Microarray Gene Expression Data with Linked Survival Phenotypes: Diffuse Large-B-Cell Lymphoma Revisited

Mark R. Segal

Division of Biostatistics, University of California, San Francisco, CA 94143-0560

Abstract: Diffuse large-B-cell lymphoma (DLBCL) is an aggressive malignancy of mature B lymphocytes and is the most common type of lymphoma in adults. While treatment advances have been substantial in what was formerly a fatal disease, less than 50% of patients achieve lasting remission. In an effort to predict treatment success and explain disease heterogeneity clinical features have been employed for prognostic purposes, but have yielded only modest predictive performance. This has spawned a series of high profile microarray-based gene expression studies of DLBCL, in the hope that molecular level information could be used to refine prognosis. The intent of this paper is to reevaluate these microarray-based prognostic assessments, and extend the statistical methodology that has been used in this context.

Methodological challenges arise in using patients' gene expression profiles to predict survival endpoints on account of the large number of genes and their complex interdependence. We initially focus on the Lymphochip data and analysis of Rosenwald et al., (2002). After describing relationships between the analyses performed and gene harvesting (Hastie et al., 2001), we argue for the utility of penalized approaches, in particular LARS-Lasso (Efron et al., 2004). While these techniques have been extended to the proportional hazards / partial likelihood framework, the resultant algorithms are computationally burdensome. We develop residual-based approximations that eliminate this burden yet perform similarly. Comparisons of predictive accuracy across both methods and studies are effected using time-dependent ROC curves. These indicate that gene expression data, in turn, only delivers modest predictions of post therapy DLBCL survival. We conclude by outlining possibilities for further work.

KEY WORDS: Diffuse large-B-cell lymphoma; Gene harvesting; Least angle regression; Microarray; Proportional hazards; Time-dependent ROC curve

1 Introduction

Diffuse large-B-cell lymphoma is an aggressive malignancy of mature B lymphocytes. It is the most common type of lymphoma in adults with an annual incidence exceeding 25,000 cases in the United States and accounting for 40% of cases of non-Hodgkins lymphoma. While treatment advances have been substantial in what was formerly a fatal disease, less than 50% of patients achieve lasting remission. In an effort to predict treatment success and explain disease heterogeneity five clinical features (age, tumor stage, serum lactate dehydrogenase concentration, performance status, number of extranodal disease sites) were synthesized into the International Prognostic Index (IPI). The modest predictive performance of the IPI and other variables has spawned a series of high profile DNA microarray investigations into DLBCL, in the hope that molecular level information could be used to refine prognosis.

DNA microarray technology, fast emerging as one of the most widely used and powerful tools for a suite of genomic applications, can profile gene expression of an organism on a whole genome scale. Due to a variety of factors the data structures and attendant research questions surrounding microarray studies are not amenable to standard "off-the-shelf" analyses. Consequently, there has been a correspondingly rapid rate of development of new statistical methodologies; see for example Parmigiani et al., (2003) and Speed (2003). Much of this attention has focused on preprocessing, clustering, identifying differentially expressed genes or discrimination. Less effort has been directed to regression problems where we have a linked, continuous phenotype (outcome). Here, we address the situation where the phenotype is a (right censored) survival time – time from DLBCL treatment to death.

The paper is organized as follows. The suite of DLBCL microarray studies is briefly overviewed next. Chronologically tracing this progression is informative with regard the evolution of statistical approaches. This provides motivation for application of the gene harvesting methodology proposed by Hastie et al., (2001) as detailed in Section 2. Results so obtained, along with some putative shortcomings of gene harvesting, further motivate use of LARS-Lasso procedures (Efron et al., 2004). Extension of these methods to censored survival phenotypes was developed by Gui and Li (2004), as outlined in Section 3.1. The resultant techniques are, however, very computationally intensive. A "residual finesse" strategy, described in Section 3.2, is established to mitigate these concerns. Results from applying these methods, as well as several others advanced in the literature for this data, are presented in Section 3.4. Parallel analyses straddling different microarray platforms are summarized in Section 3.5. One tool used to facilitate comparisons between methods and datasets is time-dependent ROC curves (Heagerty et al., 2000), briefly described in Section 3.3. Finally, Section 4 offers some concluding discussion and possibilities for further work.

