Liu et al. BMC Bioinformatics 2010, 11:606
http://www.biomedcentral.com/1471-2105/11/606

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Kernel based methods for accelerated failure time model with ultra-high dimensional data

Zhenqiu Liu[1*], Dechang Chen[2], Ming Tan[1], Feng Jiang[3], Ronald B Gartenhaus[1]

Abstract

Background: Most genomic data have ultra-high dimensions with more than 10,000 genes (probes). Regularization methods with L_1 and L_p penalty have been extensively studied in survival analysis with high-dimensional genomic data. However, when the sample size $n \ll m$ (the number of genes), directly identifying a small subset of genes from ultra-high ($m > 10,000$) dimensional data is time-consuming and not computationally efficient. In current microarray analysis, what people really do is select a couple of thousands (or hundreds) of genes using univariate analysis or statistical tests, and then apply the LASSO-type penalty to further reduce the number of disease associated genes. This two-step procedure may introduce bias and inaccuracy and lead us to miss biologically important genes.

Results: The accelerated failure time (AFT) model is a linear regression model and a useful alternative to the Cox model for survival analysis. In this paper, we propose a nonlinear kernel based AFT model and an efficient variable selection method with adaptive kernel ridge regression. Our proposed variable selection method is based on the kernel matrix and dual problem with a much smaller $n \times n$ matrix. It is very efficient when the number of unknown variables (genes) is much larger than the number of samples. Moreover, the primal variables are explicitly updated and the sparsity in the solution is exploited.

Conclusions: Our proposed methods can simultaneously identify survival associated prognostic factors and predict survival outcomes with ultra-high dimensional genomic data. We have demonstrated the performance of our methods with both simulation and real data. The proposed method performs superbly with limited computational studies.

Background

Survival prediction and prognostic factor identification play a very important role in medical research. Survival data normally include the censoring variable that indicates whether some outcome under observation (like death or recurrence of a disease) has occurred within some specific follow-up time. The modeling procedures must take into account such censoring. It is even more difficult to develop a proper statistical learning method for survival prediction.

Several models for survival predictions have been proposed in statistical literature. The most popular one is the Cox proportional hazards model [1-3], in which model parameters are estimated with partial log likelihood maximization. Another one is the accelerate failure time (AFT) model [4-6]. AFT is linear regression model in which the response variable is the logarithm or a known monotone transformation of a failure (death) time. There are mainly two approaches in literature for fitting a AFT model. One is the the Buckley-James estimator which adjusts censored observations using the Kaplan Meier estimator [7,8], and the other is a semiparametric estimation of AFT model with an unspecific error distribution [9-11]. However, the semiparametric Bayesian approach based on complex MCMC procedures is computationally intensive and tends to have inaccurate results, and the Stute's weighted least squares (LS) estimator only implicitly accounts for the censored time. The model has not been widely used in practice due to the difficulties in computing the model parameters [12], and there is no nonlinear AFT model in the literature.

* Correspondence: zliu@umm.edu
[1]University of Maryland Greenebaum Cancer Center, 22 South Greene Street, Baltimore, MD 21201, USA
Full list of author information is available at the end of the article

Kernel based methods such as support vector machines (SVM) have been extensively studied recently in the framework of classification and regression [13] in the area of pattern recognition and statistical learning. The concept of kernel formulated as an inner product in the feature space allows us to build nonlinear extensions of many linear models [14]. It would have been a potential alternative if it were not for the complexity of censoring. Moreover, LASSO type penalty and its generalized versions have been proposed for gene (variable) selection with high dimensional genomic profiles with censored survival outcomes [15-18]. However, since the sample size $n \ll m$ (the number of variables), methods based the primary formulation with a huge m ($m > 40,000$) are not efficient. Consequently, in current microarray analysis, what people really do is select a couple of thousands (or hundreds) of genes using filter-based methods (such as T-test) and then apply the LASSO-type penalty to further reduce the number of disease associated genes. This two-step procedure will lead to missing biologically important genes and introducing bias. The dual solution with kernel proposed in this article attempts to resolve these inadequacies by solving a much smaller $n \times n$ matrix.

In this paper, we propose a nonlinear kernel ridge regression for censored survival outcome prediction under the framework of AFT model. We also develop an efficient dual solution with adaptive kernel ridge regression for ultra-high dimensional genomic data analysis.

Unlike the weighted least square method, our model explicitly accounts for censoring. The proposed models are evaluated with simulation and real data and the prediction error of the test data.

Results and Discussion

Simulation Data

Simulation studies are conducted to evaluate the performance of the proposed methods under different assumptions. The following describes a method to generate input data with censored survival outcomes that emulates the mechanisms presented by the actual data.

1. Sample 12 -dimensional input data $\mathbf{x}$ with 100 training and test samples respectively from a multivariate normal distribution with mean zero and variance-covariance matrix Σ. The pairwise correlation between the ith and the jth input variables in Σ is $r^{|i-j|}$ and different correlations (r = 0.2, 0.4, 0.6, and 0.8) will be chosen to assess the performance of the proposed method.
2. Choose the model parameters $\mathbf{w} = [1, 1, 1, 1, 1, 1, -1, -1, -1, -1, -1, -1]^T$, and generate the event time from $\log T = \mathbf{w}^t \mathbf{x}^k + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$ and σ is determined by the signal to noise ratio (SNR = μ_{surv}/σ). For instance, with the mean log survival time of 3, and $SNR = 3:1$, we have $\sigma = 1$. $SNR = 3:1$ is used in all of our simulations. Finally, k indicates the kth power of input variables, so the log survival time is associated with the input variables nonlinearly.
3. The censoring variables are generated as uniformly distributed and independent of the events. Letting $d_i = (rand + C)T_i$, the censoring status will be $\delta_i = T_i < d_i$. Different Cs give a different portion of censored data. Roughly 40% - 60% censored data are produced in our simulations.

