Threshold Gradient Descent Method for Censored Data Regression with Applications in Pharmacogenomics

J. Gui and H. Li

THRESHOLD GRADIENT DESCENT METHOD FOR CENSORED DATA REGRESSION WITH APPLICATIONS IN PHARMACOGENOMICS*

J. GUI AND H. LI

Department of Statistics and Rowe Program in Genetics
University of California, Davis, CA 95616, USA
E-mail: jgui@ucdavis.edu, hli@ucdavis.edu

An important area of research in pharmacogenomics is to relate high-dimensional genetic or genomic data to various clinical phenotypes of patients. Due to large variability in time to certain clinical event among patients, studying possibly censored survival phenotypes can be more informative than treating the phenotypes as categorical variables. In this paper, we develop a threshold gradient descent (TGD) method for the Cox model to select genes that are relevant to patients' survival and to build a predictive model for the risk of a future patient. The computational difficulty associated with the estimation in the high-dimensional and low-sample size settings can be efficiently solved by the gradient descent iterations. Results from application to real data set on predicting survival after chemotherapy for patients with diffuse large B-cell lymphoma demonstrate that the proposed method can be used for identifying important genes that are related to time to death due to cancer and for building a parsimonious model for predicting the survival of future patients. The TGD based Cox regression gives better predictive performance than the L_2 penalized regression and can select more relevant genes than the L_1 penalized regression.

1. Introduction

With the sequencing of the human genome near completion and with the development of high-throughput technologies, we are now able to obtain information about an individual's entire genome or the entire genomic profiles of a tumor. Very high-dimensional genetic and genomic data are being generated in pharmaceutical industries and in biomedical and clinical research. Examples of high-throughput data include whole genome wide SNP data, microarray-based gene expression data and proteomic data. Traditional environmental risk factors such as diet, age and lifestyle can influ-

ence a person's response to treatments, it is believed that understanding an individual's genetic makeup or individual tumor's genomic profiles would provide the key for explaining such variation and for creating personalized drugs with greater efficacy and safety. For example, with the DNA microarray technology, one can simultaneously measure expression profiles for thousands of genes in cancer tissues, which offers the possibility of a powerful, genome-wide approach to the genetic basis of different types of tumors. Recent studies [1,2] have demonstrated great succuss in predicting cancer class using the gene expression data. Different classes of cancer may correspond to different clinical outcomes of a given treatment. In addition, studies also demonstrated that additional predictive power can be obtained by incorporating genomic information in addition to the traditional predictive factors such as tumor grades, sizes and stages [2].

In pharmacogenomics, an important area of research is to relate the high-dimensional genetic, genomic or proteomic data to various phenotypes such as continuous drug response levels, response to treatment which can be categorical or censored clinical outcomes such as time to cancer recurrence or death after treatment. Due to large variability in time to cancer recurrence among cancer patients, studying possibly censored survival phenotypes can be more informative than treating the phenotypes as binary or categorical variables. Since the follow up time is limited, some patients' exact survival time can't be measured. For those patients, we only have their right censored survival time. The emphasis of this paper is to develop methods for predicting patient's time to clinical event using high-dimensional genetic or genomic data.

The most popular method in regression analysis for censored survival data is the Cox regression model [3]. However, due to the very high dimensional space of the predictors, e.g., the genes with expression levels measured by microarray experiments, the standard maximum Cox partial likelihood method cannot be applied directly to obtain parameter estimates. There are mainly two solutions to this problem. One approach is based on dimension deduction such as singular value decomposition (SVD) or the partial Cox egression (PCR) [4]. The other approach is to use the penalized partial likelihood. This includes L_2 and L_1 penalization. Li and Luan [5] was the first to investigate the L_2 penalized estimation of the Cox model in the high-dimensional low-sample size settings and applied their method to relate the gene expression profile to survival data. As pointed out by Li and Luan [5], one limitation of L_2 penalization is that it uses all the genes in the prediction and does not provide a way of selecting relevant genes.

