Regularized ROC Estimation: With Applications to Disease Classification Using Microarray Data

Shuangge Ma and Jian Huang [1]

June 2005

The University of Iowa

Department of Statistics and Actuarial Science

Technical Report No. 345

Shuangge Ma is Postdoctoral Fellow, Department of Biostatistics, University of Washington–Seattle. (Email: shuangge@u.washington.edu) The work of Ma is partly supported by N01-HC-95159 from the National Heart, Lung, and Blood Institute. Jian Huang is Professor, Department of Statistics and Actuarial Science, and Program in Public Health Genetics, University of Iowa. (Email: jian@stat.uiowa.edu). The work of Huang is supported in part by the NIH grant HL72288.

Summary Microarrays capable of monitoring gene expression on a large scale are becoming a routing tool in biomedical research. An important application of microarrays is to discover genomic biomarkers, among tens of thousands of genes assayed, for disease classification and prediction. Thus there is a need for novel statistical methods that can efficiently use such high-throughput genomic data, are suitable for disease classification, and are able to select important biomarkers among thousands of genes. We propose a sigmoid approximation to a widely used accuracy measure based on the receiver operating characteristic (ROC) curve, the area under the ROC curve (AUC), as the objective function for these purposes. We show that the proposed method consistently estimate the AUC when the number of covariates is fixed and the sample size is large. The proposed sigmoid objective function makes it computationally feasible for variable selection in high-dimensional settings. We investigate gradient directed techniques, including the L_1 boosting based LASSO (least absolute shrinkage and selection operator) and the TGDR (threshold gradient descent regularization), that can identify the important genomic biomarkers from the data. The proposed methods yield parsimonious models with excellent classification performance measured by the AUC. We demonstrate the proposed methods with two cancer studies that use Affymetrix genechip expression data.

KEYWORDS: Area under the ROC curve; Biomarker; Gradient descent regularization; High-dimensional covariate; LASSO; Variable selection.

1. INTRODUCTION

Global gene expression profiling using microarrays has the potential to lead to a better understanding of the molecular features corresponding to different phenotypes (Alon et al. 1999; West et al. 2001; Spang et al. 2001). When microarrays are used for disease classification, each DNA sequence represented in microarrays can be considered as a biomarker. Of special interest is to identify biomarkers that can best separate populations with different phenotypes, for example people with and without certain type of cancer, and hence lead to a more precise treatment selection. Dudoit, Fridlyand and Speed (2002) provide a thorough review of currently available classification methodologies for genomic data. Classification of phenotypes using genomic data is challenging due to high dimensional covariates (usually greater than 1000) and relatively small number of observations (usually less than 200).

One-dimensional error measures such as overall misclassification error often used in other classification problems are rarely used in medicine (Pepe, Janes, Longton, Leisenring, and Newcomb 2004a). A unique feature of disease classification is that it is important to assess both false-positive and false-negative errors, since these two types of errors usually have different consequences for patients and hard to quantify. A common practice is to use receiver operating characteristic (ROC) curve to evaluate a classification procedure (Baker 2003; Pepe 2003). Pepe, Cai and Zhang (2004b) proposed using the area under curve (AUC) of a ROC curve as a summary measure of the effectiveness of a classification rule. It has been shown that classification based on the linear risk scores estimated with the ROC technique is optimal in the sense that no other classification rule can have better point-wise accuracy under a generalized linear model assumption. Another important advantage of the ROC approach is its adaptability to case control designs which are widely used in epidemiologic studies.

The method of Pepe et al. (2004b) is effective when the number of biomarkers is relatively

small. However, they did not address the problem of variable selection. It appears difficult to apply their method to situations where the number of predictors is greater than the sample size and variable selection is a must, such as studies using microarray data. Because it is often biologically reasonable to assume that only a small proportion of biomarkers are related to phenotypes, variable selection becomes an important issue along with establishing the classification rule. The AUC objective function of the empirical ROC curve in Pepe et al. (2004b) is a sum of indicator functions hence a discontinuous step function. Computationally, it is extremely difficult to optimize such a function in a high-dimensional space and do variable selection.

When the number of predictors is larger than the sample size, model reduction becomes essential for gaining stability and parsimony in establishing the classification function. Biased regression methods such as those described in Hastie, Tibshirani and Friedman (2001) have been employed for model reduction with genomic data, see for example, partial least squares (Nguyen and Rocke 2002; Li and Gui 2004), principal component regression (Ma, Kosorok and Fine 2005), among others. The basic idea is that by using low dimensional projections of the covariates as surrogates for the true covariates, one may achieve estimators with reduced mean squared errors. An alternative to using low dimensional projections is to use variable selection techniques to choose the "important" covariates. This may be accomplished by using penalization methods, for example, the least absolute shrinkage and selection operator (LASSO, Tibshirani 1996), the least angle regression (LARS, Efron et al. 2004), and the gradient directed regularization method (Friedman and Popescu 2004). See for example, Gui and Li (2004, 2005) for using such methods in analyzing censored survival data with microarray covaraites. Ghosh and Chinnaiyan (2004) proposed a hybrid method by first constructing a quantitative score for the binary disease status and then using this score in the linear discriminant analysis framework discussed by Su and Liu (1993). They used LASSO in

the construction of the quantitative score and biomarker selection. However, they did not use AUC as the objective function.

Because of the difficulty in using the empirical AUC as the objective function directly, we propose using the sigmoid function as an approximation to the AUC criterion (Yan et al. 2003). This approximation makes it computationally feasible to estimate the classification rule and carry out variable selection simultaneously using high dimensional genomic data. Using a continuous objective function to approximate a discontinuous one has proven to be effective in many statistical and machine learning methods, for example, in the support vector machine (SVM) method, the continuous hinge loss can be considered an approximation to the indicator error function (Hastie et al. 2001). Using the sigmoid AUC criterion, we consider two regularized methods for the ROC classification. The first method is the LASSO. Because efficient algorithms such as LARS (Efron et al. 2004) are not directly applicable to the sigmoid AUC function, we propose using a gradient directed L_1 boosting algorithm for the LASSO. As an alternative to the LASSO, we propose applying the threshold gradient descent regularization (TGDR) method (Friedman and Popescu 2004) to the same data setting.

The article is organized as follows. In section 2, we introduce the stochastic model and the ROC curve. The sigmoid approximation is discussed in section 3. Asymptotic properties of the sigmoid estimators are established in the same section assuming fixed dimensional covariates and large sample size. The LASSO and the TGDR algorithms are investigated in sections 4. In section 5, we apply the two regularized estimating approaches to two cancer studies that use Affymetrix genechip expression data. Possible extensions and concluding remarks are in section 6. Proofs of the theorems are provided in the Appendix.