1.1 Microarray Studies of DLBCL

One of the early success stories showcasing microarray technology was the work of **Alizadeh** et al., (*Nature*, 2000). This paper introduced the "Lymphochip", a custom designed cDNA microarray that was enriched with genes that are preferentially expressed in lymphoid cells as well as with genes implicated in processes pertinent to cancer or immunology. We reanalyze data deriving from Lymphochip experiments in subsequent sections. Substantively, Alizadeh et al., use Lymphochip gene expression data to effect molecular subtyping of DLBCL, identifying two novel and distinct forms corresponding to cell of origin: "germinal center B-like" and "activated B-like", with the former exhibiting significantly better overall survival. Such molecular subtyping represents a consequential advance in explaining the considerable clinical heterogeneity implicit in the DLCBL diagnostic category that had resisted attempts based morphologic and pre-microarray molecular parameters.

Importantly, subtype delineation did not make recourse to the linked survival data, but rather was based solely on gene expression data using only unsupervised methods. In particular, hierarchical clustering using average linkage and correlation distance (Eisen et al., 1998) was used to generate the now familiar "heat map" matrix. This graphically depicts gene expression profiles across differing samples, with the gene order rearranged according to the cluster dendrogram. Informal, visual inspection of this heat map led to the identification of "signatures" – gene clusters possessing ostensibly coherent expression patterns that can be labeled according to the function, or cell type where expressed, of the constituent genes. Subsequently, survival differences between select signatures are evaluated and served to establish these as novel molecular subtypes.

It is worth re-emphasizing the informal nature of this process. No criteria or algorithm is invoked in the initial signature selection. While the eye-brain interface is adept at such visual tasks, the dimensions and variation common to microarray studies imparts difficulty to this clearly subjective signature extraction process. The heterogeneity of clusters so extracted motivated the development of "tight clustering" (Tseng and Wong, 2004). Further, the use of unsupervised methods with after-the-fact assessment of phenotype (here survival) differences has been criticized relative to direct, supervised methods (Smyth et al., 2003). We revisit these concerns subsequently, but note here that it is possible to imbue the signature selection task with some formalism. Firstly, a method for determining an appropriate number of clusters is applied. Examples of such techniques include permutation based (Tibshirani et al., 2001; Fridlyand and Dudoit, 2002) or model based (Yeung et al., 2001) approaches. Secondly, the putative labeling of these clusters can be pursued using tools such as GoMiner, MAPPFinder and EASE that identify enriched functional or other categories relative to nominated, annotated databases.

Shipp et al., (*Nature Medicine*, 2002) use gene expression data obtained from use of Affymetrix Hu6800 oligonucleotide microarrays (Affymetrix, Santa Clara, CA). Again, they have linked survival phenotypes. However, in addition to arguing for the value of microarray-based profiling, they forcefully emphasize differences between unsupervised and supervised analyses. Failure to reproduce Alizadeh et al's association between cell-of-origin subtypes and survival outcome is partly ascribed to these differences. Yet, while advocating supervised methods, Shipp et al acknowledge that their approach of reducing the survival phenotype into a dichotomy (cured versus fatal/refractory disease), as opposed to using actual survival times, is limiting. In dealing with a dichotomous outcome the machine learning term "supervised" equates to the statistical term "classification". The classifier employed by Shipp et al is so-called "weighted voting" which was previously used in a celebrated microarray classification study to discriminate between acute lymphocytic leukemia and acute mylogenous leukemia (Golub et al., 1999). It has been shown (Dudoit et al., 2002; Tibshirani et al., 2002) that weighted voting essentially coincides with a penalized version of linear discriminant analysis wherein the sample covariance matrix is replaced by its diagonal. Some such (strong) regularization has been widely demonstrated to be effective in microarray-based prediction/classification problems on account of dimensional considerations – the number (p) of covariates (genes) greatly exceeds the number (n) of samples (arrays). The bias-variance tradeoff implications of this $p \gg n$ configuration necessitate regularization and/or simple methods and this frequently dovetails with interpretability concerns whereby interest is in selecting small subsets of important genes. We subsequently (Section 3.2) adopt a similar strategy, employing regularization via L_1 penalization that simultaneously confers gene selection, and using diagonal approximations to enhance computational practicality.

