	0.953 (0.002)
0.9		Truth							0.968 (0.001)
	100	PSPR	32.4	0.5	3.6	5	5.1	0.375	0.958 (0.005)
		Cox	0	21.4	14.6	0	62.1	3.904	0.963 (0.015)
		AFT	0	13.6	11.1	0	48.3	0.059	0.966 (0.002)
	200	PSPR	9.0	0.3	4.7	48	2.6	0.150	0.964 (0.003)
		Cox	0	22.2	15.0	0	65.1	3.974	0.967 (0.001)

Table 1: Simulation: $\beta = (2.1, 1.8, 1.9, 2.2, 2.7, rep(0, 45))$

ρ	Sample Size	Method	FNR %	FPR %	Model Size	Correct Model %	Wrong Variable %	Bias	AUC
0		Truth							0.868 (0.003)
	100	PSPR	2.2	1.3	5.5	58	8.5	0.063	0.851 (0.018)
		Cox	0	23.6	15.6	0	66.4	3.934	0.850 (0.010)
		AFT	0	26.6	17.0	0	68.5	0.098	0.841 (0.012)
	200	PSPR	0	0.6	5.3	83	4.0	0.017	0.863 (0.005)
		Cox	0	28.6	17.9	0	69.7	3.976	0.861 (0.005)
0.5		Truth							0.909 (0.002)
	100	PSPR	8.4	0.6	4.9	48	5.2	0.118	0.897 (0.014)
		Cox	0	20.6	14.2	0	62.6	3.950	0.903 (0.004)
		AFT	0	18.7	13.4	0	54.6	0.062	0.901 (0.008)
	200	PSPR	0.4	0.1	5	92	1.0	0.025	0.907 (0.004)
		Cox	0	23.5	15.6	0	65.3	3.975	0.906 (0.003)
0.9		Truth							0.935 (0.002)
	100	PSPR	45.2	0.9	3.1	0	10.4	0.560	0.924 (0.006)
		Cox	1.8	13.8	11.1	0	50.1	3.779	0.929 (0.015)
		AFT	1.3	12.0	10.3	7	44.9	0.194	0.930 (0.004)
	200	PSPR	34.6	0.4	3.5	1	4.5	0.379	0.929 (0.003)
		Cox	0	15.7	12.1	0	55.5	3.908	0.933 (0.002)

Table 2: Simulation: $\beta = (1.05, 0.9, 0.95, 1.1, 1.35, rep(0, 45))$

ρ	Sample Size	Method	FNR %	FPR %	Model Size	Correct Model %	Wrong Variable %	Bias	AUC
0		Truth							0.934 (0.002)
	100	PSPR	1.2	0	5.3	84	4.1	0.029	0.924 (0.034)
		Cox	0	3.3	37.5	0	86.1	3.953	0.912 (0.011)
	200	PSPR	0	0	5.0	97	0.5	0.003	0.932 (0.002)
		Cox	0	4.9	53.7	0	90.4	3.985	0.926 (0.003)
0.5		Truth							0.956 (0.001)
	100	PSPR	5.2	0	5.0	68	4.1	0.070	0.944 (0.019)
		Cox	0	3.1	35.9	0	85.5	3.965	0.947 (0.004)
	200	PSPR	0	0	5.0	99	0.2	0.006	0.954 (0.002)
		Cox	0	4.6	51	0	89.9	3.986	0.951 (0.002)
0.9		Truth							0.968 (0.001)
	100	PSPR	43.8	0	3.1	0	8.9	0.503	0.952 (0.009)
		Cox	0	2.5	29.6	0	82.0	3.913	0.963 (0.003)
	200	PSPR	33.2	0	3.7	1	7.9	0.369	0.958 (0.005)
		Cox	0	3.5	39.8	0	86.8	3.965	0.965 (0.002)