2. STOCHASTIC MODEL AND ROC CURVES

2.1 Stochastic Model

Consider a biological study where the outcome Y is a Bernoulli random variable i.e., $Y \in \{0, 1\}$. Y may denotes presence or absence of cancer or certain type of tumor. Without loss of generality, we refer to $Y = 1$ as the diseased class and $Y = 0$ as the healthy class. Assume there are n subjects in the study. For the i^{th} subject, values of d covariates are measured: $\mathbb{X}_i = (X_{i,1}, \ldots, X_{i,d})$. For example, $\mathbb{X}_i$ may denote the gene expression profiles for a subject. We model the relationship between $\mathbb{X}$ and the phenotype Y with the generalized linear model $P(Y = 1|\mathbb{X}) = G(\beta'\mathbb{X})$, where G is an unknown increasing link function, β is the d-vector of unknown regression parameter and β' denotes the transpose of β.

2.2 ROC Curves

Let $\mathbb{X}^D$ denote the covariates of a diseased subject (i.e, $Y = 1$) and $\mathbb{X}^H$ the covariates of a healthy subject. To evaluate the performance of a classifier based on the linear risk scores $\beta'\mathbb{X}$, we employ the traditional measurements of classification accuracy that are used in medicine, namely the true and false positive rates (TPR and FPR). Also known as sensitivity and 1-specificity, respectively, TPR and FPR are defined as

$$\text{TPR(c)=P}(\beta'\mathbb{X} \geq c|Y = 1) \text{ and FPR(c)=P}(\beta'\mathbb{X} < c|Y = 0),$$

for any pre-specified cutoff c.

The TPR and FPR can be summarized by the ROC curve, which is a two-dimensional plot of $\{(FPR(c), TPR(c)) : -\infty < c < \infty\}$. The ROC curve demonstrates the balance between the true positive and false positive rates. Classification rules that have $(FPR(c), TPR(c))$ closer to $(0, 1)$ indicate perfect discriminators, while those with $(FPR(c), TPR(c))$ near the 45^o line cannot discriminate between the diseased and the healthy classes. The overall performance of a classifier can be measured by the AUC of a ROC plot. It has been proved that the upper limit of the AUC (and hence the best classifier) is equal to $P(\beta_0'\mathbb{X}^D \geq \beta_0'\mathbb{X}^H)$, where β_0 is the true parameter value (Baker 2003).

For the n subjects, denote $\mathbb{D}$ and $\mathbb{H}$ as the index sets for diseased and healthy subjects with sizes n_D and n_H, respectively. For any given β and its corresponding ROC generated with the linear risk scores $\beta'\mathbb{X}$, the empirical AUC is

$$AUC(\beta) = \frac{1}{n_D n_H} \sum_{i \in \mathbb{D}; j \in \mathbb{H}} I(\beta'\mathbb{X}_i > \beta'\mathbb{X}_j). \tag{1}$$

The ROC estimator is defined as the maximizer of $AUC(\beta)$ (Pepe et al. 2004b). Note the ROC estimator is a special case of the maximum rank correlation (MRC) estimator in Han (1987). If no further assumption is made on the link function, the ROC estimator is only identifiable up to a scale constant. That is, only relative effects, instead of absolute effects, can be estimated from the AUC objective function. Without loss of generality, we assume $|\beta_{(1)}| = 1$, where $\beta_{(1)}$ denotes the first component of β, i.e, the first biomarker is the "anchor biomarker". We suggest a simple way of determining the anchor biomarker in section 5. Another often used constraint to ensure identifiability is to set $\|\beta\|_2 = 1$.

3. THE SIGMOID MAXIMUM RANK CORRELATION ESTIMATOR

3.1 Estimation

Since the objective function in (1) is not continuous, maximization of (1) is difficult. One way to overcome this difficulty is to approximate the discontinuous function by a smooth function. One possibility proposed by Yan et al. (2003) it to use the sigmoid function $s(x) = 1/(1 + exp(-x))$ as an approximation to the indicator function $I(x) = 1(x >)$ in (1). For large $|x|$, $s(x)$ is a good approximation to $I(x)$. However, for x near 0, this approximation is poor. Thus this approximation will introduce systematic bias in the estimation of β. An effective way to reduce the bias is to introduce a sequence of strictly positive numbers σ_n satisfying $\lim_{n\to\infty} \sigma_n = 0$, and use $s_n(x) = s(x/\sigma_n)$ to approximate $I(x)$. Similar approach

was proposed by Horowitz (1992) in the context of maximum score estimator for the binary response model. Horowitz considered a general class of distribution-like kernel functions for approximation of I. Let K be a differentiable distribution function on the real line, so that it satisfies $\lim_{x\to-\infty} K(x) = 0$ and $\lim_{x\to\infty} K(x) = 1$. Then $K_n(x) = K(x/\sigma_n)$ can be used to approximate $I(x)$. Since the sigmoid function is also the logistic distribution function, s_n is a special case of K_n. Because of the good approximation property and simplicity of s_n, we use s_n in our estimation. We refer to the resulting estimate $\hat{\beta}$ using the sigmoid approximation as the sigmoid maximum rank correlation (SMRC) estimator, i.e.,

$$\widehat{\beta} = argmax \left\{ R_n(\beta) = \frac{1}{n_D n_H} \sum_{i\in\mathbb{D};j\in\mathbb{H}} s_n(\beta'(\mathbb{X}_i - \mathbb{X}_j)) \right\}. \tag{2}$$

As in the original AUC estimator, we set $|\widehat{\beta}_{(1)}| = 1$ for identifiability.

Since the scaled sigmoid function is an excellent approximation to the indicator function, intuitively the SMRC estimator should be close to the ROC estimator. In the standard case of large n and small d, we show that $\widehat{\beta}$ is consistent and asymptotically normal under appropriate conditions. Consequently, the estimated sigmoid AUC objective function is a consistent estimator of the AUC. Although these results are not applicable to the case of small n and large d, it does provide justification for using s_n as a reasonable smooth approximation to the original AUC objective function.

3.2 Asymptotic Properties for Fixed d and Large n

Consider the statistical model defined in section 2.1. Let $(Y_1, \mathbb{X}_1), \ldots, (Y_n, \mathbb{X}_n)$ be a random sample of $(Y, \mathbb{X})$. Without loss of generality, we assume $\beta_{(1)} = 1$. For notational convenience, let $\theta = (\beta_{(2)}, \ldots, \beta_{(d)})'$ and $\widehat{\theta}_n = (\widehat{\beta}_{(2)}, \ldots, \widehat{\beta}_{(p)})'$, where $\widehat{\beta}$ is the SMRC estimate defined in (2). To indicate that θ is the actual parameter, let $\beta(\theta) = (1, \theta')'$. Then we can write

$$H_n(\theta) = \frac{1}{n(n-1)} \sum_{i \neq j} I(Y_i > Y_j) I(\beta(\theta)' \mathbb{X}_i - \beta(\theta)' \mathbb{X}_j > 0), \text{ and}$$

$$R_n(\theta) = \frac{1}{n(n-1)} \sum_{i \neq j} I(Y_i > Y_j) s_n(\beta(\theta)' \mathbb{X}_i - \beta(\theta)' \mathbb{X}_j).$$

In this formulation, the response Y can also be continuous or ordinal variables. With minor modifications in the assumptions, the results below are applicable to this more general situation. The following assumptions are used in the consistency proof.