We analyze the simulation data with the proposed DKRR algorithm and build the model with training data, evaluate the performance of the model with the test data. The performance of the DKRR algorithm with different kernels and different correlation structures are shown in Figure 1. As shown in the upper panels of Figure 1, when the the survival data are simulated with $k = 1$ and the true model is linear, the linear model has the best performance with the the average relative root mean squared error (RRMSE) around 0.1. Models with the radial basis function (rbf) kernel have the second best performance with different correlation structures (r = 0.2 -0.8). Models with the third order polynomial have the worst performance with the mean RRMSE around 0.4. On the other hand, when the survival data are generated with a quadratic model with $k = 2$ as shown in the lower panels of Figure 1, Model with second order polynomial kernel and rbf kernel are the two top performers with the average test RRMSE around 0.2, and the linear model performs the worst with the largest average test RRMSE around 0.6. These results indicate that model specification is very important. A misspecified model may lead to inaccurate predictions. Finally, there are no statistical significant differences for input variables with different correlations (r = 02 -0.8).

To evaluate the performance of AKRR method, the survival data are generated from linear model with r = 0.4, and different $\mathbf{w}$s. The generated input data have the dimensions of 100, 1000, 10000, 50000, and 100000, but only 12 variables at the positions of 1, 11, 21, 31, ..., 101, 111 are nonzero with the values of $[w_1, w_{11}, w_{21}, ..., w_{101}, w_{111}]^T = [1, 1, 1, 1, 1, 1, -1, -1, -1, -1, -1, -1]^T$, $[0.2, 0.2, 0.2, 0.2, 0.2, 0.2, -0.2, -0.2, -0.2, -0.2, -0.2, -0.2]^T$, or $[0.1, 0.1, 0.1, 0.1, 0.1, 0.1, -0.1, -0.1, -0.1, -0.1, -0.1, -0.1]^T$ respectively. The rest coefficients are set to 0. The random noise and rest of the variables are generated from the distribution of $N(0, \sigma^2)$, and σ is determined by the mean survival time and the signal to noise ratio (SNR = 3:1). The test RRMSEs with different input dimensions are shown in Figure 2. Figure 2 shows that the test RRMSEs have not changed significantly when

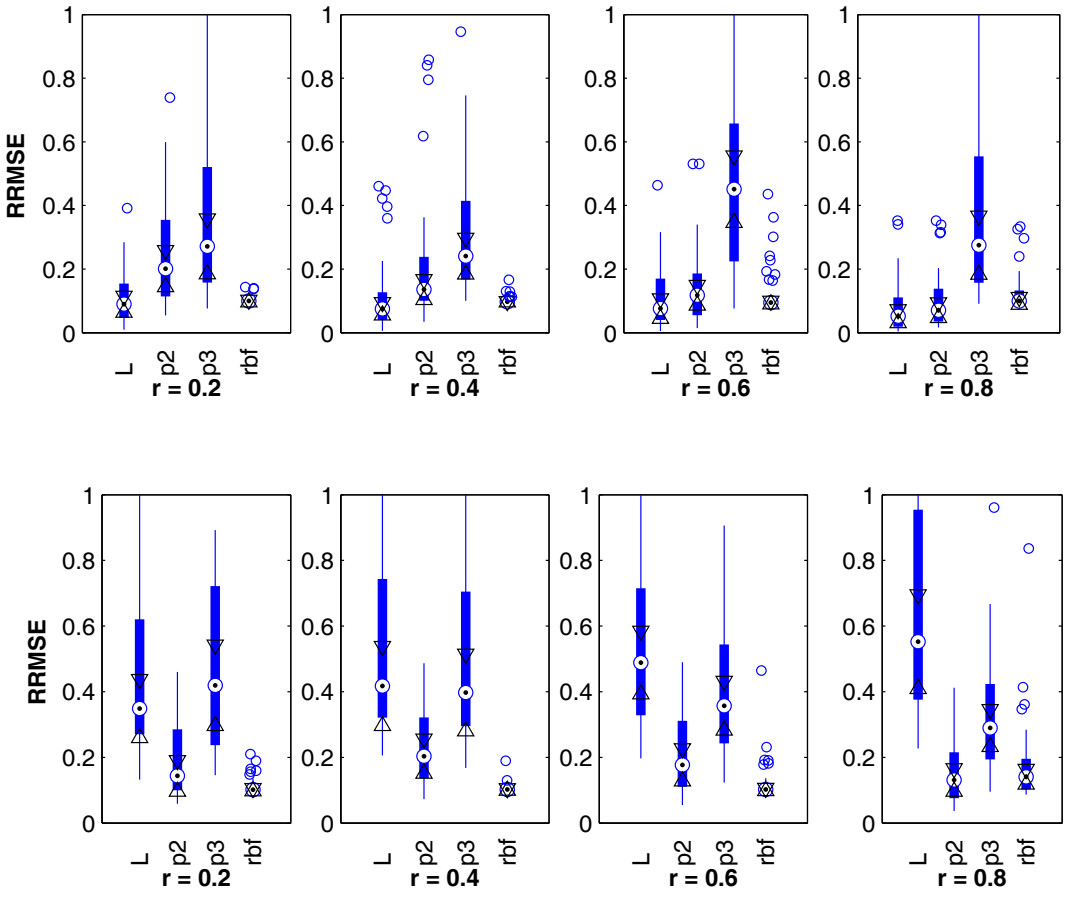

Figure 1 Test RRMSE with Different Correlation Structures. Test Relative Root Mean Squared Error (RRMSE) with Different Models and Different Correlation Structures: L - linear; p2 - second order polynomial kernel; p3 - third order polynomial kernel; and rbf - radial basis function kernel. The upper panels show the performance with the linear model (k = 1) and the lower panels show the performance with quadratic model(k = 2).

