Clustering Threshold Gradient Descent Regularization: with Applications to Microarray Studies

Shuangge Ma [a] and Jian Huang [b]

[a]Department of Biostatistics, University of Washington, Seattle, WA, USA [b] Departments of Statistics and Actuarial Science, and Biostatistics, University of Iowa, Iowa City, IA, USA

ABSTRACT

Motivation: An important application of microarray technology is to discover important genes and pathways that are correlated with clinical outcomes such as disease status and survival. While a typical microarray experiment surveys gene expressions on a global scale, there may be only a small number of genes that have significant influence on a clinical outcome of interest. In addition, expression data have cluster structures and the genes within a cluster have coordinated influence on the response, but the effects of individual genes in the same cluster may be different. Accordingly, we seek to build statistical models with the following properties. First, the model is sparse in the sense that only a subset of the parameter vector is non-zero. Second, the cluster structures of covariates (genes) are properly accounted for.

Results: For microarray studies with smooth objective functions and well defined cluster structure for genes, we propose a clustering threshold gradient descent regularization (CTGDR) method, for simultaneous cluster selection and within cluster gene selection. We apply this method to regression models for binary classification and censored survival data with microarray gene expression data as covariates, assuming known cluster structures of gene expressions. Compared to the standard TGDR and other regularization methods, the CTGDR takes into account the cluster structure and carries out feature selection at both the cluster level and within-cluster gene level. We demonstrate the CTGDR on two studies of cancer classification using microarray data and two studies of correlating survival of lymphoma patients with microarray data.

Availability: Research R code is available upon request from the authors.

Contact: shuangge@u.washington.edu or jian@stat.uiowa.edu

1 INTRODUCTION

Microarray technology provides a way of monitoring gene expressions on a large scale. Tremendous efforts have been devoted to discover genes and pathways that are accountable for variations of clinical outcomes. An understanding of the molecular biology that underlies such variations might provide a more accurate method of diagnosis and might suggest new therapeutic approaches, see for example, Alizadeh et al. (2000), Garber et al. (2001), and Rosenwald et al. (2003), among others. Two types of clinical outcomes have been of special interest. The first type is categorical outcome, which includes the presence or absence of tumor as in Alon et al. (1999) or different types of tumors as in Alizadeh et al. (2000). The second type is survival outcome, which usually corresponds to the occurrence time of certain event such as cancer. See for example Rosenwald et al. (2003) and Dave et al. (2004).

Classification and survival analysis using microarray data are challenging due to the large number of genes and relatively small sample size in a typical study. Various model reduction or variable selection methods have been proposed, including the singular value decomposition (Golub and Van Loan 1996), partial least squares (Nguyen and Rocke 2002), principal component analysis (Ma et al. 2006), LASSO-LARS (Gui and Li 2005a), and threshold gradient descent regularization (TGDR, Gui and Li 2005b; Ma and Huang 2005) among others. The essence of the aforementioned regularization techniques is to identify a small number of representative features–individual genes or linear combinations of genes, and build predictive models based on those representative features. In the feature selection stage, all genes are treated equally and the intrinsic gene structures are usually ignored.

It has been demonstrated that cluster structure exists in gene expression data (Eisen et al. 1998), and the clusters based on expression data tend to correspond to certain gene pathways (Clare and King 2002; Tavazoie et al. 1999; Yeung et al. 2001). Cluster analysis methods have been employed in gene expression studies as a dimension reduction tool (Alizadeh et al. 2000; Dave et al. 2004). With this approach, a small number of clusters based on gene expression data are first constructed, using methods such as the k-mean or hierarchical method (Johnson and Wichern 2002). The means of the expressions of genes within the same clusters are then computed and used as covariates for downstream model building. A limitation of this approach is that feature selection is carried out only at the cluster level. Once a cluster is used in the final model, all genes within that cluster are included. Although genes within the same cluster may have similar functions, it is not necessarily true they will all be associated with a specific clinical outcome. Including "noisy" genes may lead to ill-behaved models. Gene selection within clusters is still of interest to yield more interpretable models. Wei and Li (2006) proposed a nonparametric pathway-based regression approach for genomic data that explicitly make use of available pathway information in their model. They used the gradient-based boosting algorithm (GDB, Friedman 2001) for model fitting and the importance score (Breiman et al. 1984; Friedman 2001) for ranking pathways and genes. However, they did not explicitly consider variable selection at either the cluster or individual gene levels.

Regularization methods such as the LASSO and TGDR can be used for variable selection. Although capable of selecting a small number of important genes, these methods do not incorporate cluster structure. On the other hand, standard model fitting approaches using cluster analysis results as input explicitly take into account cluster structure, but cannot carry out individual gene selection. To combine the strength of the aforementioned approaches, we propose a clustering TGDR (CTGDR) method that incorporates cluster structure into TGDR-based variable selection. The proposed CTGDR carries out feature selection at two levels: at the cluster level and the individual gene level in each cluster. When there exist well defined gene clusters, the CTGDR is capable of selecting important clusters and genes within clusters simultaneously. Thus it takes advantages of both the cluster-based and regularized variable selection methods.