(A1). The true parameter value θ_0 is an interior point of Θ, where Θ is a compact subset of $\mathbb{R}^{d-1}$.

(A2) Let S_x denote the support of the covariate vector $\mathbb{X}$. (i) S_x is not contained in any proper linear subspace of $\mathbb{R}^d$ and, (ii) the first component of $\mathbb{X}$ has an everywhere positive density, conditional on the other components.

The assumptions (A1) and (A2) were used by Han (1987) to prove the consistency of the MRC estimator. Specifically, (A1) was used for establishing uniform convergence of H_n to its population version. (A2) is assumed to ensure identifiability of θ.

Theorem 1. (Consistency) Suppose assumptions (A1) and (A2) hold and $\sigma_n \to 0$ as $n \to \infty$. Then $\widehat{\beta} \to_P \beta_0 \quad \text{as } n \to \infty$.

The proof is postponed to the Appendix. We now describe notations needed for stating the asymptotic normality result of the SMRC estimator. We use notations similar to those of Sherman (1993), who established the asymptotic normality of the MRC estimator. Let χ denote the last $d-1$ components of the covariate vector $\mathbb{X}$. Let $g^0(\cdot|r)$ denote the conditional density given $\chi = r$, and g^0 the marginal density of $\beta_0' \mathbb{X}$. Write $\chi_0 = E(\chi|\beta_0' \mathbb{X})$. Denote $S(y,t) = E[I(y > Y) - I(y < Y)|\beta_0' \mathbb{X} = t]$, and $S_2(y,t) = (\partial/\partial t) S(y,t)$. We have

$$E[S(Y,t)|\beta_0'\mathbb{X}=t]=0. \tag{3}$$

This equation is the key to the asymptotic unbiasedness of the SMRC. It holds because for each t, $S(Y,t)$ has a symmetric distribution conditional on $\beta_0'\mathbb{X}=t$, see Sherman (1993). Let

$$A=E[(\chi-\chi_0)(\chi-\chi_0)'S^2(Y,\beta_0'\mathbb{X})(g^0(\beta_0'\mathbb{X}))^2], \tag{4}$$

$$2B=E[(\chi-\chi_0)(\chi-\chi_0)'S_2(Y,\beta_0'\mathbb{X})g^0(\beta_0'\mathbb{X})]. \tag{5}$$

We note that A and B are exactly the same as the Δ and V in Theorem 4 of Sherman (1993), respectively. In addition to assumptions (A1) and (A2), we need the following assumptions for asymptotic normality.

(A3) The scaling factor $\sigma_n=o(n^{-1/2})$.

(A4) (i) The matrices A and B exist and B is nonsingular. (ii) Let $q(y,t,r)=S(y,t)g^0(t|r)$. The partial derivative of q with respect to t exists and is bounded over the support of $(Y,\mathbb{X})$.

This assumption (A3) is to ensure that the bias introduced by using the sigmoid approximation is asymptotically negligible. (A4) is related to assumption (A4) of Sherman (1993). Here, (A4)(i) is made to ensure that the asymptotic variance matrix exists. (A4)(ii) is a regularity condition. Because we smooth the indicator function, existence and boundedness of only the first derivative of q are assumed. For the MRC estimator, second derivatives of a related function are needed in the proof.

Theorem 2. (Asymptotic Normality) Suppose that assumptions (A1) - (A4) hold. Then $\sqrt{n}(\widehat{\theta} - \theta_0) \to_D N(0, B^{-1}AB^{-1})$ as $n \to \infty$.

This result shows that the SMRC estimator has the same asymptotic distribution as the MRC estimator. Although the result is stated for the sigmoid function, it continues to hold when s is replaced by any symmetric distribution function with a continuous second derivative, see the proof in the Appendix. We can estimate $B^{-1}AB^{-1}$ consistently using the derivatives of $R_n(\theta)$. As can be seen from the proof of Theorem 2 given in the Appendix, a consistent estimator is $B_n^{-1}A_nB_n^{-1}$, where

$$A_n = \frac{1}{n(n-1)} \sum_{i \neq j} I(Y_i > Y_j) \left[\frac{\partial}{\partial \theta} s_n(\beta(\widehat{\theta})' \mathbb{X}_i - \beta(\widehat{\theta})' \mathbb{X}_j) \right]^2, \quad B_n = -\frac{\partial^2}{\partial \theta^2} R_n(\widehat{\theta}).$$

We note that for the MRC estimator, construction of a consistent variance estimator requires use of finite differences to approximate certain derivatives due to the fact that H_n is not differentiable, see Sherman (1993).

4. REGULARIZED ROC CLASSIFICATION

We now describe the LASSO and the TGDR algorithms for estimation and variable selection using the smoothed AUC objective function. V-fold cross validation is used for tuning parameter selection in both the LASSO and the TGDR. We postpone the discussion of determining the anchor biomarker and the sign of its corresponding coefficient to section 5. For simplicity of notations and without loss of generality, we assume the first biomarker is the anchor and $\beta_{(1)} = 1$ and we still use β to denote the remaining coefficients $(\beta_{(2)}, \ldots, \beta_{(d)})'$.

4.1 The LASSO Estimate

The LASSO estimate is defined as the maximizer of (2) under the L_1 constraint $|\hat{\beta}|_{L_1} \leq u$, for a data-dependent tuning parameter u, which indirectly determines how many estimated

coefficients are zero. The L_1 constraint is equivalent to adding a L_1 penalty to the objective function and ignoring the constraint (Tibshirani 1996). Since the L_1 penalty is not differentiable, usual derivative-based minimization techniques, for example the Newton-Raphson method, cannot be used to obtain the estimate in (2) under the L_1 constraint. In most previous studies, the minimization relies on the quadratic programming (QP) or general non-linear program, which are known to be computationally intensive and cannot be applied directly to the settings when the sample size is much smaller than the number of predictors. The LARS algorithm (Efron et al. 2004) provides an efficient way for computing the LASSO estimator in least squares regression, but is not directly applicable in the present setting.

Recent study by Kim and Kim (2004), which relates the minimization step for the LASSO estimate to the L_1 boosting algorithm, provides a computationally more feasible solution. For the LASSO estimate defined in (2) under the L_1 constraint with a fixed u, this algorithm can be implemented in the following steps:

1. Initialization $\beta = (0, \ldots, 0)$ and $m = 0$.

2. With the current estimate of β, compute $g(\beta)$, the negative derivative of $R_n(\beta)$ with respect to β. Denote the p^{th} component of $g(\beta)$ as $g_{(p)}(\beta)$.