the input dimension increases from 100 to 100000, which indicates that AKRR method performs well even with a huge number of variables. The frequencies of first 12 component variables being selected out of 100 random simulations with different **w** are given in Table 1. Table 1 shows that AKRR can correctly identify the survival associated variables with high accuracy. AKRR identifies all 12 variables with over 88% ratios and identifies 10 out of 12 variables with over 96% ratios when $|w_i| = 1$. Moreover, the performances are still very impressive when the associations between survival time and covariates are weak. AKRR identifies 10 out of 12 variables with over 95% and 94% ratios when $|w_i| = 0.2$ and $|w_i| = 0.1$ respectively. Table 2 gives more details about the average number of variables being selected and the ratios of correctly-detected, over-fitting, and under-fitting. The optimal parameters in the parenthesis are decided by 10-fold cross-validation with the training data only. p^* is chosen from the values of 0.6, 0.7, 0.8, 0.9, 1, since our computational experiments show that AKRR seems to converge to the same solution when $p \geq 0.6$ with different initializations for the same data set. λ^* is chosen from 0:0.001:1. The average number of selected variables varies from 11.43-12.61 around the true number 12. AKRR identifies exactly the same 12 variables with the ratios of 75%, 54%, and 52% for $|w_i| = 1$, 0.2, and 0.1 respectively. In all three cases, AKRR chooses the number of variables in the range of 11-13 with over 90% ratio.

For comparison purposes, we also implement the primal version of LASSO for AFT model with Gauss-Seidel method to optimize **w** directly. The computational time

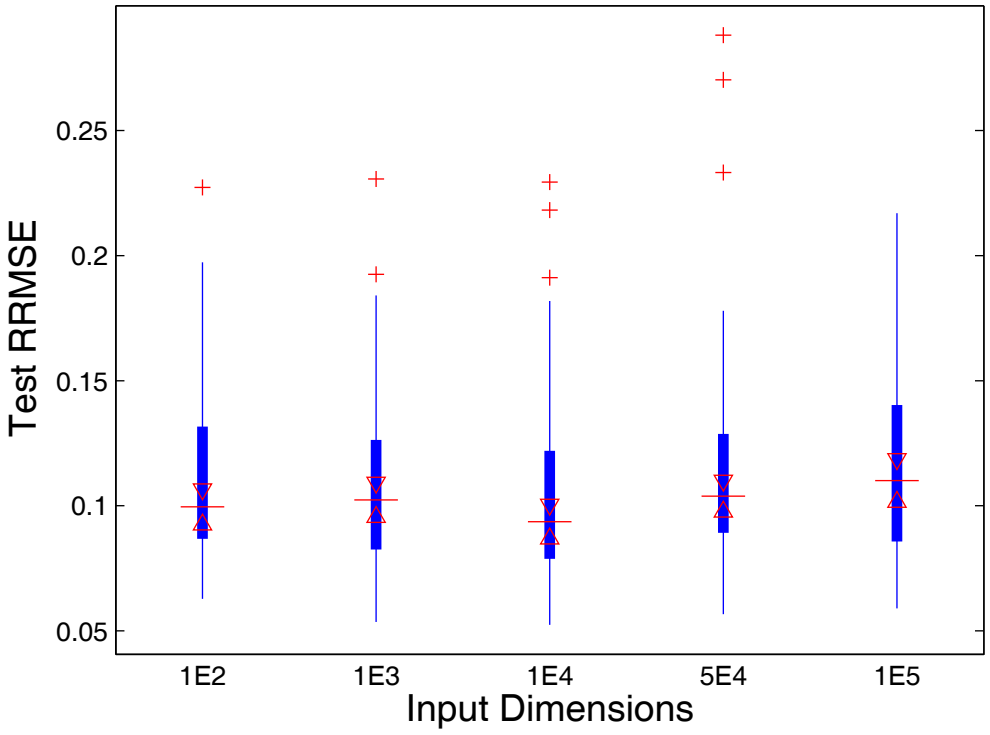

Figure 2 RRMSE with Different Input Dimensions. Test Relative Root Mean Squared Error (RRMSE) with Different Input Dimensions. The input dimensions vary from 100 to 100,000.

for different input dimensions is listed in Table 3. Table 3 shows that AKRR is so computational efficient that it only takes 17.5 seconds for one run to identify variables from 100,000 candidate variables, while LASSO might take days. With 50000 variables, AKRR only needs 7.5 seconds on average to converge, while LASSO fails to converge after 2 hours. When the input dimension is large, AKRR is much more efficient. This is reasonable since the computational time of AKRR is mainly associated with the sample size and dual variables. This method will be fast even with ultra-high dimensional input as long as the sample size is small, which is common in genomic data analysis.

Table 1 Frequencies of Correctly Identified variables with Different Parameters Out of 100 Simulations

Parameters	$\|w_i\| = 1$	$\|w_i\| = 0.2$	$\|w_i\| = 0.1$
w_1	100	100	99
w_{11}	100	98	99
w_{21}	97	99	98
w_{31}	98	99	99
w_{41}	98	98	94
w_{51}	88	77	81
w_{61}	90	86	77
w_{71}	98	95	99
w_{81}	98	98	97
w_{91}	96	96	95
w_{101}	99	99	98
w_{111}	99	100	100

Diffuse Large B-cell Lymphoma Data

We now consider one diffuse large B-cell lymphoma (DLBCL) data [19] evaluating gene expression profiles associated with the patient's survival. In this study, tumor-biopsy specimens and clinical data were obtained

Table 2 Model performance with Simulation Data and Different Parameter Values

$\|w_i\|$s (λ^*, p^*)	Av. # of Vars	Exactly-match	Overfitting	Underfitting
1 (0.01, 0.6)	12.61	75%	21%	4%
0.2 (0.002, 0.6)	11.52	54%	3%	43%
0.1 (0.001, 0.6)	11.43	52%	2%	46%