The rest of this paper is organized as follows. In section 2, we present the data and models we consider. We use logistic regression for binary classification and Cox model for right censored survival analysis as examples. The CTGDR algorithm is described in section 3. Tuning parameter selection and evaluation are also discussed. We present two classification examples in section 4 and two survival analysis examples in section 5. The article ends with a discussion in section 6. Part of the data analysis results are presented in the Supplementary Data.

2 DATA AND MODEL SETTINGS

Let Z be a length d vector of gene expressions, and let Y be the clinical outcome of interest. We assume that Y is associated with Z through a parametric or semiparametric model $Y \sim \phi(\beta' Z)$ with a regression function ϕ and unknown regression coefficient β. In addition, we assume there exists a smooth objective function and a proper estimate of β can be obtained by maximizing that function. In regularized estimation, gene selection is achieved if some components of β are estimated to be exactly zero. We are particularly interested in the classification and survival analysis problems using microarray gene expression data as covariates due to their extensive applications in medical studies.

For the classification problems, Y is the categorical variable indicating the disease status. For simplicity, we focus on binary classification only. Suppose that $Y = 1$ denotes the presence and $Y = 0$ indicates the absence of disease. We assume the commonly used logistic regression model, where the logit of the conditional probability is logit$(P(Y = 1|Z)) = \alpha + \beta' Z$, where β is the length d vector of unknown regression coefficient and α is the unknown intercept. Based on a sample of n iid observations $X_i = (Y_i, Z_i), i = 1, \dots, n$, the maximum likelihood estimator is defined as $(\hat{\alpha}, \hat{\beta}) = argmax_{\alpha,\beta} R_n(\alpha, \beta)$, where

$$\begin{aligned} R_n(\alpha,\beta) &= \sum_{i=1}^{n} Y_i \log \frac{\exp(\alpha+\beta' Z_i)}{1+\exp(\alpha+\beta' Z_i)} + \qquad (1) \\ & \quad (1-Y_i)\log \frac{1}{1+\exp(\alpha+\beta' Z_i)} . \end{aligned}$$

We are mainly interested in the estimation of β. The intercept α will always be kept in the model. For simplicity, we denote $R_n(\alpha, \beta)$ as $R_n(\beta)$.

For right censored survival data, $Y = (T, \Delta)$, where $T = min(U, V)$ and $\Delta = I(U \le V)$. Here U and V denote the event time of interest and the censoring time, respectively. The most widely used model for censored data is the Cox proportional hazards model (Cox, 1972) which assumes that conditional hazard function $\lambda(u|Z) = \lambda_0(u)\exp(\beta' Z)$, where λ_0 is the unknown baseline function and β is the regression coefficient. Based on a sample of n iid observations $X_i = (Y_i, Z_i), i = 1, \dots, n$, the maximum partial likelihood estimator is defined as the value $\hat{\beta}$ that maximizes

$$R_n(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp(Z_i'\beta)}{\sum_{j\in r_i}\exp(Z_j'\beta)} \right\}^{\delta_i},$$

where $r_i = \{j : T_j \ge T_i\}$ is the risk set at time T_i.

In the above examples, the objective functions R_n are smooth and depend only on data and the unknown regression coefficient β. Assume there are L well defined clusters associated with the gene expressions, and gene $j = 1, \dots, d$ belongs to one of the clusters $C(j) \in \{1, \dots, L\}$. The clusters can be defined based on biological functions or statistical associations or both. We also allow overlap between different clusters. We assume that the clusters have been defined *a priori*.

3 CLUSTERING TGDR

The CTGDR can be consider a generalization of the TGDR, which is introduced by Friedman and Popescu (2004) in the context of linear regression analysis and has been employed in microarray studies by Gui and Li (2005b) and Ma and Huang (2005). For completeness of this paper, we first briefly review the TGDR algorithm.

3.1 TGDR algorithm

Denote $\Delta\nu$ as the small positive increment as in ordinary gradient descent methods (Friedman and Popescu 2004). In the implementation of this algorithm, we choose $\Delta\nu = 1 \times 10^{-4}$. Denote $\nu_k = k \times \Delta\nu$ as the index for the point along the parameter path after k steps. Let $\beta(\nu_k)$ denote the parameter estimate corresponding to the index ν_k. For any fixed threshold value $0 \le \tau \le 1$, the TGDR algorithm consists of the following iterative steps:

1. Initialize $\beta(0) = 0$ and $\nu_0 = 0$.
2. For the current estimate β, compute the negative gradient $g(\nu) = -\partial R_n(\beta)/\partial\beta$. Denote the j^{th} component of $g(\nu)$ as $g_j(\nu)$. If $\max_j\{|g_j(\nu)|\} = 0$, stop the iterations.
3. Compute the threshold vector $f(\nu)$ of length d, where the j^{th} component of $f(\nu)$: $f_j(\nu) = I\{|g_j(\nu)| \ge \tau \times \max_l |g_l(\nu)|\}$.
4. Update $\beta(\nu + \Delta\nu) = \beta(\nu) - \Delta\nu \times g(\nu) \times f(\nu)$ and update ν by $\nu + \Delta\nu$, where the product of f and g is component-wise.
5. Steps 2–4 are repeated k times. The number of iterations k is determined by cross validation as described below.