An alternative is to use the L_1 penalized estimation, which was proposed by Tibshirani [6] and was called the least absolute shrinkage and selection operator (Lasso). Using newly developed least angle regression (LARS) by Efron *et al.* [7], Gui and Li [8] proposed an efficient way to estimate L_1 penalized Cox regression model, which was called the LARS-Lasso procedure. One limitation of the LARS-Lasso procedure is that the number of genes selected cannot be greater than the sample size. In addition, if there is a group of variables among which the pair-wise correlations are very high, the LARS-Lasso procedure tends to select only one variable from the group. This could be a potential limitation when the goal is to select all important genetic or genomic features which are related to the risk of a clinical event.

Friedman and Popescu [9] have recently proposed a stepwise optimization method called threshold gradient descent (TGD) and have demonstrated it application in linear regression and classification problems. They showed that with different threshold value, TGD can approximate the estimates of partial least square, ridge regression, Lasso and LARS. The TGD based methods provide a data-driven approach for selecting the penalty function. Friedman and Popescu [9] further demonstrated that with small threshold, the TGD method approximates the PLS or ridge estimates and has better predictive power for data simulated with small variability in true coefficients. On the other hand, the TGD with large threshold value can approximate the Lasso or LARS estimates, which provide better predictive performance when there is high variability of the coefficients. In this paper, we extend the TGD method to censored survival data and to the Cox regression model. We demonstrate the method by analyzing a real data set of diffuse large B-cell lymphoma (DLBCL) survival times and gene expression levels [2]. Finally, we give a brief discussion of the methods and conclusions.

2. Statistical Models and Methods

2.1. *Cox proportional hazards model*

Suppose that we have a sample size of n from which to estimate the relationship between the survival time and the genetic/genomic profiles such as the gene expression levels $X_1, \cdots, X_p$ of p genes. Due to censoring, for $i = 1, \cdots, n$, the ith datum in the sample is denoted by $(t_i, \delta_i, x_{i1}, x_{i2}, \cdots, x_{ip})$, where δ_i is the censoring indicator and t_i is the survival time if $\delta_i = 1$ or censoring time if $\delta_i = 0$, and $x_i = \{x_{i1}, x_{i2}, \cdots, x_{ip}\}'$ is the vector of the genetic/genomic profiles of p genes for the ith sample. Our aim is to build the following Cox regression model for the hazard of

cancer recurrence or death at time t

$$\lambda(t) = \lambda_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)$$
$$= \lambda_0(t) \exp(\beta' X), \tag{1}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function, $\beta = \{\beta_1, \cdots, \beta_p\}$ is the vector of the regression coefficients, and $X = \{X_1, \cdots, X_p\}$ is the vector of genetic/genomic profiles with the corresponding sample values of $x_i = \{x_{i1}, \cdots, x_{ip}\}$ for the ith sample. We define $f(X) = \beta' X$ to be the linear risk score function.

Based on the available sample data, the Cox's partial likelihood [3] can be written as

$$PL(\beta) = \prod_{k \in D} \frac{\exp(\beta' x_k)}{\sum_{j \in R_k} \exp(\beta' x_j)},$$

where D is the set of indices of the events (e.g., deaths) and R_k denotes the set of indices of the individuals at risk at time $t_k - 0$. Our goal is to find the coefficient vector β which minimizes the negative log partial likelihood function. However, in the settings when $p >> n$, there is no unique solution to this optimization problem. In addition, one expects the high variability of the estimates based on different random samples drawn from the population distribution. A common remedy is to regularize this optimization problem by addition a penalty $\lambda P(\beta)$ to the negative partial likelihood, i.e.,

$$\hat{\beta}(\lambda) = \operatorname{argmin}_\beta \{-\log PL(\beta) + \lambda P(\beta)\}. \tag{2}$$

Here the negative log partial likelihood is treated as a loss function $l(\beta)$ which we want to minimize. The most popular penalties include the L_2 penalty where $P(\beta) = \sum \beta_j^2$ and the L_1 penalty where $P(\beta) = \sum |\beta_j|$.