Table 3 Computational Time (in Seconds): AKRR vs LASSO

Input Dimensions	AKRR	LASSO
100	0.4801	0.6378
1000	0.5844	6.4577
10000	1.7500	978.23
50000	7.5255	>7200
100000	17.4545	- -

retrospectively from 240 patients with untreated diffuse large-B-cell lymphoma who had no previous history of lymphoma, according to a protocol approved by the National Cancer Institute institutional review board. The median follow-up time was 2.8 years overall (7.3 years for survivors), and 57 percent of patients died during this period. The median age of the patients was 63 years, and 56 percent were men. CDNA microarray data with 7,399 probes were collected. We divide the data into two equal parts with 120 training data and 120 test data. We utilize the two-fold cross validation scheme to choose the optimal λ and evaluate our method. We randomly split the data into two roughly equal-sized subsets and build the model with one subset and test it with the other. To avoid the bias arising from a particular partition, the procedure is repeated 100 times, each time we split the data randomly into two folds and do cross validation. The relevance count is utilized to count how many times a gene is selected in the cross validation. Clearly the maximum relevance count for a gene is 200 with the two-fold cross validation and 100 repeating. The optimal λ^* is in the range of 0.26-0.3, and the optimal p^* is set to 0.7 in all the experiments. The test RRMSE is 0.07 on average, which is better than the average test RRMSE (0.101) with LASSO based primal model. This is reasonable, since AKRR has one additional parameter p. Genes associated with survival time are shown in Table 4. We identify 23 probes with over 100 relevant counts. Those 23 probes are corresponding to 21 known genes. All of the selected genes play important roles in apoptotic processes and/or the development and progress of various cancers. 17 out of 21 genes are associated with different lymphoma according to PubMed. For example, BMP6 is the top gene in other category associated with poor outcome and HLA-C gene is from the major histocompatibility class (MHC) II family, both genes were also identified by Rosenwald et al. 2002. Moreover, CD86, CD79a, and CD19 are well known antigens and MHC II signatures associated with favorable survival outcomes. We then perform pathway analysis using DAVID (david.abcc.ncifcrf.gov) and identify 5 lymphoma associated pathways: NOD-like Receptor Signaling Pathway, Pathways in Cancer, Allograft Rejection, Focal Adhesion, and Graft-versus-host Disease. Four out 5 pathways (except for NOD-like Receptor Signaling Pathway) are known to be associated with DLBCL from PubMed.

Follicular Lymphoma (FL) Data

Follicular lymphoma is a common type of Non-Hodgkin Lymphoma (NHL). It is a slow growing lymphoma that arises from B-cells, a type of white blood cell. It is also called an "indolent" or "low-grade" lymphoma for its slow nature, both in terms of its behavior and how it looks under the microscope. A study was conducted to predict the survival probability of patients with gene expression profiles of tumors at diagnosis [20]. Fresh-frozen tumor biopsy specimens and clinical data were obtained from 191 untreated patients who had received a diagnosis of follicular lymphoma between 1974 and 2001. The median age of patients at diagnosis was 51 years (range 23 - 81) and the median follow up time was 6.6 years (range less than 1.0 - 28.2). The median follow up time among patients alive was 8.1 years. Four records with missing survival information were excluded from the analysis. Affymetrix U133A abd U133B microarray genechips were used to measure gene expression levels from RNA samples. A log 2 transformation was applied to the Affymetrix measurement. Detailed experimental protocol can be found from the original paper. There are total of 42928 probes. It is time consuming to directly apply standard LASSO methods to this problem without an initial reduction of dimensions. Our method takes less than 10 seconds for one run. Similar two-fold cross validation scheme with 100 random partitions is utilized to this data. The optimal λ^* is in the range of 0.1 - 0.12 with the optimal $p^* = 0.6$. The test RRMSE is 0.09. The final results are shown in Table 5.

Thirteen probes with over 100 relevance counts are identified. Those 13 probes are corresponding to 11 known genes associated with lymphoma and related diseases. For instance, gene C4A localizes to the major histocompatibility complex (MHC) class III region on chromosome 6. It plays a central role in the activation of the classical pathway of the complement system. C4A anaphylatoxin is a mediator of local inflammatory process. It induces the contraction of smooth muscle, increases vascular permeability, and causes histamine release from mast cells and basophilic leukocytes. C4A is on the pathway of Systemic Lupus Erythematosus (SLE). Patients with SLE can increase the risk of certain cancers, including non-Hodgkin's lymphoma. We find that C4A is negatively associated with survival time according the estimated coefficient of C4A. ALDH2 is another well studied gene which is significantly associated with acetaldehyde-induced micronuclei and alcohol-induced facial flushing. Defects in ALDH2 are a cause of acute alcohol sensitivity and alcohol induced

Table 4 Genes Associated with Survival Time for DLBCL Data

Count	GenBank	Symbal	Description
200	X59618	RRM2	ribonucleotide reductase M2 polypeptide
200	X15187	HSP90B1	tumor rejection antigen (gp96) 1
200	M60315	BMP6	bone morphogenetic protein 6
176	U04343	CD86	CD86 antigen (CD28 antigen ligand 2, B7-2 antigen)
181	X07203	MS4A1	membrane-spanning 4-domains, subfamily A, member 2
198	S75217	CD79A	CD79A antigen (immunoglobulin-associated alpha)
200	M28170	SD19	CD19 antigen
138	U45878	BIRC3	baculoviral IAP repeat-containing 3
146	U10485	LRMP	lymphoid-restricted membrane protein
176	U07620	MAPK10	mitogen-activated protein kinase 10
179	LC_30727		
153	M63438	HLA-C	immunoglobulin kappa constant
164	U46767	CCL13	small inducible cytokine subfamily A (Cys-Cys), member 13
142	X14723	CLU	clusterin
200	M27492	IL1R1	interleukin 1 receptor, type I
183	J05070	MMP9	matrix metalloproteinase 9
200	X61118	LMO2	LIM domain only 2 (rhombotin-like 1)
200	M81750	MNDA	myeloid cell nuclear differentiation antigen
115	X57809	IGL@	heat shock 70 kD protein 1A
162	J03746	MGST1	microsomal glutathione S-transferase 1
200	D38535	ITIH4	inter-alpha (globulin) inhibitor H4
200	M21574	PDGFRA	platelet-derived growth factor receptor, alpha polypeptide
187	ESTs	ESTs	

cancers. There are accumulating evidences that ALDH2-deficient individuals are at much higher risk of esophageal cancer and malignant lymphoma. Our study indicates that the up-regulated ALDH2 is positively associated with patient survival outcomes. Six other genes are also associated with different cancers including follicular lymphoma.