The tuning parameters τ and k jointly determine the property of β. When $\tau \approx 0$, β is dense even for small values of k. When $\tau \approx 1$, β is sparse for small k and remains so for a relatively large number of iterations, but will become dense eventually. At the extreme when $\tau = 1$, the TGDR usually increases in the direction of a single covariate in each iteration. This mimics the incremental forward stage-wise strategy in Hastie, Tibshirani and Friedman (2001). When τ is in the middle range, the characteristics of β are between those for $\tau = 0$ and $\tau = 1$. For $\tau \ne 0$, variable selection can be achieved with cross validated, finite k, by having certain components of β exactly zero. We refer to Friedman and Popescu (2004) for more detailed discussions.

3.2 CTGDR method

The TGDR described above is capable of individual gene selection but not accounting for the cluster structures. We now describe the proposed CTGDR algorithm, starting with a simple modification of the TGDR.

3.2.1 Naive CTGDR

Naive CTGDR Algorithm I. This algorithm modifies step 3 of the TGDR as follows:

$$f_j^1(\nu) = I\left\{\sum_{m\in C(j)} |g_m(\nu)| \geq \tau_1 \times \max_{C(k)} \sum_{l\in C(k)} |g_l(\nu)|\right\}, \qquad (2)$$

where $0 \leq \tau_1 \leq 1$ is the threshold tuning parameter. The other steps in the TGDR are kept unchanged.

Compared to the original TGDR, algorithm I uses "cluster gradients" to replace the individual gradients. The combined effects of genes in the same clusters are considered and compared with the combined effects of other clusters. This algorithm is similar to the traditional clustering based approaches in the sense that gene selection is achieved on a cluster basis, and if the combined effect of genes in a cluster is important, then all the genes within this cluster will be included in the final model (Dave et al. 2004). The key difference is that the naive CTGDR I estimated coefficients of genes in the same clusters may be different. So genes within the same clusters still have different contributions in the final model, whereas in traditional cluster based methods, all genes within the same clusters have the same coefficient and hence equal contributions to the outcome.

Algorithm I does feature selection at the cluster level. If a cluster is selected, then all the genes in this cluster will be included. The total number of genes in the final model can be large. Consider for an example a hypothetical study with 2000 genes and five clusters of equal sizes are constructed. Then using algorithm I, it is possible three or four clusters are selected. The total number of genes in the final model will be greater than 1000. Although the prediction performance may still be satisfactory, this makes the final estimation results hard to interpret from a gene discovery point of view. Since it is often the case that only a subset of genes within each cluster have important impact on the outcome of interest, gene selection within cluster is still needed.

Naive CTGDR Algorithm II. This algorithm partly solves the drawbacks of algorithm I. Denote $\tau_2 \in [0, 1]$ as the threshold tuning parameter. We replace f in step 3 of the TGDR with

$$f_j^2(\nu) = I \quad |g_j(\nu)| \geq \tau_2 \times \max_{l\in C(j)} |g_l(\nu)| \quad , \qquad (3)$$

so that each gene is compared with other genes within the same cluster and only important genes from each cluster are selected. The rationale is that genes from different clusters/pathways may not be directly comparable. So a fair comparison will be for genes within the same clusters. Within each cluster, we use the TGDR to identify important genes.

We have employed algorithm II to the examples in sections 4 and 5. We were able to identify a smaller number of genes ($\sim$200, much fewer than that from naive algorithm I) with satisfactory prediction performance. However, algorithm II has it own drawbacks. It is roughly equivalent to carrying out the TGDR in each cluster separately and the final model includes genes selected from all clusters. The underlying assumption is that all clusters are associated with the outcome of interest. Previous cluster based methods as in Dave et al. (2004) and Alizadeh et al. (2000) show that this is not necessarily true. Cluster selection may still be necessary.

3.2.2 CTGDR algorithm The naive CTGDR algorithm I carries out cluster selection, but does not select important genes within each cluster. On the other hand, the naive CTGDR algorithm II does gene selection in each cluster separately, but does not select clusters. The advantages and drawbacks of the naive CTGDR algorithms motivate the following CTGDR algorithm.

Let $\tau_1,\ \tau_2 \in [0, 1]$ be two threshold parameters. In step 3 of the TGDR algorithm, define

$$f_j(\nu) = f_j^1(\nu) \times f_j^2(\nu), \qquad (4)$$

where $f^1(\nu)$ is defined in (2) with threshold value τ_1 and $f^2(\nu)$ is defined in (3) with threshold value τ_2, respectively.