2.2. *The threshold gradient descent algorithm*

As observed in Friedman and Popescu [9], for a given penalty function $P(\beta)$, the procedure represented by (2) produces a family of estimates, $\hat{\beta}(\lambda)$, each is indexed by a particular value of the tuning parameter λ. This family lies on a one-dimensional path of finite length in the p dimensional space of all joint parameter values β. Model selection procedure such as cross-validation (see next section) can be use for selecting a point (i.e., a λ parameter) on that path. Different penalty $P(\beta)$ therefore corresponds to different path. To account for different possible true values of β, Friedman and Popescu [9]

further suggested a gradient directed path finding algorithm for estimating β. Specifically, let $l(\beta) = -\log PL(\beta)$, $\eta = X\beta$, and define

$$\mu_i = -\partial l/\partial \eta_i = \delta_i - \exp(\eta_i) \sum_{k \in C_i} \frac{d_k}{\sum_{j \in R_k} \exp(\eta_j)},$$

where $C_i = \{k : i \in R_k\}$ denotes the risk sets containing individual i and d_k is the number of events at time t_k. Then $\mu = (\mu_1, \cdots, \mu_n)$ is the negative gradient of the loss function with respective to $\{\eta_1, \cdots, \eta_n\}$. The negative gradient with respective to β is therefore $g = -\partial l/\partial \beta = X\mu$. Starting from $\hat{\beta} = 0$, the gradient directed paths can be updated as

$$\hat{\beta}(\nu + \Delta\nu) = \hat{\beta}(\nu) + \Delta\nu h(\nu),$$

where $\Delta\nu > 0$ is an infinitesimal increment and $h(\nu)$ is the direction in the parameter space tangent to the path evaluated at $\hat{\beta}(\nu)$. This tangent vector at each step represents a descent direction. In order to direct the path towards parameter points with diverse values, Friedman and Popescu [9] suggested to define $h(\nu)$ as

$$h(\nu) = \{h_j(\nu)\}_1^p = \{f_j(\nu).g_j(\nu)\}_1^p,$$

where

$$f_j(\nu) = I[|g_j(\nu)| \geq \tau \cdot max_{1 \leq k \leq p}|g_k(\nu)|],$$

where $I[.]$ is an indicator function, and $0 \leq \tau \leq 1$ is a threshold parameter that regulates the diversity of the values of $f_j(\nu)$; larger values of τ lead to more diversity [9]. $g(\nu)$ is the negative gradient evaluated at $\beta = \hat{\beta}(\nu)$.

For any threshold value $0 \leq \tau \leq 1$, the threshold gradient decent path finding algorithm for the Cox model involves the following five steps,

(1) Set $\beta(0) = 0$, $\nu = 0$.
(2) Calculate η, μ, $g(\nu) = -\partial l/\partial \beta$ for the current β.
(3) Calculate $f_j(\nu) = I[|g_j(\nu)| \geq \tau \cdot max_{1 \leq k \leq p}|g_k(\nu)|]$
(4) Update $\beta(\nu + \triangle\nu) = \beta(\nu) + \triangle\nu \cdot g(\nu) \cdot f(\nu)$, $\nu = \nu + \triangle\nu$.
(5) Repeat 2-4. Cross validation (see next section) is then employed to determine a point on the path (ν) and to terminate the iterations.

2.3. *Model selection through the cross validated partial likelihood*

For a given gradient threshold τ, to determine the value of the tuning parameter ν in the final model, one can choose ν which minimizes the

cross-validated partial likelihood (CVPL), which is defined as

$$CVPL(\nu) = -\frac{1}{n} \sum_{i=1}^{n} [l(\hat{f}^{(-i)}(\nu) - l^{(-i)}(\hat{f}^{(-i)}(\nu)],$$

where $\hat{f}^{(-i)}(\nu)$ is the estimate of the score function based on the threshold gradient descent with tuning parameter ν from the data without the ith subject. The terms $l(f)$ and $l^{(-i)}(f)$ are the log partial likelihoods with all the subjects and without the ith subject, respectively. The optimal value of ν is chosen to maximize the sum of the contributions of each subject to the log partial likelihood. When the threshold τ is unknown, we can perform a two-dimensional parameter search using CVPL for τ and ν simultaneously.