Table 5 Genes Associated with Survival Time for FL Data

count	ProbeID	Symbal	Description
200	231760_at	C20orf51	chromosome 20 open reading frame 51
200	232932_at		
200	235856_at	C4A	complement component 4A (Rodgers blood group)
187	224280_s_a	LOC56181	family with sequence similarity 54, member B
200	201425_at	ALDH2	aldehyde dehydrogenase 2 family (mitochondrial)
180	214694_at	M-RIP	Myosin phosphatase Rho-interacting protein
200	214713_at	YLPM1	YLP motif containing 1
200	218477_at	TMEM14A	transmembrane protein 14A
200	220669_at	HSHIN1	HIV-1 induced protein HIN-1
195	203970_s_a	PEX3	peroxisomal biogenesis factor 3
200	208470_s_a	HPR	haptoglobin-related protein; haptoglobin
175	210920_x_a		
200	215444_s_a	TRIM31	tripartite motif-containing 31

Conclusions

We proposed kernel based methods for nonlinear AFT model and variable selection for ultra-high dimensional data. Our evaluations with simulation and real data illustrate that the proposed methods can effectively reduce the dimension of the covariates with sound prediction accuracy. In many studies, both clinical and genomic data are available. Due to the ultra-high dimension in genomic data, directly applying LASSO based methods to genomic data is usually not feasible. Our proposed method provides an efficient solution for it. Kernel based nonparametric methods have been well studied in statistical learning, but there are not many studies for survival analysis. In this paper, we provide a basis for further explorations in this field.

Methods

To formulate the model, consider a set of n independent observations $\{T_i, \delta_i, \mathbf{x}_i\}_{i=1}^n$, where δ_i is the censoring indicator, T_i is the survival time (event time) if $\delta_i = 1$ or censoring time if $\delta_i = 0$, and $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{im})^t$ is the m-dimensional input vector of the ith sample. Letting $\mathbf{w} = (w_1, w_2, ..., w_m)^t$ be a vector of regression coefficients

and $\varphi(\mathbf{x}_i)$ is the nonlinear transform of $\mathbf{x}_i$ in feature space, the AFT model is defined as

$$M(\mathbf{x}_i) = \mathbf{w}^t\phi(\mathbf{x}_i), \quad i = 1, \ldots, n, \tag{1}$$

where $M(\mathbf{x}_i) > \log T_i$ if $\delta_i = 0$ and $M(\mathbf{x}_i) = \log T_i$ if $\delta_i = 1$. Because there are both equality and inequality constraints in the model, new methods need to be developed.

Kernel Ridge Regression (KRR)

The kernel ridge regression for right censored survival data is as follows:

$$\begin{aligned} &\min \frac{1}{2n}\sum_{i=1}^{n}\xi_i^2 + \frac{\lambda}{2}\mathbf{w}^t\mathbf{w} \\ &\text{s.t. } |\mathbf{w}^t\phi(\mathbf{x}_i) - \log T_i| < \xi_i, \\ &\quad \text{if } \delta_i = 1; \\ &\quad \mathbf{w}^t\phi(\mathbf{x}_i) > \log T_i - \xi_i, \\ &\quad \text{if } \delta_i = 0; \\ &\quad \xi_i \geq 0, \ \forall\ 1 \leq i \leq n. \end{aligned} \tag{2}$$

When ties in the event times are presented, variables associated with each tied time appear in the constraints independently. We can define an index function $I(\delta_i) = 1$ if $\delta_i = 1$, and for censored data with $\delta_i = 0$, $I(\delta_i)$ is defined as $I(\delta_i) = 1$ if $\log T_i \geq \mathbf{w}^t\varphi(\mathbf{x}_i)$ and 0 otherwise. Then equation (2) is equivalent to the following quadratic function:

$$\begin{aligned} &J(\mathbf{w}) = \\ &\frac{1}{2n}\sum_{i=1}^{n} I(\delta_i)\left\{\mathbf{w}^t\phi(\mathbf{x}_i) - \log T_i\right\}^2 \\ &+ \frac{\lambda}{2}\mathbf{w}^t\mathbf{w}, \end{aligned} \tag{3}$$

where $\lambda \geq 0$. If we set the gradient of $J(\mathbf{w})$ with respect to $\mathbf{w}$ to zero, then the solution for $\mathbf{w}$ is a linear combination of the vectors $\varphi(\mathbf{x}_i)$:

$$\begin{aligned} &\mathbf{w} = \\ &-\frac{1}{n\lambda}\sum_{i=1}^{n} I(\delta_i)\left\{\mathbf{w}^t\phi(\mathbf{x}_i) - \log T_i\right\}\phi(\mathbf{x}_i) \\ &= \sum_{i=1}^{n} a_i\phi(\mathbf{x}_i) = \Phi^t\mathbf{a}, \end{aligned} \tag{4}$$

where Φ is the design matrix, whose i^{th} row is given by $\varphi(\mathbf{x}_i)^t$, and $\mathbf{a} = (a_1, a_2, \ldots, a_n)^t$ are the dual variables, defined by

$$a_i = -\frac{I(\delta_i)}{n\lambda}\left\{\mathbf{w}^t\phi(\mathbf{x}_i) - \log T_i\right\} \tag{5}$$