3. Find p^* that minimizes $min_p(g_{(p)}(\beta), -g_{(p)}(\beta))$. If $g_{(p^*)}(\beta) = 0$, then stop the iteration.

4. Otherwise denote $\gamma = -sign(g_{(p^*)}(\beta))$. Find $\hat{\alpha} \in [0, 1]$ that minimizes $R_n((1-\alpha)\beta + \alpha \times u \times \gamma\eta_{p^*})$, where η_{p^*} is a length d vector that has the p^*th element equals to 1 and the rest components equal to 0.

5. For the p^{th} component of β: $\beta_{(p)} = (1-\hat{\alpha})\beta_{(p)}$ for $p \neq p^*$, and $\beta_{(p^*)} = (1-\hat{\alpha})\beta_{(p^*)} + \gamma u\hat{\alpha}$. Let $m = m + 1$.

6. Repeat steps 2–5 until convergence or a fixed number of iterations N has been reached.

The β at convergence is the LASSO estimate (Kim and Kim 2004). We conclude convergence if the absolute value of $g_{(p^*)}(\beta)$ computed in step 3 is less than a pre-defined criteria, and/or if $R_n(\beta)$ is larger than a pre-defined threshold. Note we exclude the anchor covariate from the above iterative estimation.

Comparing with the QP, the L_1 boosting algorithm only involves evaluations of simple functions. Data analysis experiences show the computational burden for the L_1 boosting is minimal. One attractive feature of the L_1 boosting algorithm is that the convergence rate is independent of the dimension of input, which is particularly valuable for high dimensional genomic data (Kim and Kim 2004). In addition, it has been known that for boosting methods, over-fitting usually does not pose a serious problem (Friedman, Hastie and Tibshirani 2001). So the overall iteration N can be taken to be a large number to ensure convergence.

4.2 The TGDR Estimate

The TGDR approach first establishes a parameter path in the high dimensional coefficients space using the gradient descent method, and then identify the best model along the parameter path with certain cross validation techniques (Friedman and Popescu 2004).

For any fixed threshold $0 \leq \tau \leq 1$, the TGDR algorithm constructs the parameter path as follows. Suppose that we search along a discrete grid $\nu_k = k\Delta_\nu, k = 0, 1, \ldots$ in the path, where Δ_ν is positive and infinitesimal. For any $k \geq 0$, let β_k denote the parameter estimate along the parameter path corresponding to the index ν_k. Denote the negative gradient of $R_n(\beta)$ evaluated at β_k as $g(\nu_k)$. Moreover, we define $f(\nu_k)$, whose j^{th} component is $f_j(\nu_k) = I(|g_j(\nu)| \geq \tau \cdot \max |g_k(\nu)|)$. Denote the component-wise product of $f(\nu_k)$ and $g(\nu_k)$ as $h(\nu_k)$. Let the initial value $\beta_0 = 0$. In the threshold gradient descent search,

$$\beta_k = \beta_{k-1} + \Delta_\nu h(\nu_{k-1}),\ k = 1, 2, \ldots. \tag{6}$$

When $p < n$, the TGDR can generate a full path connecting the initial value with the ordinary M-estimate for any fixed τ. For the data settings discussed in section 2 and the sigmoid function, the TGDR estimate may diverge or converge to a local maxima/minima as $k \to \infty$.

Stable estimates are expected with non-zero τ, since covariates (biomarkers) with small gradients are excluded from the model. If a covariate (or a group of covariates) is known to be related to the phenotypes *a priori*, we can exclude it from the threshold step, i.e., f for this covariate can be set to 1. Similar strategy can be employed with the LASSO estimates. See Kim and Kim (2004) for relevant discussions.

4.3 Tuning Parameters Selection

We propose using the following V-fold cross validation (Wahba 1990) to determine the tuning parameters: u for the LASSO and k for the TGDR with fixed τ. For a pre-defined integer V, partition the data randomly into V non-overlapping subsets of equal sizes. Choose u for the LASSO (k for the TGDR) to maximize the cross-validated objective function

$$CV\ score = \sum_{v=1}^{V} \left[R_n(\hat{\beta}^{(-v)}) - R_n^{(-v)}(\hat{\beta}^{(-v)}) \right], \tag{7}$$

where $\hat{\beta}^{(-v)}$ is the LASSO (TGDR) estimate of β based on the data without the v^{th} subset for a fixed u (k) and $R_n^{(-v)}$ is the function R_n defined in (2) evaluated without the v^{th} subset. Since the CV score function in (7) contains differences of covariates from the two phenotypes, the usual leave-out-one cross validation is not applicable here.

For the TGDR, the optimal threshold τ can be chosen with similar cross validation techniques. Since the performance of the TGDR estimates for different threshold values is of interest, we employ cross validation with respect to k only in this article. We discuss the

effect of different τ with empirical studies in section 5. Related discussions can also be found in Gui and Li (2005).

4.4 Characteristics of LASSO and TGDR

The LASSO and the TGDR are both gradient directed iterative algorithms. One difference is that the LASSO estimate increases in the direction of one covariate, while the TGDR estimate may increase in the direction of multiple covariates at each iteration. The effectiveness of the LASSO technique for model reduction with genomic data has been demonstrated by Ghosh and Chinnaiyan (2004). The TGDR has been successfully employed for right censored survival data with high dimensional covariates and the Cox model in Gui and Li (2005).

In a linear regression model, Friedman and Popescu (2004) show that the TGDR can provide a path connecting the solutions roughly corresponding to the PLS/RR (ridge regression) and the solutions roughly corresponding to the LASSO by varying the threshold values. Moderate to large threshold values create paths that involve more diverse absolute coefficient values than the PLS/RR solutions but less than the LASSO solutions. When two covariates with the same norm are strongly correlated, their corresponding gradients are close. So the TGDR yields similar estimates for strongly correlated covariates. This property is not shared by the LASSO. One drawback of the TGDR is that the TGDR estimates may be less stable compared with the LASSO estimates, since the TGDR tends to identify more covariate effects. To out best knowledge, there is no study investigating the similarity and distinction between and the LASSO and the TGDR. More studies are needed to draw definitive conclusions regarding the relative performance of the two proposed approaches in different situations.

5. EXAMPLES

5.1 Datasets

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 chip. A selection of 2000 genes with 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/.*

Estrogen Data. These datasets were first presented by West et al. (2001) and Spang et al. (2001). Their common expression matrix monitors 7129 genes in 49 breast tumor samples. The data are available at *http://mgm.duke.edu/genome/dna_micro/work/* and were obtained by applying the Affymetrix gene chip technology. The response describes the lymph nodal (LN) status, which is an indicator for the metastatic spread of the tumor, an important risk factor for disease outcome. 25 samples are positive (LN+) and 24 samples are negative (LN-).