The next article of note is that of **Rosenwald** et al., (*New England Journal of Medicine*, 2002). This paper not only utilizes supervised methods applied directly to survival outcomes but does so in the context of a relatively large study comprizing some 240 patients. Expression profiling again makes recourse to Lymphochip DNA microarrays. It is to this dataset that our primary re-analyses are applied. A detailed description of the analytic approach adopted by Rosenwald et al., along with salient results, is provided in the next section, where relationships with gene harvesting are indicated.

The paper by **Wright** et al., (*PNAS*, 2002) was primarily concerned with attempting to reconcile the disparate findings (Alizadeh et al., *vs.* Shipp et al.,) with regard the role (or lack thereof) of cell-of-origin subtypes. A central methodologic component is the use of "compound covariates" (called linear predictor scores) to discriminate between subtypes. Such a base classifier is closely connected to weighted voting and shrunken centroid approaches; see Tibshirani et al., (2002). The latter is distinguished by its use of soft, rather than hard, threshholding in constructing the classifier, a strategy that has been advocated in other settings (Donoho and Johnstone, 1994), and which we employ subsequently. Wright et al., demonstrate that subtypes and attendant survival differences can be elicited from the (Affymetrix) expression profiling data obtained by Shipp et al.

Finally, we summarize the work of **Lossos** et al., (*NEJM*, 2004). Motivated by the desire to devise a simple, clinically practicable prediction model for survival in DLBCL patients based on gene expression, they retreat from microarray-based expression profiling in favor of quantitative reverse-transcriptase polymerase chain reaction (RT-PCR) expression measurements applied to a select and targeted list of 36 genes. This list derives in part from the abovementioned microarray studies. The fact that no overlap exists among the genes included in the prediction models derived by Shipp et al., and Rosenwald et al., further motivates Lossos et al to devise an alternate prediction scheme. Univariate Cox proportional hazards models are used to associate each of the 36 genes with survival for 66 patients. A somewhat arbitrary thresholding of the associated Wald z statistic is used to select the 6 most significant genes. These are then included in a multivariate proportional hazards model and the resultant coefficients used as the basis for a survival prediction model and, via stratification, the construction of prognostic groups. Validation of these predictions and groups is assessed by applying the derived 6 gene model to the Shipp et al., and Rosenwald et al., microarray data. We further evaluate both the 6 gene model and this approach to validation in Sections 3.4 and 3.5.

2 Gene Harvesting for Lymphochip

As indicated, Rosenwald et al., (2002) employ the Lymphochip cDNA microarray in developing predictive models of survival (from time of chemotherapy) in DLBCL patients. These Lymphochips contain 7399 features that represent 4128 genes. Loosely, we will use the terms genes and features interchangeably. Initially, a hierarchical clustering of the expression data is performed in order to elicit gene signatures. This clustering, which employs average linkage and correlation distance, was applied to a superset of the data for which survival prediction is pursued. Importantly, both training and test (validation) DLBCL cases are used.

The process whereby signatures are identified is, as mentioned, quite subjective. Indeed, in the concluding remarks of the article cited re the signature extraction process (Shaffer et al., 2001), it is stated “The concept of a gene expression signature is somewhat fluid and operationally defined.” However, it is unclear what constitutes “operationally defined” here – no algorithmic or computable definitions are proffered; rather there is sole reliance on visual inspection of heatmaps. Beyond the concerns with this visual approach that were stated previously and led to the development of tight clustering, there are further issues surrounding the clustering itself. Even with regard hierarchical clustering there are choices for linkage type and distance metric, there being nothing sacrosanct about average linkage and correlation distance. Some sensitivity analysis and stability assessment of designated signatures with respect to differing specifications seems warranted. And, of course, there are numerous alternative clustering schemes.

In parallel, univariate – one gene at a time – Cox proportional hazards regressions are fitted to relate expression to survival. This is done on the training data ($n_{train} = 160$). The partitioning into training and test data is somewhat unorthodox with (some) selection based on survival phenotype being applied, as opposed to random assignment. The 35 genes that achieved univariate significance, and which satisfied variability and missingness criteria, were then molded into a multivariate predictor as follows. Where possible, the 35 genes were assigned to the previously identified signatures: 9 to the major histocompatibility complex (MHC) class II signature, 3 to the germinal-center B-cell signature, 3 to the proliferation signature, and 6 were to the lymph-node signature. The remaining 14 genes were not associated with any particular gene-expression signature. Expression values for those genes ascribed to each of the four signatures were then averaged (within patients) yielding four signature expression profiles. An exception was made for the MHC class II signature where only the 4 (of 9) most significant genes were used, due to high between-gene correlations.