In (4), the term $f^1(\nu)$ carries out cluster selection, while $f^2(\nu)$ carries out within-cluster gene selection. So the combined f can carry out feature selection at both the cluster level and within cluster level. Further flexibility is introduced by allowing two possibly different threshold values. In this algorithm, if a gene or a cluster is known to be associated with the clinical outcome *a priori*, then it can be excluded from the thresholding step.

The three tuning parameters k, τ_1 and τ_2 jointly determine the properties of the CTGDR estimates, as can be seen from numerical studies in sections 4 and 5 (see Tables 1 and 2). Roughly speaking, the tuning parameters τ_1 and τ_2 have similar effects as the tuning parameter τ for the standard TGDR in section 3.1. If τ_1 and τ_2 are both close to 1, then the estimate remains sparse for a relatively large k, but will become dense eventually. If τ_1 and τ_2 are both close to 0, the estimate is dense for even a very small k. τ_1 and τ_2 determine the degree of sparsity on cluster level and within cluster level, respectively, with larger thresholding values leading to more parsimonious models with fixed k. With nonzero τ_1 and τ_2, the model with small to moderate k usually has a small number of clusters and a small number of genes within each selected cluster.

3.2.3 Possible extensions In the above CTGDR algorithm, the cluster gradient is simply defined as the sum of absolute values of individual gradients. This is the default definition when there is no extra information on the clusters. If there exists external knowledge of the relative importance of clusters, then we can modify the indicator function in (2) as

$$I\left\{w_j \sum_{m\in C(j)} |g_m(\nu)| \geq \tau_1 \times \max_{C(k)} w_k \sum_{l\in C(k)} |g_l(\nu)|\right\}, \qquad (5)$$

where w_js are positive weights measuring the relative importance of cluster j. A simple choice of w_j is the inverse of cluster size, so that the relative importance of clusters is not affected by cluster size. Choosing different weights has considerable impact on the cluster selection results. If external knowledge about the relative importance of genes within the same cluster is present, then the cluster gradient can be defined as the weighted sum of individual gradients, with more stable and more important genes having larger weights. Further flexibility is introduced by considering weighted gradients. We leave study of the weighted scheme to a future article.

3.3 Tuning parameter selection

We select the tuning parameters k and (τ_1, τ_2), which jointly determine the characteristics of the estimator, using the following two-step approach.

First we choose the tuning parameter k for any fixed (τ_1, τ_2) using V-fold cross validation (Wahba 1990) as follows. Partition the data randomly into V non-overlapping subsets of equal sizes. Choose k to maximize the cross-validated objective function

$$CV(k) = \sum_{v=1}^{V} \left[R_n(\beta^{(-v)}) - R_n^{(-v)}(\beta^{(-v)}) \right], \qquad (6)$$

where $\beta^{(-v)}$ is the CTGDR estimate of β based on the data without the v^{th} subset for a fixed k and $R_n^{(-v)}$ is the objective function R_n evaluated without the v^{th} subset.

After cross validation over k, model features for different τ_1 and τ_2 can be obtained as for example shown in Tables 1 and 2. We choose parsimonious models with relatively large CV score. An AIC type score as in Huang et al. (2005) can be used as model selection criterion. Cross validation over τ_1 and τ_2 can also be considered, i.e, we can select the model with the largest CV score over all possible k, τ_1 and τ_2. However, this approach may lead to models with slightly larger CV scores, but a lot more genes, which correspond to unstable models. Beyond selecting the model (corresponding to the cross validated tuning parameters) with the best predictive performance, V-fold cross validation also provides partial protection against overfitting (Nguyen and Rocke 2002).

3.4 Evaluation

Unlike in standard classification or survival analysis where the association between clinical outcome and covariates is of primary interest, studies given in sections 4 and 5 put more emphasis on variable selection and prediction based on selected genes. So we consider the following cross validation based approach for evaluating prediction performance, as suggested by Ma and Huang (2005).

1. We first partition the data randomly into a training set of size n_1 and a testing set of size n_2 with $n_1 + n_2 = n$. In this article, we set $n_1 \sim 2/3n$.
2. Compute the CTGDR estimate based on the training set only. Using this training set estimate, we compute a prediction index for the testing set.
3. To take into account the possibility of an extreme prediction performance due to a rare partition, we repeat this process B (for example 1000) times. Each time a new partition is made and the prediction index is computed.

For classification studies, the prediction index can be the prediction error or the prediction AUC from a ROC analysis (Ma and Huang 2005). For censored survival studies, we first create two risk groups based on dichotomizing the estimated linear risk scores $\hat{\beta}' Z_i$ at the median risk score for the testing set. Note that we can define multiple risk groups based on the quantiles of the linear risk scores. We then use the logrank statistic to test whether the survival curves of the different risk groups are different. A large value of the logrank statistic indicates that the high and low risk groups are well separated, and suggests satisfactory prediction performance of the CTGDR estimate.