Substituting $\mathbf{w} = \Phi^t\mathbf{a}$ into a_i, we obtain

$$\begin{aligned} a_i &= -\frac{I(\delta_i)}{n\lambda}\left\{\phi^t(\mathbf{x}_i)\Phi^t\mathbf{a} - \log T_i\right\} \\ &= -\frac{I(\delta_i)}{n\lambda}\left\{K(\mathbf{x}_i, .)\mathbf{a} - \log T_i\right\}, \end{aligned} \tag{6}$$

where $K = (K(\mathbf{x}_i, \mathbf{x}_j))_{n \times n} = (\varphi(\mathbf{x}_i)^t\varphi(\mathbf{x}_j))_{n \times n}$ is a kernel matrix which can be defined explicitly and $K(\mathbf{x}_i, .) = \varphi(\mathbf{x}_i)^t\Phi^t$ is the i-th row of the kernel matrix. Popular kernels include:

- Linear kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^t\mathbf{x}_j,$$

- Radial basis function (Gaussian) kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{|\mathbf{x}_i - \mathbf{x}_j|^2}{\sigma^2}\right),$$

- Polynomial kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \left(\mathbf{x}_i^t\mathbf{x}_j + p_2\right)^{p_1},$$

- Sigmoid kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh\left(\beta\mathbf{x}_i^t\mathbf{x}_j\right).$$

Our kernel ridge regression algorithm based on the dual equation (DKRR) (6) is as follows:

The Dual Kernel Ridge Regression (DKRR) Algorithm

Given λ, training data $\{\mathbf{x}_i, \log T_i, \delta_i\}_{i=1}^{n}$, test data $\{\mathbf{x}_k, \log T_k, \delta_k\}_{k=1}^{n_k}$, and a small ε.

Calculate the kernel matrices $K = \left[K(\mathbf{x}_i, \mathbf{x}_j)\right]_{n \times n}$ and $K_{te} = \left[K(\mathbf{x}_k, \mathbf{x}_i)\right]_{n_k \times n}$.

Center the kernels and the survival times: $K = \left(I_n - \frac{1}{n}1_n1_n^t\right)K\left(I_n - \frac{1}{n}1_n1_n^t\right)$ and

$K_{te} = \left(K_{te} - \frac{1}{n}1_{n_k}1_n^t K\right)\left(I_n - \frac{1}{n}1_n 1_n^t\right)$, where 1_n: a vector with n 1's. and I_n: an identity matrix, and $\log T_k = \log T_k - \frac{\sum_{k=1}^{n_k}\log T_k}{n_k}$. Let $\mathbf{a}^0$ = [0, ..., 0, 0]^t, and j = 0

WHILE 1,

- FOR i = 1 to n,

$$I(\delta_i) = \begin{cases} 1 & \text{if } \delta_i > 0, \\ 1 & \text{if } \delta_i = 0, \\ & \& K(\mathbf{x}_i, .)\mathbf{a}^j \le \log T_i, \\ 0 & \text{otherwise.} \end{cases}$$

$$a_i^{j+1} = -\frac{I(\delta_i)}{n\lambda}\left\{K(\mathbf{x}_i, .)\mathbf{a}^j - \log T_i\right\}$$

$$\mathbf{a}^{j+1} = \left[a_1^{j+1}, \ldots, a_i^{j+1}, a_{i+1}^j \ldots, a_n^j\right]^t$$

END

- $j = j + 1$
- IF $|\mathbf{a}^{j+1} - \mathbf{a}^j| < \varepsilon$, BREAK.
- $\mathbf{a}^j = \mathbf{a}^{j+1}$

END

This dual kernel ridge regression (DKRR)algorithm designed for a quadratic error function with linear constraints is a convex function with convex constraints. Theoretically this algorithm will always converge and global optimal solution is guaranteed irrelevant to initial value $\mathbf{a}^0$. In our computational experiments, the differences of the estimated parameters with different initial values are very small (less than 0.01 with the infinity norm).

Adaptive Kernel Ridge Regression

When the number of variables is greater than the sample size n, regularization is needed to obtain a stable estimator $\mathbf{w}$. We propose a $L_p = \sum_{i=1}^n |w_i|^p \ (p < 1)$ penalty for variable selection and estimation simultaneously. Unlike LASSO, L_p penalty and its different approximation schemes (i.e., adaptive LASSO) possess the oracle property [21,22]. Here the oracle property of a method means that it can correctly identify the nonzero coefficients with probability converging to one and that the estimators of nonzero coefficients are asymptotically normal with the same means and covariances as what they would have the zero coefficients be known in advance. We therefore propose the following penalized AFT model:

$$\begin{aligned} J(\mathbf{w}) &= \\ &\frac{1}{2n}\sum_{i=1}^n I(\delta_i)\left\{\mathbf{w}^t\phi(\mathbf{x}_i) - \log T_i\right\}^2 \\ &+ \frac{\lambda}{2}\sum_{i=1}^m |w_i|^p \\ &= \frac{1}{2n}\sum_{i=1}^n I(\delta_i)\left\{\mathbf{w}^t\phi(\mathbf{x}_i) - \log T_i\right\}^2 \\ &+ \frac{\lambda}{2}\sum_{i=1}^m \frac{|w_i|^2}{|w_i|^{2-p}}, \end{aligned} \tag{7}$$

where $\lambda \ge 0$. With $n \ll m$, linear kernel is more appropriate, since model with linear kernel has less over-fitting. We will take $\varphi(\mathbf{x}_i) = \mathbf{x}_i$, introduce an auxiliary (latent) variable vector $\mathbf{u} = [u_1, u_2, \ldots, u_m]^t$, and develop an adaptive procedure based on equation (7). Equation (7) can be rewritten as:

$$\begin{aligned} J(\mathbf{w}, \mathbf{u}) &= \\ &\frac{1}{2n}\sum_{i=1}^n I(\delta_i)\left\{\mathbf{w}^t\mathbf{x}_i - \log T_i\right\}^2 \\ &+ \frac{\lambda}{2}\sum_{i=1}^m \frac{|w_i|^2}{|u_i|^{2-p}}, \end{aligned} \tag{8}$$