5.2 Data Pre-processing

Data pre-processing is usually needed when using genomic data to exclude biomarkers with severe missingness and/or little variations. The proposed approaches do not incorporate missing data automatically, so pre-processing of missing data as in Dudoit, Fridlyand and Speed (2002) are needed when they are present. One widely accepted approach is to fill in missing values with corresponding sample medians or means. The Colon data has been partially pre-processed. For the Estrogen data, 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, Fridlyan and Speed, 2002). A base 2 logarithmic transformation is then applied. Compared with the log 10 transformation in Dudoit et al. (2002), the log 2 transformation can be easily translated into fold changes (with 2-fold change as the basic unit).

We identify the anchor biomarker as follows. 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 biomarkers are ranked based on the t-statistics computed with the adjusted standard errors. Compared with the adjusted t-test, standard t-test usually picks up a lot genes with very small fold change (although very significant in terms of p-values), mostly due to very small standard errors. This issue has been discussed in Cui et al. (2005), where shrinkage methods in marginal tests were investigated. The biomarker with the largest absolute values of the adjusted t-statistic is chosen as the anchor biomarker. For the anchor biomarker, if the sample mean of the diseased class is larger, then $\beta_{(1)} = 1$, otherwise $\beta_{(1)} = -1$. Although there is no guarantee that biomarkers marginally strongly associated with the phenotypes will have nonzero coefficients in the joint model, the probability of that is high.

After the anchor biomarker is chosen, we determine the scale parameter σ_n as follows. As discussed in section 3.2, when $\sigma_n = o(n^{-1/2})$, the SMRC estimators with $s_n(x)$ as an approximation to the indicator function are $\sqrt{n}$-consistent and asymptotically normal distributed. Computationally, when $|x/\sigma_n|$ is greater than 5, the sigmoid approximation is usually accurate enough in most practical situations (Gammerman 1996). Denote the sample means of gene expressions of the diseased and normal groups as $\bar{\mu}_D$ and $\bar{\mu}_H$ for the anchor biomarker, and denote the sample standard errors of the anchor biomarker in the two groups as $\bar{\sigma}_D$ and $\bar{\sigma}_H$, respectively. If we assume that the biomarkers are independent and joint normally distributed, then we can choose σ_n to satisfy

$$(8) \qquad P\left(\frac{x_{1D} - x_{1H}}{\sigma_n} > 5\right) > 0.95, \text{ where } x_{1D} \sim N(\bar{\mu}_D, \bar{\sigma}_D^2) \text{ and } x_{1H} \sim N(\bar{\mu}_H, \bar{\sigma}_H^2).$$

Even if the normality and independence assumptions are not satisfied, (8) still provides a useful guideline for choosing σ_n. When there are many covariates, we can still use the same value used in the case of a single covariate as determined by (8). The rationale for this is as follows.

If we only have the anchor biomarker, then (8) guarantees that a majority of the terms x/σ_n have absolute values greater than 5. Now consider multi-biomarker cases. If we neglect the correlation structures among biomarkers, then for biomarker j, $P(\beta_{(j)}(x_D - x_H) > 0) > 0.5$. So when we add more covariates, the absolute values tend to become bigger. When we take the correlation structures into consideration, the above argument is not necessarily true for all biomarkers, but will still tend to be true on average. Empirical studies support the validity of this approach.

The 500 biomarkers with the largest absolute values of the adjusted t-statistics are used for analysis. Note there is no theoretical or computational limitation on how many covariates can be used in the regularized estimation. We choose to work with 500 biomarkers only to gain further stability. The biomarkers are then standardized to have zero means and unit variances.

5.3 Empirical Results

We apply the two regularized estimating approaches to the Colon and the Estrogen datasets. Five-fold cross validation is applied for tuning parameter selections.

For the LASSO approach, the cross validated tuning parameters are $u = 22.5$ for the Colon data and $u = 13.0$ for the Estrogen data, respectively. With the cross validated tuning parameters, the LASSO estimates have AUCs 0.954 (Colon data) and 0.958 (Estrogen data), respectively. For the TGDR, we show in Table 1 the cross validated k, number of nonzero coefficients and corresponding estimate of AUC for each fixed τ. It can be seen that generally k increases and the number of nonzero coefficients decrease as τ increase. However, the change of AUC is not significant. Our extensive empirical studies show that generally more iterations are needed for large threshold values of τ, which will also lead to more parsimonious models. When the values of AUC are comparable, parsimonious statistical models are preferred. So

with the TGDR, we choose $\tau = 1.0$ for both datasets.

In tables 2 and 3, we list the genes with non-zero coefficients and corresponding estimates under the LASSO and the TGDR. Since the gene expressions have been normalized to have unit variances, estimates of covariates effects from each method are directly comparable. Larger absolute values of coefficients indicate stronger influences. Detailed gene descriptions, the correlation and the clustering structures for genes with non-zero coefficients are available upon request from the authors. We can see that the TGDR identifies more genes than the LASSO. Most of the genes identified by the LASSO are also identified by the TGDR. For genes identified by both approaches, the estimates have the same signs, which indicates the same biological conclusions. Both approaches have large AUC values and hence satisfactory classification performances.

6. CONCLUDING REMARKS

Microarrays that can assay tens of thousands of genes are becoming a routine tool in studies of many diseases, and are providing a large number of potential biomarkers for disease classification and prediction. It is important to develop computationally efficient models and methods that are appropriate for disease classification and biomarker selection. The AUC under the ROC curve is extensively used for evaluating classification performance in biomedical studies, therefore, it is desirable to construct classification rules that optimize AUC. In this article, we propose using the smooth sigmoid objective function as an approximation to the discontinuous AUC, so that it is computationally feasible for high-dimensional genomic data. Two gradient directed regularization methods with cross validated tuning parameters are investigated. Applications of the proposed methods to two studies using Affymetrix gene chip data indicate that both the LASSO and the TGDR can yield parsimonious models with satisfactory classification performance as measured by AUC.

We have provided theoretical results for the SMRC estimates under the assumption of fixed d and large n. It is shown that if the scale parameter in the sigmoid function is properly chosen, the SMRC estimates have the same efficiency as the MRC estimates. The asymptotic results provide partial justification for using the sigmoid estimates when $d >> n$. Theoretical properties of the SMRC estimator under the when $d >> n$ are of interest for future study.

The AUC/ROC approach considered in this paper only deals with two-class problems. Since the AUC objective function can be considered a special case of the maximum rank correlation method of Han (1987), which is applicable to a more general response variable including ordinal polytomous outcomes and continuous responses, it can be generalized to multi-class classification problems. Suppose y is such a response variable, then the maximum rank correlation estimator is defined by

$$\hat{\beta} = argmax \left\{ M(\beta) = \sum_{i,j} I(y_i \geq y_j) I(\beta' \mathbb{X}_i \geq \beta' \mathbb{X}_j) \right\}.$$

The first indicator terms are not parameter-dependent. So we can approximate the second indicator terms with the sigmoid function and apply the proposed LASSO and the TGDR with minor modifications.