For censored survival analysis, we also consider model fitting evaluation based on the time-dependent ROC, which was proposed by Heagerty et al. (2000) in the context of the medical diagnosis and has been used as criteria for censored data regression with microarray gene expression data (Gui and Li 2005b). The essential idea is to treat the event indicator as binary outcome for each time point and evaluate the classification performance at each time using the standard ROC technique. In the ROC approach, the AUC can be used as the evaluation/comparison criteria and a larger AUC at time u indicates better predictability of the survival outcome at time u as measured by sensitivity and specificity.

Table 1. Colon and Estrogen data. Model features for different tuning parameters. variable: number of selected genes; cluster: number of selected clusters.

			Colon			Estrogen	
τ_1	τ_2	CV	variable	cluster	CV	variable	cluster
1.0	1.0	**-24.4**	**16**	**5**	-17.6	18	5
	0.9	-24.6	30	5	-16.9	24	5
	0.8	-25.3	34	4	-16.3	39	5
	0.7	-25.5	76	6	-15.2	58	5
0.9	1.0	-24.7	16	5	**-16.7**	**18**	**6**
	0.9	-26.2	35	6	-16.2	28	6
	0.8	-26.9	49	7	-15.0	47	6
	0.7	-27.9	85	6	-13.5	85	9

4 BINARY CLASSIFICATION EXAMPLES

Colon data. In this dataset, expression levels of 40 tumor and 22 normal colon tissues for 6500 human genes are measured using the Affymetrix gene chips. A selection of 2000 genes with the highest minimal intensity across the samples has been made by Alon et al. (1999), and these data are publicly available at *http://microarray.princeton.edu/oncology/*. The colon data have been analyzed in several previous studies using other statistical approaches, see for example Dettling and Buhlmann (2003), Pochet et al. (2004), Ben-Dor et al. (2000), Nguyen and Rocke (2002) and Ma and Huang (2005).

Estrogen data. This dataset was first presented by West et al. (2001) and Spang et al. (2001). It includes expression values of 7129 genes of 49 breast tumor samples. The expression data were obtained using the Affymetrix gene chip technology and are available at *http://mgm.duke.edu/genome/dna_micro/work/*. The response describes the response of the estrogen receptor (ER). Among the 49 samples, 25 are positive (ER+) and 24 are negative (ER−). We threshold the raw data with a floor of 100 and a ceiling of 16000. Genes with $\max(expression)/\min(expression) < 10$ and/or $\max(expression) - \min(expression) < 1000$ are also excluded (Dudoit et al. 2002). 3332 (46.7%) genes pass the first step screening. A base 2 logarithmic transformation is then applied. The estrogen data have also been studied by Dettling and Buhlmann (2003) and Ma and Huang (2005).

Although there is no limitation on the number of genes that can be used in the CTGDR, we first identify 500 genes for each dataset based on marginal significance to gain further stability as in Ma and Huang (2005). Compute the sample standard errors of the d biomarkers $se_{(1)}, \ldots, se_{(d)}$ and denote their median as $med.se$. Compute the adjusted standard errors as $0.5(se_{(1)} + med.se), \ldots, 0.5(se_{(d)} + med.se)$. Then the genes are ranked based on the t-statistics computed with the adjusted standard errors.

The 500 genes with the largest absolute values of the adjusted t-statistics are used for classification. The adjusted t-statistic is similar to a simple shrinkage method discussed in Cui et al. (2005). For each of these two datasets, we constructed 25 clusters based on gene expressions using the k-mean approach. The cluster sizes are not exactly equal but comparable. Details of the cluster structures are available from the authors.

We apply the CTGDR to the clustered data obtained above. We consider the tuning parameters τ_1 and τ_2 taking values in the grid 0, 0.1,..., 1.0. Partial model features for different thresholding values and cross validated k are shown in Table 1. For the colon data, the model with $\tau_1 = 1.0$ and $\tau_2 = 1.0$ is identified as the final model. For the estrogen data, the final model has $\tau_1 = 0.9$ and $\tau_2 = 1.0$. 16 and 18 genes are identified, respectively, representing 5 and 6 clusters. The estimated coefficients and gene description for the final models are given in the Supplementary Data.

We evaluate the prediction performance of the CTGDR using the approach discussed in section 3.4. Based on 1000 partitions, the prediction errors for the testing sets have means 0.12 (0.07) and 0.09 (0.07), respectively, where the numbers in "()" are the standard errors. For the colon data, the CTGDR provides a similar prediction as the SMRC in Ma and Huang (2005, mean classification error 0.14), and better performance than boosting (Dettling and Buhlmann 2003, mean classification error: LogitBoost 0.16; AdaBoost 0.18), classification tree (Dettling and Buhlmann 2003, mean classification error 0.15) and SVM (Pochet, et al. 2004, mean classification error 0.18). For the estrogen data, the CTGDR still provides a satisfactory prediction. However, it is less optimal than the SMRC in Ma and Huang (2005, mean classification error 0.06) and boosting in Dettling and Buhlmann (2003, mean classification error: LogisBoost 0.04; AdaBoost 0.04).