$$\text{and, } \mathbf{u} = \mathbf{w}. \tag{9}$$

With equation (8) and (9), we will find the first order derivative for $\mathbf{w}$ with a fixed $\mathbf{u}$ and then update $\mathbf{u} = \mathbf{w}$. After taking the first order derivative, we have the following equation:

$$\begin{aligned} \mathbf{w} &= \\ &-\frac{1}{n\lambda}\sum_{i=1}^n I(\delta_i)\left\{\mathbf{w}^t\mathbf{x}_i - \log T_i\right\}\left(\mathbf{x}_i \odot |\mathbf{u}|^{2-p}\right) \\ &= \sum_{i=1}^n a_i\left(\mathbf{x}_i \odot |\mathbf{u}|^{2-p}\right) = X_{\mathbf{u}}^t\mathbf{a}, \end{aligned} \tag{10}$$

where $\odot$ represents the componentwise product of two vectors and

$$X_{\mathbf{u}} = \begin{pmatrix} \mathbf{x}_1^t \ \odot \ \left(|\mathbf{u}|^{2-p}\right)^t \\ \vdots \\ \mathbf{x}_n^t \ \odot \ \left(|\mathbf{u}|^{2-p}\right)^t \end{pmatrix},$$

and

$$a_i = -\frac{I(\delta_i)}{n\lambda}\left\{\mathbf{w}^t\mathbf{x}_i - \log T_i\right\}. \quad (11)$$

We substitute $\mathbf{w} = X_{\mathbf{u}}^t\mathbf{a}$ and define a new kernel function $K_{\mathbf{u}} = XX_{\mathbf{u}}^t$. Then we have $K_{\mathbf{u}}(\mathbf{x}_i, .) = \mathbf{x}_i^t X_{\mathbf{u}}^t$, which is the ith row of $K_{\mathbf{u}}$. So,

$$\begin{aligned} a_i &= -\frac{I(\delta_i)}{n\lambda}\left\{\mathbf{x}_i^t X_{\mathbf{u}}^t\mathbf{a} - \log T_i\right\} \\ &= -\frac{I(\delta_i)}{n\lambda}\left\{K_{\mathbf{u}}(\mathbf{x}_i, .)\mathbf{a} - \log T_i\right\}. \end{aligned} \quad (12)$$

The adaptive kernel ridge regression algorithm based on dual variables **a** with equation (10) and (12) is as follows:

Adaptive Kernel Ridge Regression (AKRR) Algorithm

Given a λ, $p \in (0,1]$, training data $\{\mathbf{x}_i, \log T_i, \delta_i\}_{i=1}^n$, and a small ε and η.

Initializing $\mathbf{w} = \mathbf{u} = \text{rand}(m, 1)$, and $\mathbf{a} = [0, ..., 0]^t$

Setting $\mathbf{u}(u_i == 0) = 10e - 5$ and $j = 1$.

While $|\mathbf{w} - \mathbf{u}| > \varepsilon$

- $\mathbf{u} = \mathbf{w}$,
- $K_{\mathbf{u}} = XX_{\mathbf{u}}^t$
- FOR $i = 1$ to n,

$$I(\delta_i) = \begin{cases} 1 & \text{if } \delta_i > 0 \\ 1 & \text{if } \delta_i = 0, \\ & \& K(\mathbf{x}_i, .)\mathbf{a}^j \le \log T_i, \\ 0 & \text{otherwise.} \end{cases}$$

$$a_i^{j+1} = -\frac{I(\delta_i)}{n\lambda}\left\{K_{\mathbf{u}}(\mathbf{x}_i, .)\mathbf{a}^j - \log T_i\right\}$$

$$\mathbf{a}^{j+1} = [a_1^{j+1}, \ldots, a_i^{j+1}, a_{i+1}^j \ldots, a_n^j]^t$$

END

- $j = j + 1$
- $\mathbf{w} = X_{\mathbf{u}}^t\mathbf{a}$.

END

$\mathbf{w}(\mathbf{w} < \eta) = 0$

Unlike other LASSO based methods which seek to find optimal **w** directly, AKRR algorithm updates the m-dimensional **w** through updating a much smaller n-dimensional dual variable **a**. This method is computationally highly efficient when $n \ll m$, which is common in genomic data. Although the proposed method is based on the dual problem, the primal variable **w** is explicitly updated in the computation. Theoretically AKRR algorithm will always converge to global optimal solution when $p = 1$ irrelevant to initial values of **w**, **u**, and **a**, as the error function is convex under L_1 penalty, but only local optimal solution is guaranteed when $p < 1$. However, in our computational experiments with simulation and real data, even though we may have different optimal solutions with different initializations only when $p \le 0.5$, most selected features are still the same in different runs. AKRR always reach the same optimal solution in all of our experiments when $p \ge 0.6$. One possible explanation is that the error function may still be near convex or convex almost everywhere when p is large. Therefore it may be possible that we enjoy both the oracle property with less bias and the global optimal solution with larger p ($0.6 \le p < 1$). Theoretical study for the near convex error function, however, is out of the scope of this paper. To prevent the results stick to a local optimal solution when $p \le 0.5$, we run AKRR 30 times and the best solution is chosen from the run with smallest test error. Even though AKRR does choose different variables with different ps, a small subset (≥ 5) of most important genes are always selected in our experiments. The model performance can be evaluated with cross-validation and the relative root mean squared error $\left(\text{RRMSE} = \sqrt{\frac{\sum_i((y_i - \gamma_i)/\gamma_i)^2}{n}}\right)$ of the test data. There are two parameters p and λ for the adaptive kernel ridge regression (AKRR) algorithm. One efficient way is to set $p = 0.1, 0.2, ..., 1$ alternatively, and then search for the best λ through cross-validation. The range of can be determined by the path of the optimal solution. $\lambda_{\min} = 0$ and $\lambda_{\max}$ is set to be the smallest value with all zero estimated parameters by multiple trials. We search the optimal λ from $\lambda \in (0, 1]$ in this paper. Usually we have a larger $\lambda_{\max}$ for $p = 1$, and smaller $\lambda_{\max}$ when p is smaller.