Table 3: Simulation: $\beta = (2.1, 1.8, 1.9, 2.2, 2.7, rep(0, 995))$

ρ	Method	FNR %	FPR %	Model Size	Correct Model %	Wrong Variable %	Bias	AUC
Uniform error								
0	Truth							0.883 (0.002)
	PSPR	1.2	2.0	5.8	51	12.0	0.062	0.866 (0.015)
	Cox	0	20.5	14.2	0	62.3	3.934	0.864 (0.010)
	AFT	0	23.9	15.7	0	63.9	0.076	0.872 (0.015)
0.5	Truth							0.923 (0.002)
	PSPR	13.6	0.7	4.7	34	5.9	0.170	0.906 (0.015)
	Cox	0	18.9	13.5	0	58.5	3.923	0.913 (0.008)
	AFT	0	14.6	11.6	0	50.4	0.091	0.919 (0.011)
0.9	Truth							0.946 (0.001)
	PSPR	45.0	0.8	3.1	0	9.9	0.556	0.936 (0.006)
	Cox	2.4	23	15.2	1	55.3	3.649	0.934 (0.023)
	AFT	7.3	10.9	9.5	0	43.7	0.305	0.944 (0.007)
Point error								
0	Truth							0.909 (0.002)
	PSPR	0.4	0.6	5.3	81	4.1	0.022	0.894 (0.016)
	Cox	0	23.8	15.7	0	65.3	3.946	0.879 (0.012)
	AFT	0	21.6	14.7	0	62.6	0.114	0.852 (0.020)
0.5	Truth							0.940 (0.001)
	PSPR	9.2	0.5	4.8	49	3.3	0.111	0.922 (0.016)
	Cox	0	25.2	16.4	0	63.8	3.905	0.919 (0.022)
	AFT	2.0	15.9	12.1	0	51.6	0.196	0.900 (0.017)
0.9	Truth							0.959 (0.001)
	PSPR	43.0	0.5	3.1	0	6.5	0.508	0.943 (0.006)
	Cox	1.6	20.2	14.0	0	53.7	3.719	0.943 (0.020)
	AFT	14.7	9.7	8.6	7	37.9	0.479	0.932 (0.015)
Laplace error								
0	Truth							0.874 (0.003)
	PSPR	3.2	1.9	5.7	50	11.7	0.083	0.852 (0.025)
	Cox	0.2	20	14.0	0	60.7	3.908	0.849 (0.016)
	AFT	0	20.4	14.2	0	60.4	0.161	0.833 (0.025)
0.5	Truth							0.915 (0.002)
	PSPR	15.2	1.0	4.7	22	7.8	0.195	0.895 (0.017)
	Cox	0	17.6	12.9	1	54.6	3.883	0.900 (0.020)
	AFT	2.0	19.7	13.8	0	54.6	0.234	0.884 (0.020)
0.9	Truth							0.939 (0.002)
	PSPR	47.0	0.8	3.0	0	10.2	0.583	0.927 (0.007)
	Cox	5.2	15.5	11.7	1	49.7	3.669	0.929 (0.022)
	AFT	16.7	7.5	7.5	3	38.9	0.419	0.926 (0.008)

Table 4: Simulation 1: mis-specified model

ρ	Sample	β_0					
	Size	2.1	1.8	1.9	2.2	2.7	other
0	100	13	2	10	11	64	0
	200	8	2	1	9	80	0
0.5	100	1	12	38	43	6	0
	200	0	7	37	53	3	0
0.9	100	0	5	59	36	0	0
	200	0	0	71	28	1	0

Table 5: Distribution of anchor variable in percentage: $\beta = (2.1, 1.8, 1.9, 2.2, 2.7, rep(0, 45))$

ρ	Sample	β_0					
	Size	1.05	0.9	0.95	1.1	1.35	other
0	100	19	1	6	17	56	1
	200	8	3	5	12	72	0
0.5	100	1	3	27	61	8	0
	200	0	6	36	52	6	0
0.9	100	0	5	51	43	1	0
	200	0	0	68	32	0	0

Table 6: Distribution of anchor variable in percentage: $\beta = (1.05, 0.9, 0.95, 1.1, 1.35, rep(0, 45))$

ρ	Sample	β_0					
	Size	2.1	1.8	1.9	2.2	2.7	other
0	100	13	6	7	7	64	3
	200	8	5	4	5	78	0
0.5	100	0	10	25	58	7	0
	200	0	9	30	59	2	0
0.9	100	0	4	65	31	0	0
	200	0	0	64	36	0	0

Table 7: Distribution of anchor variable in percentage: $\beta = (2.1, 1.8, 1.9, 2.2, 2.7, rep(0, 995))$