5 SURVIVAL ANALYSIS EXAMPLES

Follicular Lymphoma data. Follicular lymphoma is the second most common form of non-Hodgkin's lymphoma, accounting for about 22 percent of all cases. A study was conducted to determine whether the survival probability of patients with follicular lymphoma can be predicted by the gene-expression profiles of the tumors at diagnosis (Dave et al. 2004). Fresh-frozen tumor-biopsy specimens and clinical data from 191 untreated patients who had received a diagnosis of follicular lymphoma between 1974 and 2001 were obtained. The median age at diagnosis was 51 years (range 23 to 81), and the median follow up time was 6.6 years (range less than 1.0 to 28.2). The median follow up time among patients alive at last follow up was 8.1 years. Eight records with missing survival information are excluded from the downstream analysis. Detailed experimental protocol can be found in Dave et al. (2004).

Affymetrix U133A and U133B microarray genechips were used to measure gene expression levels from RNA samples. A log2 transformation was applied to the Affymetrix measurements. We first filter the 44928 gene measurements with the following criteria: (1) the max expression value of each gene across 191 samples must be greater than 9.186 (the median of the maximums of all probes). (2) the max-min should be greater than 3.874 (the median of the max-min of all probes). (3) Compute correlation coefficients of the uncensored survival times with gene expressions. Select the genes whose correlation with survival time is greater than 0.2. There are 729 genes that pass this screening process. We normalize genes across samples to have mean 0 and variance 1.

Table 2. Follicular and MCL data. Model features for different tuning parameters. variable: number of selected genes; cluster: number of selected clusters.

		Follicular			MCL		
τ_1	τ_2	CV	variable	cluster	CV	variable	cluster
1.0	1.0	-182.5	129	6	**-83.4**	**106**	**9**
	0.9	-183.4	149	6	-83.2	141	7
	0.8	-182.2	180	6	-83.7	170	4
	0.7	-181.5	268	10	-83.1	238	5
0.9	1.0	**-178.6**	**127**	**7**	-83.7	113	10
	0.9	-180.5	169	9	-83.5	143	10
	0.8	-180.5	202	10	-83.9	201	12
	0.7	-180.0	239	9	-83.5	392	12

Mantel Cell Lymphoma data. Rosenwald et al. (2003) reported a study using microarray expression analysis of mantle cell lymphoma (MCL). The primary goal of this study was to discover genes that have good predictive power of patient's survival risk. Among 101 untreated patients with no history of previous lymphoma included in this study, 92 were classified as having MCL, based on established morphologic and immunophenotypic criteria. Survival times of 64 patients were available and other 28 patients were censored. The median survival time was 2.8 years (range 0.02 to 14.05 years). Lymphochip DNA microarrays (Alizadeh et al., 2000) were used to quantify mRNA expression in the lymphoma samples from the 92 patients. The gene expression data that contains expression values of 8810 cDNA elements is available at *http://llmpp.nih.gov/MCL.*

We pre-process the data as follows to exclude noises and gain further stability: (1) Compute the variances of all gene expressions; (2) Compute correlation coefficients of the uncensored survival times with gene expressions; and (3) Select the genes with variances larger than the first quartile and with correlation coefficients larger than 0.2. 1451 out of 8810 genes (16.5%) pass the above initial screening. We standardize these genes to have zero mean and unit variance. We follow previously suggested first stage screening methods. So they are slightly different for the two datasets.

For each dataset, we obtain 20 clusters using the k-mean method, following a protocol similar to that in Dave et al. (2004). Cluster structures are available upon request. We employ the proposed CTGDR and partial model features for different threshold values are shown in Table 2. Using the two-step approach proposed in section 3.3, we identify the models with $(\tau_1, \tau_2) = (0.9, 1.0)$ for Follicular Lymphoma data and $(1.0, 1.0)$ for MCL data, respectively. Detailed descriptions of the genes appeared in the final model are provided in the Supplementary Data.

We use the approaches discussed in section 3.4 for evaluation and comparison. For comparison, we consider estimates obtained using Dave's approach and the standard TGDR method. With the TGDR, the models with $\tau = 1.0$ are selected for both models. However, they only have 42 (Follicular) and 28 (MCL) genes. To make a fair comparison, we also consider the TGDR models with $\tau = 0.55$ (Follicular) and $\tau = 0.80$ (MCL), which have 123 and 93 genes, respectively–much closer to those for the CTGDR.