The sigmoid function has been extensively used in machine learning and neural network studies (Gammerman, 1996). Several issues related to numerical optimization remain unsolved. Gammerman (1996) noted that the empirical sigmoid objective function may have multiple local minima in neural network studies. Several ad hoc solutions have been proposed. For simple gradient directed approaches, the global minimum can be detected by varying starting values for gradient search. It appears that similar simple solutions are also applicable in the present setting. Although our own experiences show that this usually does not pose a serious problem, it is worth further investigation.

APPENDIX

We now prove Theorems 1 and 2. We note that the only special property of the sigmoid function we use in the proof of Theorem 2 is its symmetry: $s(x) + s(-x) = 1$ and that it is smooth and has a continuous second derivative. Therefore, the proofs below are valid when we use any scaled symmetric distribution function with a continuous second derivative as an approximation to the indicator function $1(x > 0)$.

Proof of Theorem 1. Because the maximum rank correlation estimator is consistent under (A1) and (A2) (Han 1987), it suffices to show that

$$\sup_{\theta \in \Theta} |H_n(\theta) - R_n(\theta)| \to_P 0.$$

For any $\delta > 0$, we have

$$\begin{aligned}
&|H_n(\theta) - R_n(\theta)| \\
&\leq \frac{1}{n(n-1)} \sum_{i \neq j} I(Y_i > Y_j) \, |I(\beta'(\mathbb{X}_i - \mathbb{X}_j) > 0) - s_n(\beta'\mathbb{X}_i - \beta'\mathbb{X}_j)| \, I(|\beta'(\mathbb{X}_i - \mathbb{X}_j)| \geq \delta) \\
&\quad + \frac{1}{n(n-1)} \sum_{i \neq j} I(Y_i > Y_j) \, |I(\beta'(\mathbb{X}_i - \mathbb{X}_j) > 0) - s_n(\beta'(\mathbb{X}_i - \mathbb{X}_j))| \, I(|\beta'(\mathbb{X}_i - \mathbb{X}_j)| < \delta) \\
&\equiv T_{n1} + T_{n2}.
\end{aligned}$$

On the set $\{|x| > \delta\}$, we have

$$|s_n(x) - I(x > 0)| \leq \exp(-|x|/\sigma_n) < \exp(-\delta/\sigma_n).$$

Thus when $\sigma_n \to 0$, $s_n(x) \to I(x > 0)$ uniformly on the set $\{|x| > \delta\}$. Therefore, T_{n1} converges

to 0 uniformly over Θ. The second term

$$T_{2n} \leq \frac{1}{n(n-1)} \sum_{i \neq j} I(|\beta'(\mathbb{X}_i - \mathbb{X}_j)| < \delta).$$

Since the class of indicator functions $\{I(\beta'(\mathbb{X}_i - \mathbb{X}_j)| < \delta) : \theta \in \Theta\}$ is manageable, by uniform convergence of U-processes (Theorem 7, Nolan and Pollard 1987), the right-hand side converges almost surely to $P(|\beta'(\mathbb{X}_i - \mathbb{X}_j)| < \delta)$ over Θ. However, under assumption (A2), it can be proved in a similar way as in Lemma 4 of Horowitz (1992), $P(|\beta'(\mathbb{X}_i - \mathbb{X}_j)| < \delta)$ converges to 0 uniformly over Θ by. This completes the proof of consistency.

We need the following lemma in the proof of Theorem 2. For simplicity, we write $\beta = \beta(\theta)$.

Lemma 1. Let $Z = (Y, \mathbb{X})$, $z = (y, x)$, $z_1 = (y_1, x_1)$ and $z_2 = (y_2, x_2)$ are random realizations of Z. Denote

$$\kappa_n(z, \theta) = E[I(y > Y)s_n(\beta'x - \beta'\mathbb{X})] + E[I(Y > y)s_n(\beta'\mathbb{X} - \beta'x)].$$

Let $q(y, t|r) = S(y, t)g_0(t|r)$ and $q_2 = (\partial/\partial t)q$. Then under assumption (A4)(ii),

$$\begin{aligned} \frac{\partial \kappa_n(Z, \theta_0)}{\partial \theta} &= (\chi - \chi_0)S(Y, \beta_0'\mathbb{X})g^0(\beta_0'\mathbb{X}) + O(\sigma_n), \\ \frac{\partial^2 \kappa_n(Z, \theta_0)}{\partial \theta^2} &= \int \left[-\int \frac{1}{\sigma_n} s''(y) q(Y, \sigma_n y + \beta_0'\mathbb{X}|r) dy \right] (r - \chi)(r - \chi)' dG_{-1}(r). \end{aligned}$$

Proof of Lemma 1. Since $s_n(x) + s_n(-x) = 1$, we have

$$\kappa_n(x, y, \theta) = \int [1 - s_n(\beta'x - \beta'\mathbb{X})]S(Y, \beta_0'\mathbb{X})dG(x) + \int \rho(Y, \beta_0'x)dG(x),$$

where $\rho(y,t) = E[\{y < Y\}|\beta_0'\mathbb{X} = t]$. This equation is similar to (17) of Sherman (1993). Note that the second term on the right-hand side does not depend on β. Let x_{-1} denote the last $d-1$ elements of x. Recall $\chi = \mathbb{X}_{-1}$. Let G_{-1} denote the distribution of χ. We have

$$\begin{aligned}\frac{\partial}{\partial\theta}\kappa_n(Z,\beta_0) &= \int \frac{1}{\sigma_n}s'((\beta_0'x - \beta_0'\mathbb{X})/\sigma_n)(\chi - x_{-1})S(Y,\beta_0'x)dQ(x) \\ &= \int\left[\int \frac{1}{\sigma_n}s'((t-\beta_0'\mathbb{X})/\sigma_n)s(Y,t)g^0(t|r)dt\right](\chi - r)dG_{-1}(r).\end{aligned}$$

By assumption (A4)(ii), the integral inside the bracket equals

$$\begin{aligned}\int \frac{1}{\sigma_n}s'((t-\beta_0'\mathbb{X})/\sigma_n)S(Y,t)g^0(t|r)dt &= \int s'(y)S(Y,\sigma_n y + \beta_0'\mathbb{X})g^0(\sigma_n y + \beta_0'\mathbb{X}|r)dy \\ &= \int_{-\infty}^{\infty} s'(y)dy\ S(Y,\beta_0'\mathbb{X})g^0(\beta_0'\mathbb{X}|r) + O(\sigma_n) \\ &= S(Y,\beta_0'\mathbb{X})g^0(\beta_0'\mathbb{X}|r) + O(\sigma_n),\end{aligned}$$

where in the last equality, we use $\int_{-\infty}^{\infty} s'(y)dy = 1$. So the result follows.