These steps provided the building blocks for development of a multivariate predictor. Initially, all four of the signature expression profiles are forced into a Cox proportional hazards regression. Then forward stepwise methods were used to add genes from the set of 14 genes that were not affiliated with any of the signatures. This resulted in the inclusion of one additional gene, *BMP6*. The resultant linear predictor score, based on these five terms and attendant coefficients, constitutes Rosenwald et al's gene-expression-based survival predictor. This is validated using the test data by testing the corresponding predictor (retaining training data-based coefficients) for association with survival, as well as by stratifying (into quartiles) based on predictor score, and testing between strata survival differences.

While this prediction strategy moves beyond the earlier unsupervised approach of Alizadeh et al., (2000) by direct use of survival outcomes, it nonetheless retains the subjectivity and arbitrariness surrounding use of cluster-based signatures as outlined above. Furthermore, several selection criteria are invoked (univariate significance level, variability, within-signature correlation, stepwise entry significance level). Evaluating the performance characteristics of the resultant predictor requires evaluation of the entire model-building process – this is particularly consequential in the microarray setting due to dimensionality concerns; see Hastie, Tibshirani and Friedman (2001), West et al., (2001), Simon et al., (2003) and Pittman et al., (2004).

We address the central concern – signature arbitrariness/subjectivity – by appeal to methodology devised expressly for supervised analysis of microarray data, namely gene harvesting (Hastie et al., 2001). In fact, gene harvesting represents a formalization of the approach pursued by Rosenwald et al. that overcomes some of the shortcomings. Gene harvesting commences with the same hierarchical clustering scheme as used by Rosenwald et al., (2002) to elicit signatures. But, in our application while using the same 7399 Lymphochip features, we restrict to expression data derived from the 160 DLBCL patients constituting the training data, thereby avoiding any possible “data re-use” concerns. Rather than visual extraction of signatures, gene harvesting treats the average expression profile from each node (cluster) in the resulting dendrogram as a potential covariate. So, if we start with p features, this procedure augments with an additional $p - 1$ expression profiles that serve as covariates for the supervised (regression) modeling. With survival outcomes, this modeling is forward stepwise Cox proportional hazards regression, akin to that employed by Rosenwald et al., (2002). However, rather than using significance criteria, whose known distortion in the face of greedy selection algorithms will be compounded here due to large p, gene harvesting uses cross-validation to determine the number of terms retained. Provision is also made for between-gene interactions, non-linear effects, and for biasing results toward selection of larger clusters.

Hastie et al., (2001) claim two advantages for this approach. Firstly, because of the familiarity of hierarchical clustering (e.g., Eisen et al., 1998) in unsupervised analyses of microarray expression data, the usage of clusters as covariates will be convenient for interpretation. Indeed, it is arguably this familiarity that led to the signature-based approaches. Secondly, by using clusters as covariates, selection of correlated sets of genes is favored, which in turn potentially reduces overfitting. However, it has been demonstrated that gene harvesting can give rise to artifactual results (Segal et al., 2003) as we discuss in conjunction with the results obtained. But first we give a brief overview of the gene harvesting algorithm.

The available data from Rosenwald et al., (2002) consists of the $n \times p$ matrix of gene expression values $\mathbf{x} = [x_{ij}]$ where x_{ij} is the expression level of the j^{th} feature ($j = 1, \ldots, p = 7399$) for the i^{th} DLBCL patient ($i = 1, \ldots, n = 240$). Each patient also provides a potentially right censored survival outcome y_i, along with a corresponding censoring indicator δ_i. A hierarchical clustering algorithm is applied to the expression matrix corresponding to the $n_{train} = 160$ patients constituting the training data and, for each of the resulting clusters $c_k, k = 1, \ldots, 2p - 1$, the average expression profile $\bar{x}_{c_k} = (\bar{x}_{1,c_k}, \bar{x}_{2,c_k}, \ldots, \bar{x}_{n_{train},c_k})$ where $\bar{x}_{i,c_k} = 1/|c_k| \sum_{j \in c_k} x_{ij}$ is obtained. Note that we have included the individual genes (the tips/leaves of the dendrogram) as clusters (of size 1) in this formulation – their average expression profile coinciding with the individual gene profile.