3. Application to prediction of survival time of patients with DLBCL

To demonstrate the utility of the TGD based Cox regression in relating genomic data to censored survival phenotypes, we re-analyzed a recently published data set of DLBCL by Rosenwald et al. [2]. This data set includes a total of 240 patients with DLBCL, including 138 patient deaths during the followups with median death time of 2.8 years. Rosenwald et al. divided the 240 patients into a training set of 160 patients and a validation set or test set of 80 patients and built a multivariate Cox model. The variables in the Cox model included the average gene expression levels of smaller sets of genes in four different gene expression signatures together with the gene expression level of BMP6. It should be noted that in order to select the gene expression signatures, they performed a hierarchical clustering analysis for genes across all the samples (including both test and training samples). In order to compare our results with those in Rosenwald et al., we used the same training and test data sets in our analysis. In this data set, the gene expression measurements of 7,399 genes (not unique since many genes were spotted multiple times on the arrays) are available for analysis.

Table 1.   Number of features selected for different threshold value τ

Threshold value (τ)	0.0	0.2	0.4	0.6	0.8	1.0
Number of non-zero coefficients	7399	1171	464	140	43	4

3.1. *Effects of the threshold value τ*

Table 1 shows for several threshold τ values, the number of coefficients estimated to be non zero for the training set of 160 patients. For each

threshold value τ, we obtain the optimal predictive point along the path using a 10-fold CVPL. As expected, larger values of τ give rise to fewer non-zero coefficient values, with only 4 out of 7399 features having any influence on survival for $\tau = 1.0$. Figure 1 shows the coefficient values obtained for the DLBCL training data using threshold gradient descent with threshold values $\tau = 0, 0.4, 0.6$ and 1, sorted by the gene order in the original data set of Rosenwald *et al.* All solutions produce relatively large absolute coefficient values in similar regions, with larger values of τ selecting fewer non-zero values within each one. In addition, we clearly observed that smaller τ resulted in smaller absolute coefficients of all the genes, and larger τ resulted in very different estimates of the coefficients among all the genes.

3.2. *Predictive performance*

In order to assess how well the model predicts the outcome, we employ the idea of time dependent receiver-operator characteristics (ROC) curve for censored data and area under the curve (AUC) as our criteria. These methods were recently developed by Heagerty *et al.* [10] in the context of the medical diagnosis and were proposed as criteria for censored data regression with microarray gene expression data [5,4]. Note that larger AUC at time t based on a score function $f(X)$ indicates better predictability of time to event at time t as measured by sensitivity and specificity evaluated at time t. Figure 2 (a) shows the areas under the ROC curves for different threshold values τ based on 10-fold cross-validation for the training data set. This plot suggested that the model with $\tau = 0.4, 0.6$ and 0.8 gave the best predictive results as measured by the areas under the curves. Since the model with $\tau = 0.8$ includes fewer genes in the model, one should choose this model for future prediction. Figure 2 (b) shows the areas under the ROC curves based on the predicted scores for the patients in the testing data set. This plot indicates that $\tau = 0.8$ and $\tau = 1.0$ gave almost the same predictive performance, which implies that the DLBCL data set encourages variability among the coefficients of different genes. The AUCs are between 0.66 and 0.68 in the first 10 years of followups, indicating a reasonable predictive performance. In contrast, smaller values of τ, which correspond approximately to L_2 penalized estimation, gave much lower values of AUCs, indicating worse predictive performance.