Acknowledgements

We thank the Associate Editor and the two anonymous referees for their constructive comments, which improve this manuscript significantly. This work was partially supported by the 1R03CA133899 grant from the National Cancer Institute and by the National Science Foundation CCF-0729080 grant.

Author details

[1]University of Maryland Greenebaum Cancer Center, 22 South Greene Street, Baltimore, MD 21201, USA. [2]Department of Preventive Medicine and Biometrics, Uniformed Services University of the Health Sciences, Bethesda, MD 20814, USA. [3]Department of Pathology, The University of Maryland School of Medicine, Baltimore, MD 21201, USA.

Authors' contributions

ZL conceptualized and designed method, developed the software, and wrote the manuscript. FJ and RG analyzed and interpreted the data on its biological contents. DC and MT helped in method design and manuscript

writing and revised the manuscript critically. All authors read and approved the final manuscript.

Received: 17 August 2010 Accepted: 21 December 2010
Published: 21 December 2010

doi:10.1186/1471-2105-11-606
Cite this article as: Liu et al.: Kernel based methods for accelerated failure time model with ultra-high dimensional data. *BMC Bioinformatics* 2010 **11**:606.

References

1. Cox D: **Regression models and life-tables (with discussion).** *Journal of Royal Statistical Society, Series B* 1972, **34**:187-220.
2. Gui J, Li H: **Variable Selection via Non-concave Penalized Likelihood and its Oracle Properties.** *Journal of the American Statistical Association, Theory and Methods* 2001, **96**:456.
3. Van Houwelingen H, Bruinsma T, Hart A, Van't Veer L, Wessels L: **Cross-validated Cox regression on microarray gene expression data.** *Stat Med* 2006, **25**:3201-3216.
4. Kalbfleisch J, Prentice R: *The Statistical Analysis of Failure Time Data* New York: John Wiley; 1980.
5. Wei L: **The accelerated failure time model. a useful alternative to the Cox regression model in survival analysis.** *Statistics in Medicine* 1992, **11**:1871-1879.
6. Ying Z: **A large sample study of rank estimation for censored regression data.** *Annals of Statistics* 1993, **21**:76-99.
7. Stute W, Wang J: **The strong law under random censorship.** *Annals of Statistics* 1993, **14**:1351-1365.
8. Stute W: **Distributional convergence under random censorship when covariables are present.** *Scandinavia Journal of Statistics* 1996, **23**:461-471.
9. Christensen R, Johnson W: **Modelling accelerated failure time with a Dirichlet process.** *Biometrika* 1988, **75**:693-704.
10. Kuo L, Mallick B: **Bayesian semiparametric inference for the accelerated failure time model.** *Canadian J Stat* 1997, **25**:457-472.
11. Bedrick E, Christensen R, Johnson W: **Bayesian accelerated failure time analysis with application to veterinary epidemiology.** *Stat Med* 2000, **19**:221-237.
12. Jin Z, Lin D, Wei L, Ying Z: **Rank-based inference for the accelerated failure time model.** *Biometrika* 2003, **90**:341-353.
13. Vapnik V: *Statistical Learning Theory* New York: Wiley and Sons; 1998.
14. Shawe-Taylor J, Cristianini N: *Kernel Methods for Pattern Analysis* London: Cambridge University Press; 2004.
15. Ma S, Huang J: **Additive risk survival model with microarray data.** *BMC Bioinformatics* 2007, **8**:192.
16. Sha N, Tadesse M, Vannucci M: **Bayesian variable selection for the analysis of microarray data with censored outcomes.** *Bioinformatics* 2006, **22(18)**:2262-2268.
17. Liu Z, Gartenhaus R, Chen X, Howell C, Tan M: **Survival Prediction and Gene Identification with Penalized Global AUC Maximization.** *Journal of Computational Biology* 2009, **16(12)**:1661-1670.
18. Liu Z, Jiang F: **Gene identification and survival prediction with Lp Cox regression and novel similarity measure.** *Int J Data Min Bioinform* 2009, **3(4)**:398-408.
19. Rosenwald A, Wright G, Chan W, Connors J, Campo E, Fisher R, Gascoyne R, Muller-Hermelink H, Smeland E, Giltnane J, Hurt E, Zhao H, Averett L, Yang L, Wilson W, Jaffe E, Simon R, Klausner R, Powell J, Duffey P, Longo D, Greiner T, Weisenburger DD, DBLJVJAJMELGAGTMTLM, Sanger WG, Ott G, Kvaloy S, Delabie J, Holte H, Krajci P, Stokke T, Staudt L: **The use of molecular profiling to predict survival after themotheropy for diffuse large-B-cell lymphoma.** *The New England Journal of Medicine* 2002, **346**:1937-1947.
20. Dave S, Wright G, Tan B, Rosenwald A, Gascoyne R, Chan W, Fisher R, Braziel R, Rimsza L, Grogan T, Miller T, LeBlanc M, Greiner T, Weisenburger D, Lynch J, Vose J, Armitage J, Smeland E, Kvaloy S, Holte H, Delabie J, Connors J, Lansdorp P, Ouyang Q, Lister T, Davies A, Norton A, Muller-Hermelink H, Ott G, Campo E, Montserrat E, Wilson W, Jaffe E, Simon R, Yang L, Powell J, Zhao H, Goldschmidt N, Chiorazzi M, Staudt L: **Prediction of survival in follicular lymphoma based on molecular features of tumor-in filtrating immune cells.** *N Engl J Med* 2004, **351(21)**:2159-2169.
21. Knight K, Fu W: **Asymptotics for Lasso-type estimators.** *Annals of Statistics* 2000, **28**:1356-1378.
22. Fan J, Peng H: **On Nonconcave Penalized Likelihood With Diverging Number of Parameters.** *The Annals of Statistics* 2004, **32**:928-961.