The second derivative

$$\begin{aligned}\frac{\partial}{\partial\theta^2}\kappa_n(Z,\beta_0) &= -\int s_n''(\beta_0'x - \beta_0'\mathbb{X})(x_{-1}-\chi)(x_{-1}-\chi)'S(Y,\beta_0'x)dG(x) \\ &= \int\left[-\int \frac{1}{\sigma_n^2}s''((t-\beta_0'\mathbb{X})/\sigma_n)q(Y,t|r)dt\right](r-\chi)(r-\chi)'dG_{-1}(r) \\ &= \int\left[-\int \frac{1}{\sigma_n}s''(y)q(Y,\sigma_n y + \beta_0'\mathbb{X}|r)dy\right](r-\chi)(r-\chi)'dG_{-1}(r).\end{aligned}$$

This proves Lemma 1.

Proof of Theorem 2. The proof follows the method of Sherman (1993). Let

$$f_n(z_1,z_2,\theta) = I(y_1 > y_2)\left[s_n(\beta'(x_1-x_2)) - s_n(\beta_0'(x_1-x_2))\right].$$

and $\Gamma_n(\theta) = R_n(\theta) - R_n(\theta_0)$. We can write

$$\Gamma_n(\theta) = \Gamma_{n0}(\theta) + P_n g_n(\cdot, \theta) + U_n h_n(\cdot, \cdot, \theta),$$

where $\Gamma_{n0}(\theta) = E f_n(Z_1, Z_2, \theta)$,

$$g_n(z, \theta) = P f_n(z, \cdot, \theta) + P f_n(\cdot, z, \theta) - 2\Gamma_{n0}(\theta) = \kappa_n(z, \theta) - \kappa_n(z, \theta_0) - 2\Gamma_{n0}(\theta),$$

and

$$h_n(z_1, z_2, \theta) = f_n(z_1, z_2, \theta) - P f_n(z_1, \cdot, \theta) - P f_n(\cdot, z_2, \theta) + \Gamma_{n0}(\theta).$$

Let $w(Z, \theta) = S(Y, \beta_0' \mathbb{X}) g^0(\beta_0' \mathbb{X})(\chi - \chi_0)$. We show that

$$\Gamma_{n0}(\theta) = \frac{1}{2}(\theta - \theta_0)' V (\theta - \theta_0) + O(\sigma_n |\theta - \theta_0|) + o(|\theta - \theta_0|^2), \tag{9}$$

$$P_n g(\cdot, \theta) = n^{-1/2} (\theta - \theta_0)' W_n + O(\sigma_n |\theta - \theta_0|) + o(|\theta - \theta_0|^2), \tag{10}$$

where $W_n = \sqrt{n} P_n w(\cdot, \theta_0)$, and

$$U_n h_n(\cdot, \cdot, \theta) = o_p(1) \tag{11}$$

uniformly in an $o_p(1)$ neighborhood of θ_0.

To prove (9), we note that $2\Gamma_{n0}(\theta) = E[\kappa_n(Z, \theta) - \kappa_n(Z, \theta_0)]$. By Lemma 1, equation (3)

and

$$q_2(y,t|r) = S_2(y,t)g^0(t|r) + S(y,t)\frac{\partial}{\partial t}g^0(t|r),$$

it follows that, using (A4)(ii),

$$\begin{aligned}
&E\frac{\partial^2}{\partial\theta^2}\kappa_n(Z,\theta_0) \\
&= -E\left\{\int\left[\int_{-\infty}^{\infty} s''(y)(1/\sigma_n)[q(Y,\sigma_n y+\beta_0'\mathbb{X}|r) - q(Y,\beta_0'\mathbb{X}|r)dy\right](r-\chi)(r-\chi)'dG_{-1}(r)\right\} \\
&= -E\left[\int\int_{-\infty}^{\infty} s''(y)ydy\int q_2(Y,\beta_0'\mathbb{X}|r))(r-\chi)(r-\chi)'dG_{-1}(r)\right] + O(\sigma_n) \\
&= E[(\chi-\chi_0)(\chi-\chi_0)'S_2(Y,\beta_0'\mathbb{X})g^0(\beta_0'\mathbb{X})] + O(\sigma_n).
\end{aligned}$$

So (9) holds. Here in the last equality we used $\int_{-\infty}^{\infty} s''(y)ydy = -1$ by integration by parts.

To prove (10), let $v(Z,\theta) = g^0(\beta_0'\mathbb{X})S(Y,\beta_0'\mathbb{X})(\chi-\chi_0)(\chi-\chi_0)'$. Since $g_n(z,\theta) = \kappa_n(z,\theta) - \kappa_n(z,\theta_0) - 2\Gamma_{n0}(\theta)$, by Lemma 1, we have

$$\begin{aligned}
P_n g(\cdot,\theta) &= n^{-1/2}(\theta-\theta_0)'W_n + \frac{1}{2}(\theta-\theta_0)'V_n(\theta-\theta_0) + O(\sigma_n|\theta-\theta_0|) + o_p(|\theta-\theta_0|^2) \\
V_n &= P_n v(\cdot,\theta_0) - 2V.
\end{aligned}$$

By a weak law of large numbers, $V_n = o_p(1)$. (10) follows.

To prove (11), consider the function

$$f(z_1,z_2,\theta,\sigma) = I(y_1 > y_2)\left[s((\beta' x_1 - \beta' x_2/\sigma) - s((\beta_0' x_1 - \beta_0' x_2)/\sigma)\right].$$

Then $f_n(z_1,z_2,\theta) = f(z_1,z_2,\theta,\sigma_n)$. Since $\sigma_n \to 0$, it suffices to show that $U_n f(\cdot,\cdot,\theta,\sigma) = o_p(n^{-1})$ uniformly over $o_p(1)$ neighborhoods of $(\theta_0,0)$. First, by assumption A2(ii) and using

the dominated convergence,

$$Eh^2(Z_1, Z_2, \theta, \sigma) \to 0 \ \text{ as } \ (\theta, \sigma) \to (\theta_0, 0).$$

Let $\mathcal{S}$ denote the class of functions

$$\mathcal{S} = \{I(y_1 > y_2)[s((\beta' x_1 - \beta' x_2)/\sigma) - s((\beta_0' x_1 - \beta_0' x_2)/\sigma)] : \ \theta \in \Theta, \ \sigma \in (0, 1]\}.$$

Then $\mathcal{S}$ is Euclidean by Lemma 22 (ii) of Nolan and Pollard (1987) and it is bounded by 2. Therefore, (11) follows from Theorem 3 of Sherman (1993) or Corollary 8 of Sherman (1992).