To further examine whether clinically relevant groups can be identified by the model, we divided the patients in the test data into two groups based
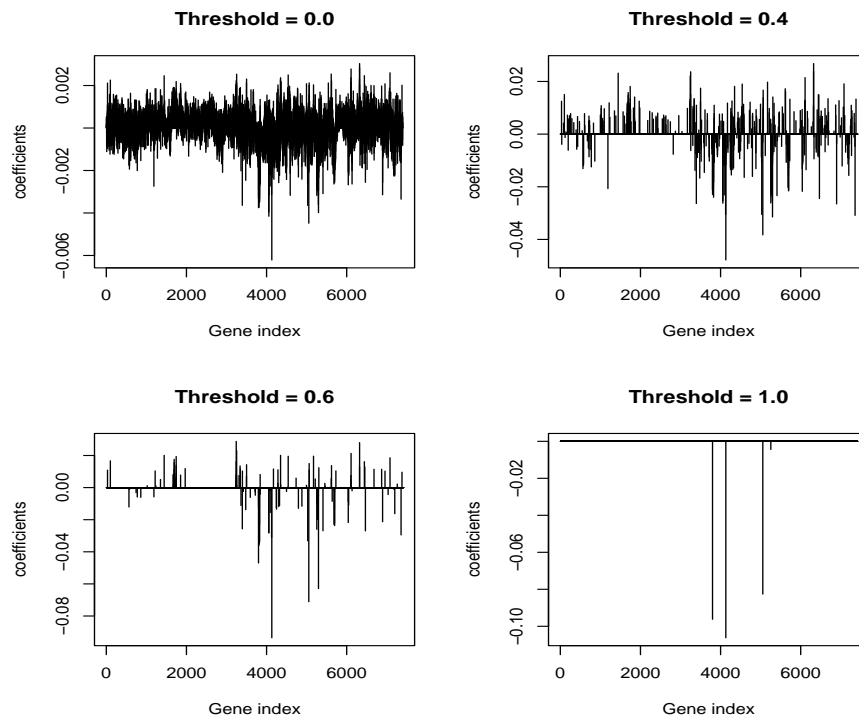
Figure 1.   Coefficient values obtained for the DLBCL data using threshold gradient descent with threshold values $\tau = 0, 0.4, 0.6$ and 1, sorted by the gene order in the original data set of Rosenwald *et al.* All solutions produce relatively large absolute coefficient values in similar regions, with larger values of τ selecting fewer non-zero values within each one.

on their estimated risk scores $\beta' X$ in the Cox model (1) using the mean score as a cutoff value. Figure 3 shows the Kaplan-Meier curves for the two groups of patients, showing very significant difference (p-value=0.0004) in overall survival between the high risk group (36 patients) and low risk group (44 patients). Similar analysis was done using L_2 penalization, less significant difference was observed (p-value=0.003).

3.3. *Genes identified*

As shown previously, cross-validation analysis for the training data set suggested that threshold value $\tau = 0.8$ should result in better predictive perfor-

Table 2.   Genes that were identified to be related to the risk of death when $\tau = 0.8$.

Gene ID	Group	Description
AA286871		tumor necrosis factor receptor superfamily, member 17
AI361763		UDP-Gal:betaGlcNAc beta 1,4- galactosyltransferase, polypeptide 2
AA213564	PS	ribosomal protein S21
AA443696	PS	ribosomal protein S21
AA714637	PS	ribosomal protein S12
AA714637	PS	ribosomal protein S12
AA837360		proapoptotic caspase adaptor protein
AA760674		COX15 homolog, cytochrome c oxidase assembly protein (yeast)
AI087048		interferon regulatory factor 4
W72411		tumor protein p63
H92332	MHC	major histocompatibility complex, class II, DQ alpha 1
AA411017	MHC	major histocompatibility complex, class II, DQ alpha 1
AA159668	MHC	major histocompatibility complex, class II, DQ alpha 1
AA411017	MHC	major histocompatibility complex, class II, DQ alpha 1
AA729055	MHC	major histocompatibility complex, class II, DR alpha
AA032179	MHC	major histocompatibility complex, class II, DR beta 5
AA714513	MHC	major histocompatibility complex, class II, DR beta 5
AA729003		T-cell leukemia/lymphoma 1A
R97095		T-cell leukemia/lymphoma 1A
AA480985	GCB	Weakly similar to germinal center expressed transcript
AA805575	GCB	Weakly similar to A47224 thyroxine-binding globulin precursor
AA278822		Fc receptor-like protein 1
AA485725		immunoglobulin kappa constant
AA487453	PS	GRO2 oncogene
AA598653	LNS	osteoblast specific factor 2 (fasciclin I-like)
LC-29222	LNS	
AA495985	LNS	small inducible cytokine subfamily A (Cys-Cys), member 18
H98765	LNS	cytochrome P450, subfamily XXVIIA, polypeptide 1
AA579913	LNS	leukocyte immunoglobulin-like receptor, subfamily B, member 1
R62612	LNS	fibronectin 1
X14420	LNS	collagen, type III, alpha 1
W87899		aryl hydrocarbon receptor
AI370252		T cell receptor beta locus
AA147638		T cell receptor beta locus
AA147638		T cell receptor beta locus
N29376		myeloid cell nuclear differentiation antigen
AA833786		hemoglobin, alpha 2
W63749		B-cell CLL/lymphoma 2
AA495936		microsomal glutathione S-transferase
H44867		mal, T-cell differentiation protein
AA292532		regulator of G-protein signalling 16
AA804793		ESTs
AA243583		KIAA0084 protein