ρ	β_0					
	1.05	0.9	0.95	1.1	1.35	other
Uniform error						
0	14	5	11	12	58	0
0.5	1	14	39	39	7	0
0.9	0	10	52	35	3	0
Point error						
0	12	7	9	14	58	0
0.5	2	14	38	35	11	0
0.9	0	7	60	32	1	0
Laplace error						
0	20	6	8	17	49	0
0.5	2	5	37	50	6	0
0.9	0	6	56	36	2	0

Table 8: Distribution of anchor variable in percentage: mis-specified model

Sample	Number of	Correlation		
Size	Variables	0	0.5	0.9
100	50	207	106	146
	500	489	315	454
	1000	496	375	479
200	50	227	247	294
	500	3127	1838	1752
	1000	3368	2893	2977

Table 9: PSPR: Total time (in seconds) averaged over 4 trials for choosing the scale parameter and computing the solution path at 100 λ values. Timing was carried out on an Intel Xeon 2.33GHz processor.

Sample Size	Number of Variables	Correlation 0	0.5	0.9
100	50	7	24	25
	500	10	7	8
	1000	10	8	9
200	50	1	2	2
	500	31	27	21
	1000	44	32	29

Table 10: Cox (glmnet): Total time (in seconds) averaged over 4 trials for computing the solution path at 100 λ values. Timing was carried out on an Intel Xeon 2.33GHz processor.

Gene	Transcript ID	PSPR Estimate	Cox Estimate
RUNX1	3930361	1	1
DFFA	2396125	-0.80	-1.30
ELOVL6	2781817	-0.91	-0.33
SLCO5A1	3139581	1.05	0.17
GGA2	3685188	0.82	
LOC400713	3840196	0.49	
CHSY3	2827867	1.22	
NOC3L	3301012	-1.40	
MST1	2674603	-1.62	
TEAD1	3321056	-0.49	
LRRC23	3402994	1.43	
LTBP1	2476515		0.95
NONO	3980888		-0.78
PDCD11	3262201		-1.21
RBM9	3959207		2.33
CTNNBL1	3884328		-0.41
ADAM6	3581642		0.45
ZNF418	3872545		0.79
FRZB	2590721		0.74
C8orf33	3121024		-0.07
SUGT1L1	3511005		1.28
RELL1	2765866		0.26
FAM55C	2634059		-1.10
ERCC6	3288709		-0.01
ZNF417	3872529		0.33

Table 11: Final model

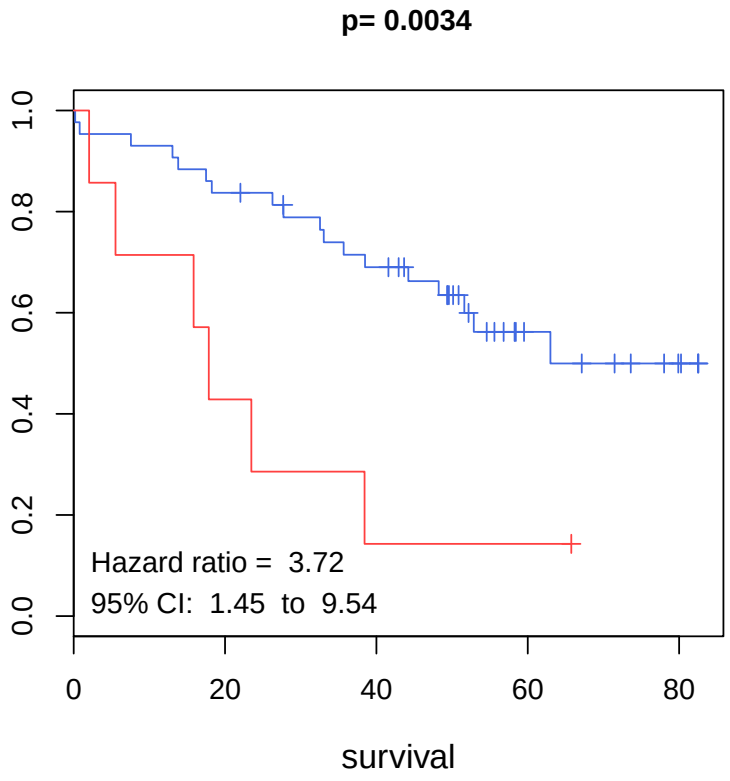

Figure 1: PSPR model: Survival comparison between high/low risk groups on the validation set (the high or low risk is defined based on whether the model-based risk score exceeds the 15th percentile in the training set)