Finally, based on (9), (10), (11), and (A3) $\sigma_n = o(n^{1/2})$, by Theorem 1 of Sherman (1993), $\widehat{\theta} - \theta_0 = O_p(n^{-1/2})$. Asymptotic normality now follows from Theorem 2 of Sherman (1993). See also Chapter 3.2, Theorem 3.2.16 of Van der Vaart and Wellner (1996).

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.

Baker, S. G. (2003) The Central Role of Receiver Operating Characteristic (ROC) Curves in Evaluating Tests for the Early Detection of Cancer. *Journal of the National Cancer Institute*, **95**, 511515.

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.

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

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression. *Annals of Statistics* **32**, 407–499.

Friedman, J., Hastie, T. and Tibshirani, R. (2000) Additive logistic regression: a statistical view of boosting. *Annals of Statistics* **28**, 337–374.

Friedman, J.H. and Popescu, B.E. (2004) Gradient directed regularization for linear regression and classification. *Technical report, Department of Statistics, Stanford University.*

Gammerman, A. (1996) *Computational Learning and Probabilistic Reasoning.* Wiley, New York.

Ghosh, D. and Chinnaiyan, A.M. (2004) Classification and selection of biomarkers in genomic data using LASSO. *Journal of Biomedicine and Biotechnology, in press.*

Gui, J. and Li, H. (2004) Penalized Cox Regression Analysis in the High-Dimensional and Low-sample Size Settings, with Applications to Microarray Gene Expression Data. submitted .

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

Han, A. K. (1987) Non-parametric Analysis of a Generalized Regression Model. *Journal of Econometrics*, **35**, 303-316.

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

Horowitz, J. L. (1992) A Smoothed Maximum Score Estimator for the Binary Response Model. *Econometrica*, **60**, 505-531.

Kim, Y. and Kim, J. (2004) Gradient LASSO for feature selection. *Proceedings of the 21st International Conference on Machine Learning.*

Li, H. and Gui, J. (2004) Partial Cox Regression Analysis for High-Dimensional Microarray Gene Expression Data. *Bioinformatics*, **20**(suppl 1), i208-i215.

Ma, S., Kosorok, M.R., and Fine, J.P. (2004) Additive risk models for survival data with high dimensional covariates. *UW Madison Biostatistics and Medical Informatics TR186.*

Nolan, D. and Pollard, D. (1987) U-processes: rates of convergence. *Annals of Statistics* **15**, 780–799.

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

Pepe, M.S. (2003) *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press, United Kingdom.

Pepe, M.S., Janes, H., Longton, G., Leisenring, W., Newcomb, P. (2004a) Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology*, **159**, 882-890.

Pepe, M.S., Cai, T. and Zhang, Z. (2004b) Combining predictors for classification using the area under the ROC curve. *University of Washington Biostatistics Working Paper Series.*

Sherman R. (1993). The limiting distribution of the maximum rank correlation estimator. Econometrica, **61**: 123-137.

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.*

Su, J.Q. and Liu, J.S. (1993) Linear combinations of multiple diagnostic markers. *JASA* **88**, 1350–1355.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, B* **58**, 267–288.

Van der Vaart, A. W. and Wellner, J. A. (1996) *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer, New York.

Wahba, G. (1990) *Spline models for observational data.* SIAM. CBMS-NSF Regional Conference Series in Applied Mathematics.

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.

Yan, L., Dodierm, R., Mozer, M.C. and Wolnieicz, R. (2003) Optimizing classifier performance via approximation to the wilcoxon-mann-witney statistic. *Proceedings of the 20th International Conference on Machine Learning.*

Table 1. TGDR: tuning parameter section features. variable: number of variables with non-zero coefficients.

		Colon			Estrogen	
τ	k	variable	AUC	k	variable	AUC
0.0	448	500	0.943	117	500	0.945
0.1	444	500	0.959	69	500	0.962
0.2	440	500	0.946	106	500	0.944
0.3	466	500	0.933	77	500	0.949
0.4	479	467	0.959	61	485	0.965
0.5	554	351	0.962	70	444	0.962
0.6	638	266	0.946	92	306	0.939
0.7	950	150	0.963	153	187	0.945
0.8	1410	74	0.964	284	104	0.966
0.9	2320	42	0.954	1280	55	0.956
1.0	4280	29	0.954	2630	24	0.958

Table 2. Colon data (genes with non-zero coefficients) gene IDs and estimates.

Gene ID	LASSO	TGDR	Gene ID	LASSO	TGDR
Hsa.467	-0.867	-0.922	Hsa.1013	-0.164	–
Hsa.18664	0.236	–	Hsa.8147	–	-1.142
Hsa.81	–	0.534	Hsa.36689	-0.862	-1.619
Hsa.24506	–	-0.456	Hsa.37937	–	-1.332
Hsa.949	0.564	–	Hsa.2487	–	1.428
Hsa.3306	0.819	0.775	Hsa.10047	–	0.448
Hsa.2856	–	0.561	Hsa.692	-0.778	–
Hsa.549	2.127	–	Hsa.8214	–	0.310
Hsa.3016	–	0.930	Hsa.5392	0.303	1.241
Hsa.341	–	0.798	Hsa.2808	-0.335	-1.157
Hsa.789	–	0.708	Hsa.24582	–	0.953
Hsa.2210	0.286	–	Hsa.1731	0.323	–
Hsa.43405	–	-0.979	Hsa.2928	3.508	1.212
Hsa.17426	0.442	1.188	Hsa.41159	0.556	–
Hsa.33268	–	0.191	Hsa.1454	-0.671	-1.737
Hsa.627	-1.000	-1.000	Hsa.6814	0.739	–
Hsa.1145	–	-0.038	Hsa.1387	–	-0.696
Hsa.2519	–	-0.292	Hsa.43331	-0.267	-1.237
Hsa.72	–	0.501	Hsa.41260	–	0.230

Table 3. Estrogen data (genes with non-zero coefficients) gene IDs and estimates.

Gene ID	LASSO	TGDR	Gene ID	LASSO	TGDR
D38437-f	0.397	–	D87468	–	-0.199
HG2247-HT2332	–	-0.328	HG4716-HT5158	-0.407	-1.218
L21998	-0.252	–	L26336	–	0.947
L40401	–	0.097	M16447	–	0.677
M24485-s	-0.500	–	M26311-s	–	-0.852
M32053	0.387	–	M62403-s	–	0.145
U01062	-0.260	-0.599	U17077	-0.309	–
U31814	-0.327	–	U41060	0.927	0.970
U42408	–	-0.842	U43944	–	-0.510
U45955	–	-0.592	U82169	–	-0.457
U84011-s	0.289	0.030	U96113	0.527	0.730
X03635	-1.000	-1.000	X06268	–	0.261
X17059-s	0.429	–	X57809-s	–	-0.411
X72841	–	0.548	X76180	–	0.696
X86693	–	0.217	X87237	-0.634	–
X92814	–	0.020	X95876	–	-0.476