mance. For $\tau = 0.8$, 43 non-unique genes have non-zero coefficients. Table 2 lists the gene IDs and their descriptions for these 43 genes. Note some genes appear more than once due to replicates on arrays. These genes are
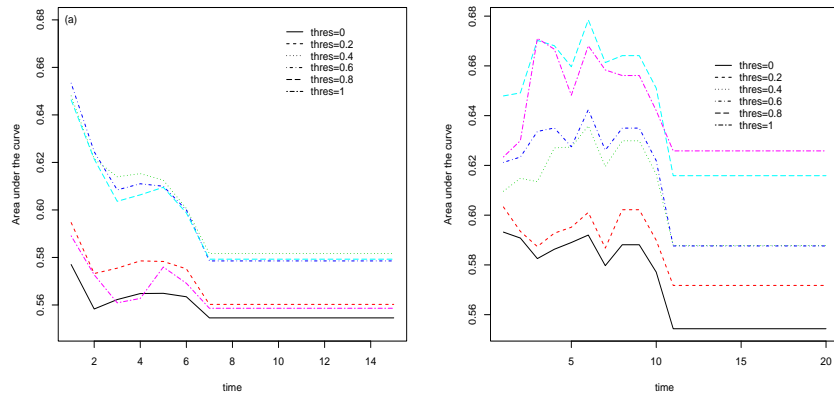
Figure 2.   ROC curves based on (a) the 10-fold cross validated scores based on the training data set and (b) the estimated scores for the 80 patients in the testing data set for different gradient threshold values τ.
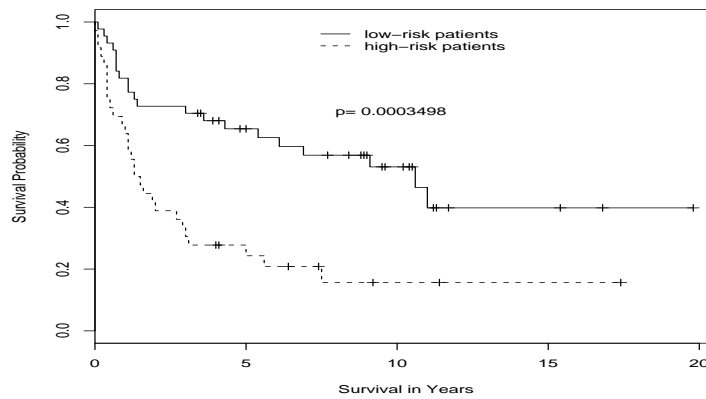

Figure 3.   The Kaplan-Meier curves for the high and low risk groups defined by the estimated scores for the 80 patients in the test data set.

related to the risk of death among the DLBCL patients. It is interesting to note that many of these 43 genes belong to the four signature groups defined by Rowsenald *et al.* using clustering analysis of genes. The four groups are MHC class II (MHC), Proliferation signature (PF), Lymph node

signature (LNS) and Germinal center B-cell signature (GCB). The genes in these groups were shown to be related to the risk of death for the DLBCL patients. The estimated coefficients for the genes in MHC, GCB and LNS signature groups were all negative except for AA495985, indicating that high expression levels of these genes reduce the risk of death among the patients with DLBCL. All the genes in the PF group have positive coefficients. This agrees with what Rosenwald *et al.* (2002) has found. In addition, since many genes were spotted several times on the arrays, it is interesting that these genes were all selected as predictors. Other genes which do not belong to the four signature groups include T-cell receptor beta locus, T-cell leukemia/lymphoma 1A and T-cell differentiation protein.