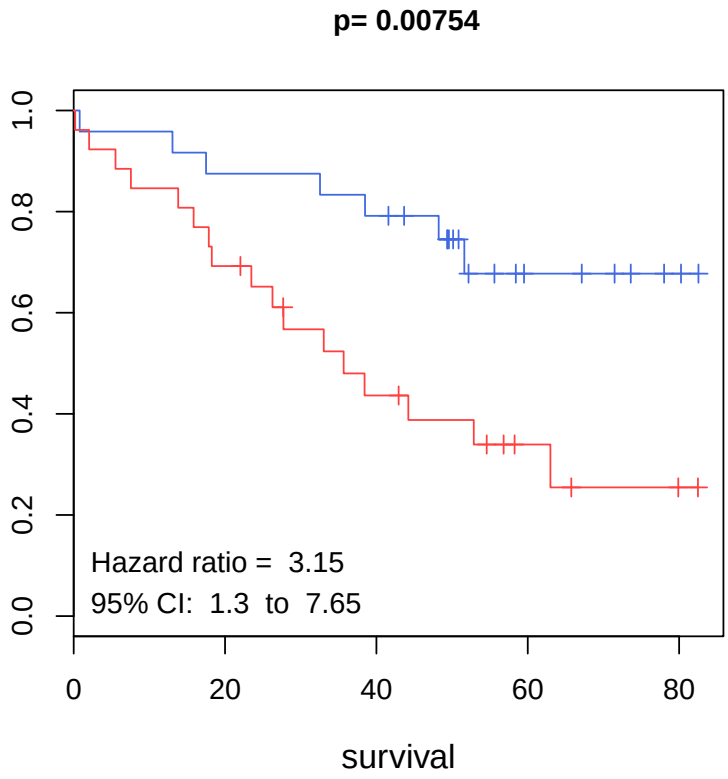

Figure 2: PSPR model: Survival comparison between high/low risk groups on the validation set (the high or low risk is defined based on whether the model-based risk score exceeds the median in the training set)

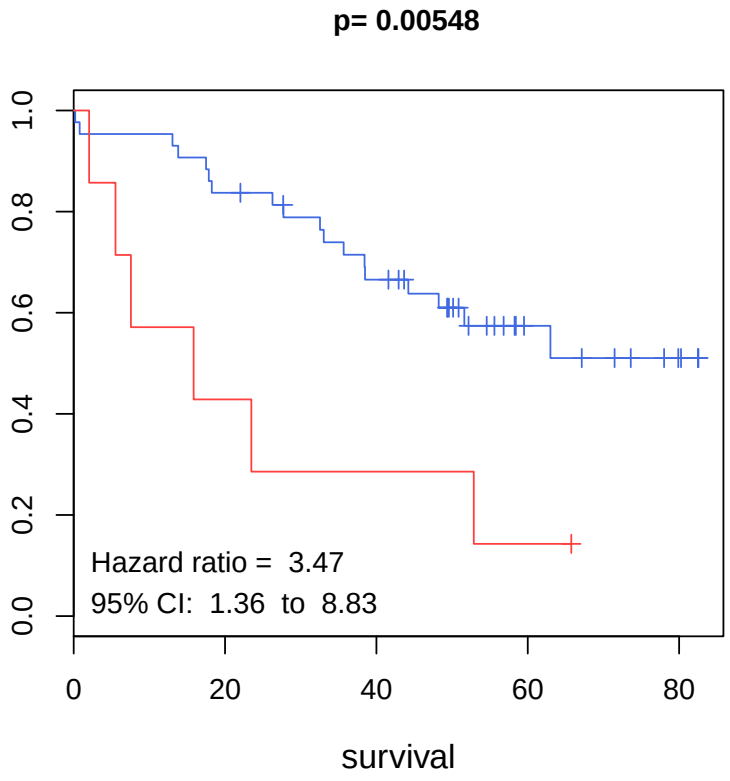

Figure 3: Cox model: Survival comparison between high/low risk groups on the validation set (the high or low risk is defined based on whether the model-based risk score exceeds the 15th percentile in the training set)