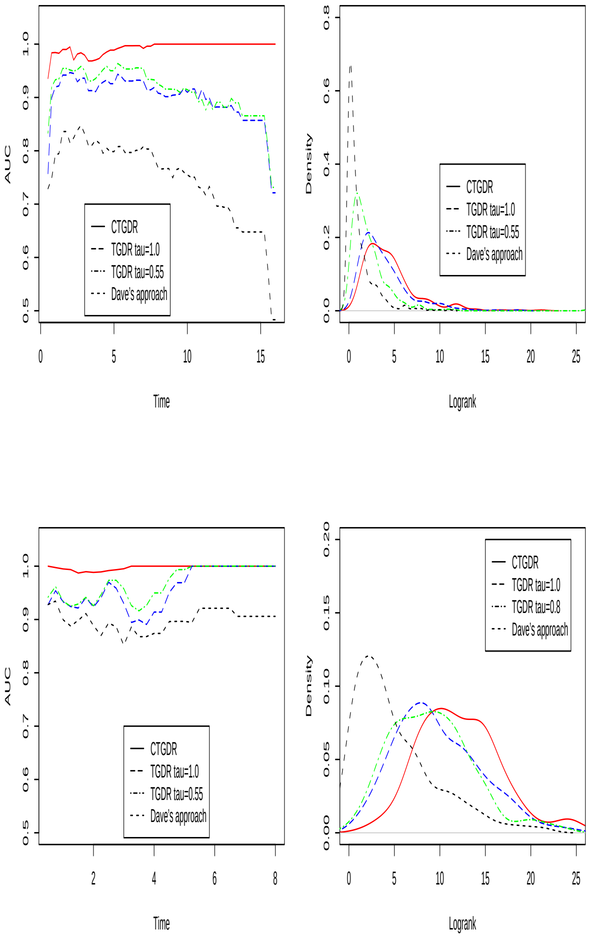

Fig. 1. Follicular lymphoma (upper panels) and MCL data (lower panels). Plot of time-dependent AUCs (left panels) and the kernel density estimates of the logrank statistics (right panels).

We show in Figure 1 the time-dependent ROC plot and the kernel density estimation of the logrank statistics. We can see that the CTGDR has dominating AUCs, which suggest better model fitting for both datasets. We note that the AUCs for the CTGDR are very close or equal to 1. This is partly caused by the relatively large number of genes identified. The logrank statistics for the Follicular data have means 4.502 (CTGDR), 3.836 (TGDR, $\tau = 1.0$), 2.258 (TGDR, $\tau = 0.55$) and 1.081 (Dave's). We compare the CTGDR logrank statistics with those from other approaches using the Wilcoxin tests. For the Follicular Lymphoma data, we obtain p-values < 0.001 for TGDR and Dave's approach, which suggests significantly better prediction performance of the CTGDR. For the MCL data, the logrank statistics have means 12.142 (CTGDR), 9.764 (TGDR, $\tau = 1.0$), 9.101 (TGDR, $\tau = 0.8$) and 5.131 (Dave's). The corresponding p-values from Wilcoxin tests are < 0.001.

6 DISCUSSIONS

The proposed CTGDR approach carries out variable selection at the cluster and individual gene levels simultaneously, and directly accounts for cluster structure in microarray gene expression data. This algorithm is quite flexible in that it can use any clustering results as input in the analysis. The CTGDR is different from the existing variable selection methods applied to microarray data which focus on feature selection at the individual gene level. We used logistic regression for classification and Cox model for survival data as examples to illustrate the application of the CTGDR. However, the CTGDR algorithm does not depend on the actual form of the objective function, as long as it is well defined and differentiable. So the CTGDR can be used in survival analysis with other models such as the accelerated failure time and additive hazards models, and classification analysis based other objective functions such as the SVM hinge loss and the ROC objective function.

We have demonstrated the proposed approach on four publicly available datasets. In these examples, we constructed the clusters based on the expression data. Co-expressed genes in the same cluster are likely to be involved in the same cellular processes, so a strong correlation of expression patterns between those genes indicates co-regulation. However, gene clusters obtained based on expression data may not completely overlap with functional groups and pathways. If there is external information on the pathways to which the genes under study belong, such external information should be used to form clusters. In most applications, it is probably the case that only partial external information are available, i.e., some genes are known to belong to certain pathways, but for most genes such information are not available. In this case, a simple solution is to use the expression data to cluster the genes without the external pathway information. Since the CTGDR can use any clustering result as input in the model fitting, we can incorporate existing biological information into the analysis. Indeed, incorporating existing pathway information should improve the variable selection and prediction performance. We note that the CTGDR does not require that there is no overlap in gene clusters. Thus it is applicable to the situation when some genes play a role in multiple functional pathways.

We have only considered classification and survival models in which the outcome variable depends on a simple linear combination of the gene expression data. The CTGDR is applicable to more complicated models which may include nonparametric and nonlinear components. It is also applicable to models with interaction at both the cluster and individual gene levels. Such models would probably be more realistic from a biological standpoint. However, it is important not to make the models overly complex given the limited amount of data in a typical study. We plan to consider such issues in future studies.