As a comparison, when $\tau = 1.0$, only four genes (H92332, LC_29222, AA480985, H98765) were selected by the model. These four genes belong to three of the four signature groups and have large absolute coefficients. This is in direct contrast with the genes selected by the TGD Cox regression with $\tau = 0.8$, where genes which are highly correlated were also selected.

4. Discussion and Conclusions

In pharmacogenomics, one important problem is to predict patient's time to cancer relapse or time to death due to cancer after treatment using genomic profiles of the cancerous cells prior to the treatment. Powerful statistical methods for such prediction allow for high-dimensional genomic data such as microarray gene expression data to be used most efficiently. In this paper, we have extended the threshold gradient descent based method [9] for censored survival data in order to identify important predictive genes for survival using high throughput genetic or genomic data. Since the risk of cancer recurrence or death due to cancer may result from the interplay between many genes, methods which can utilize data of many genes, as in the case of our proposed method, are expected to show better performance in predicting risk. We have demonstrated the applicability of our methods by analyzing time to death of the diffuse large B-cell lymphoma patients and obtained satisfactory results, as evaluated by both applying the model to the test data set and the time dependent ROC curves. Our simulations also indicated better predictive performance of the proposed method than other penalization methods (results not shown due to page limitation).

As we observed in our analysis of the DLBCL data set, if there is a group of variables or genes among which the pairwise correlations are very high, the procedure with $\tau = 1$ tends to select only one variable from the group

and does not care which one is selected. Such procedure, which is very closely related to Lasso, can give a good predictive performance. However, for genes sharing the same biological pathways, the correlations among them can be high. If the goal is to select important and relevant genes, one may want to include all these highly correlated genes, if one of them is selected. We observed that the TGD procedure with $\tau = 0.8$ precisely achieved this goal. In addition, we may expect more robust prediction since the gene expression levels of highly correlated genes are used in the model. Therefore, in practice, one should use cross validation to select the optimal threshold value τ.

Some possible extensions of the proposed methods are worth mentioning. One is to study the treatment effect adjusting for genetic or genomic profiles. This can be done by including a treatment indicator variable in the Cox regression model. However, regularization by threshold gradient is only applied to the variables related the high-dimensional genomic profiles. A profile-type penalization can then be developed. Another interesting problem is to identify certain genetic profiles which may interact with the treatment in determining the risk of certain clinical event. This can be done by including all the treatment by gene interactions in the model and using the threshold gradient descent method to select the relevant genes and their interactions. Both approaches deserve further investigation.

In summary, the proposed threshold gradient descent method for the Cox model can be very useful in pharmacogenomics in building a parsimonious predictive model of risk based on the genetic or genomic profiles. The procedure is robust and numerically stable and can also be applied to select important genes that are related to patients' survival outcome.

References

1. Golub, T.R., Slonim, D.K., Tamayo, P., *et al.* 1999 *Science* 286,531-537.
2. Rosenwald, A., Wright, G., Chan, W., *et al.* 2000 *The New England Journal of Medicine* 346,1937-1947.
3. Cox, D.R. 1972. *Journal of the Royal Statistical Society. Series B* 34,187-220.
4. Li, H. and Gui, J. 2004. *Bioinformatics* 20, i208-215.
5. Li, H. and Luan, Y. 2003. *Pacific Symposium on Biocomputing* 8,65-76.
6. Tibshirani, R. 1995. *Journal of Royal Statistical Society B* 58,267-288.
7. Efron B., Johnston I., Hastie T. and Tibshirani R. 2004. *Annals of Statistics.*
8. Gui J. and Li H. 2004. *Technical report, UC Davis.*
9. Friedman, J.H. and Popescu B.E. 2004. *Technical Report, Statistics Department, Stanford University.*
10. Heagerty, P.J., Lumley, T. and Pepe, M. 2000. *Biometrics* 56,337-344.