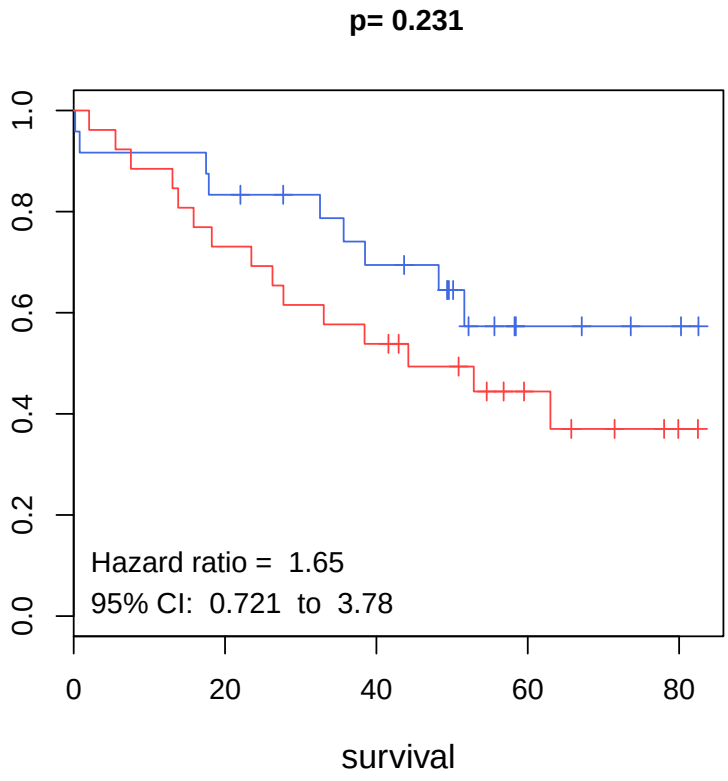

Figure 4: Cox model: Survival comparison between high/low risk groups on the validation set (the high or low risk is defined based on whether the model-based risk score exceeds the median in the training set)

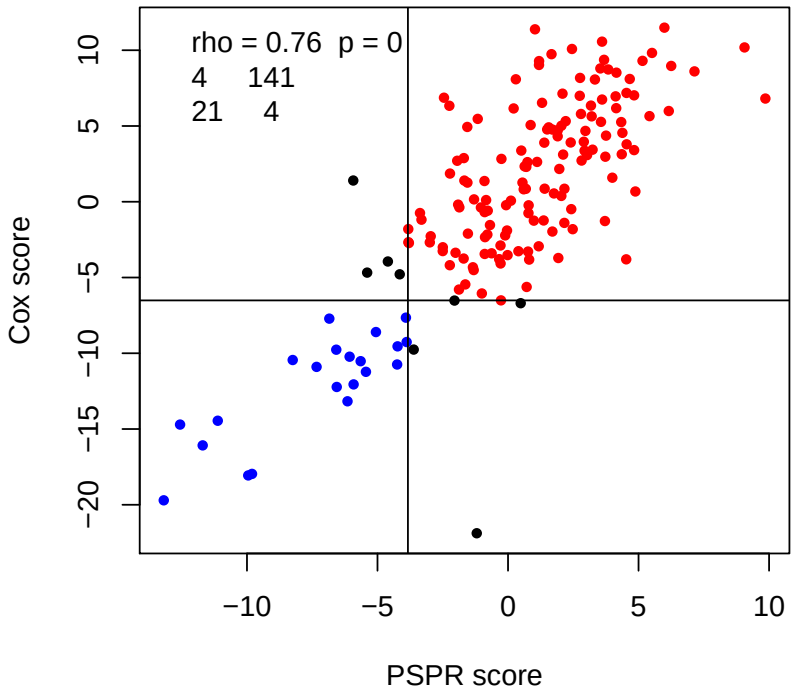

Figure 5: Cox risk score against PSPR risk score. The horizontal and vertical lines are the 15% percentile of the Cox and PSPR risk score in the training patients, respectively.