ACKNOWLEDGMENT

The work of Ma is partially supported by the NIH grant N01-HC-95159. The work of Huang is supported in part by the NIH grant HL72288.

REFERENCES

Alon, U., Barkai, N., Notterman, D., Gish, K., Mack, S. and Levine, J. (1999) Broad Patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS* **96**, 6745–6750.

Alizadeh, A.A., Eisen M.B., Davis R.E., Ma C., et al. (2000) Distinct types of diffuse large B-Cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511.

Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I. Schummer, M. and Yakhini, Z. (2000) Tissue classification with gene expression profiles. *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology.*

Breiman, L., Firedman, J.H., Olshen, R.A., and Stone, C. (1984): Classification and Regression Trees. Wadsworth Inc. Monterey, California, U.S.A.

Clare, A. and King, R.D. (2002). How well do we understand the clusters found in microarray data? *In Silico Biology* **0046**.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of Royal Statistical Society, Series B,* **34**: 187–220.

Cui, X., Hwang, G., Qiu, J., Blades, N.J. and Churchill, G.A. (2005) Improved statistical tests for differential gene expression by shrinking variance components estimates. *Bioinformatics* **6**, 59–75.

Dave, S.S., Wright, G., Tan, B. et al. (2004). Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *The New England Journal of Medicine* **351** 2159–2169.

Dettling, M. and Buhlmann, P. (2003) Boosting for tumor classification with gene expression data. *Bioinformatics* **9**, 1061–1069.

Dudoit, S., Fridyland, J.F. and Speed, T.P. (2002) Comparison of discrimination methods for tumor classification based on microarray data. *JASA* **97**, 77–87.

Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Nat. Acad. Sci. USA **95**: 14863–14868.

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

Friedman, J.H. and Popescu, B.E. (2004). Gradient directed regularization for linear regression and classification. *Technical report, Department of Statistics, Stanford University. http* $: //www - stat.stanford.edu/ \sim jhf/PathSeeker.html.$

Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, van de Rijn M, Rosen GD, Perou CM, Whyte RI, Altman RB, Brown PO, Botstein D and Petersen I. (2001). Diversity of gene expression in adenocarcinoma of the lung. *PNAS* **98**, 13784–13789.

Golub, G. and Van Loan, C. (1996). *Matrix Computations*. Johns Hopkins Univ Press, Baltimore.

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

Gui, J. and Li, H. (2005b) Threshold gradient descent method for censored data regression with applications in pharmacogenomics. *Proceedings of Pacific Symposium on Biocomputing 2005.*

Hastie, T., Tibshirani, R. and Friedman, J (2001) *The Elements of Statistical Learning*. Springer-Verlag.

Heagerty, P.J., Lumley, T. and scPepe, M.S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker, *Biometrics* **56**, 337–344.

Huang, J., Ma, S. and Xie, H. (2005). Regularized estimation in the accelerated failure time model with high dimensional covariates. *Biometrics, In press.*

Johnson, R.A. and Wichern, D.W. (2002). *Applied Multivariate Statistical Analysis.* Prentice-Hall.

Ma, S. and Huang, J. (2005). Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics*, **21** 4356–4362.

Ma, S., Kosorok, M. R. and Fine, J. P. (2006). Additive risk models for survival data with high dimensional covariates. *Biometrics*, **62** 202–210.

Nguyen, D. and Rocke, D.M. (2002a). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18**: 39–50.

Pochet, N., De Smet, F., Suykens, J. and De Moor, B. (2004) Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. *Bioinformatics* **17**, 3185–3195.

Rosenwald, A. Wright, G., Wiestner, A., Chan, W. C., et al. (2003). The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell*, **3**, 185–197.

Spang, R. Blanchette, C., Zuzan, H., Marks, J., Nevins, J. and West, M. (2001) Prediction and uncertainty in the analysis of gene expression profiles. *Proceedings of the German Conference on Bioinformatics GCB 2001.*

Tavazoie, S., Hughes, J., Campbell, M., Cho, R. and Church, G. (1999) Systematic determination of genetic network architecture. *Nature Genetics*, **22** 281-285.

Wahba, G. (1990) *Spline models for observational data. CBMS-NSF Regional Conference Series in Applied Mathematics.* Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.

Wei, Z. and Li, H. (2006). Nonparametric pathway-based regression models for analysis of genomic data. *University of Pennsylvania Biostatistics Working Papers, Year 2006, Paper 6.*

West, M. Blanchette, C., Dressmna, H., Huang, E., Ishida., s., Spang, R., Zuzan, H., Olson, J., Marks, J. and Nevins, J. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *PNAS* **98**, 11562–11467.

Yeung, K.Y., Haynor, D. and Ruzzo, W. (2001) Validating clustering for gene expression data. *Bioinformatics* **17** 309–31.