This set of $2p - 1$ average expression profiles constitutes the covariate set ($\mathcal{C}$). A forward stepwise Cox proportional hazards regression is performed as follows. Initially, the first term in the model ($\mathcal{M}$) is the covariate that maximizes a modified partial log-likelihood score statistic, hereafter termed the score. The modification consists of a variance penalty used to inflate the score statistic denominator. This is motivated by stability concerns deriving from small variances; see Tusher et al., (2001). At each subsequent stage, candidates for inclusion consist of all products between a term in $\mathcal{M}$ and a term in $\mathcal{C}$. The term chosen for inclusion is that which most improves fit, again measured by maximizing the score, but subject to biasing in favor of selection of larger clusters. This biasing is effected by selecting the largest cluster whose score is within $100 \times (1 - \alpha)\%$ of the maximal attained score, where α is a user specified tuning parameter. The process continues until some prespecified maximum number of terms, m, have been added to the model. The number of terms retained is subsequently determined by cross-validation. However, we note that the cross-validation results presented for the illustrative gene harvesting survival analysis in Hastie et al., (2001) are puzzling in that resubstitution “error” significantly exceeds cross-validated error.

Hastie et al., (2001) restrict to second order interaction terms; i.e., product terms are limited to pairwise products. This is partly motivated by interpretational considerations and borrows from the multivariate additive regression spline (MARS) methodology of Friedman (1991). The gene harvesting model for the hazard function $\lambda(\cdot)$ for the i^{th} patient is then

$$\lambda(t_i; x) = \lambda_0(t_i) \exp\{\sum_{k \in S_1} \beta_k \bar{x}_{i,c_k} + \sum_{k,k' \in S_2} \beta_{k,k'} \bar{x}_{i,c_k} \bar{x}_{i,c_{k'}}\}. \quad (1)$$

Here, $\lambda_0(t_i)$ is the baseline hazard and S_1 connotes the set of clusters that enter singly while S_2 is the set of clusters that enter as product terms. So, $m = |S_1| + |S_2|$. Estimation makes recourse to partial likelihood (Cox, 1972) which we outline further in the next section.

Results from applying gene harvesting using average linkage and Euclidean distance are presented in Table 1. Interactions were not allowed; i.e. $S_2 = \phi$. Additionally, biasing toward the selection of larger clusters was employed with $\alpha = 0.1$.

Table 1: Gene harvesting using average linkage.

Step	Cluster	Score	Size
1	3498	3.83	14
2	3383	3.06	9
3	2357	3.61	2
4	g5730	3.34	1
5	3450	2.76	2
6	6223	2.47	4

The first cluster selected, #3498 with 14 members, contains exclusively MHC class II genes of particular subtypes. We note that MHC class II was one of the four signatures chosen by Rosenwald et al., (2002) and that there is overlap with regard subtypes.

The potential for gene harvesting to produce artifactual solutions derives from the selection of clusters for which the *average* gene expression profile is strongly associated with outcome, but the cluster itself is heterogeneous so that individual constituent gene profiles are not associated with outcome, and the average does not provide a meaningful summary. That this potential can be realized was demonstrated in Segal et al., (2003) and is exacerbated by (i) the small sample sizes typically encountered in microarray studies, and (ii) gene harvesting's use of all clusters as candidate covariates, irrespective of their coherence. Here, the MHC class II cluster identified by gene harvesting was not artifactual in that there are strong correlations between all 14 constituent genes and all are individually associated with survival.

An additional concern for gene harvesting is sensitivity to the hierarchical clustering scheme and/or distance metric used. However, here the clusters chosen were essentially invariant to linkage method (average, single, complete).

Nonetheless, selection of the MHC class II cluster was predicated on biasing covariate selection in favor of larger clusters. If this biasing is turned off ($\alpha = 0$) then the resultant selections are all singletons, i.e., individual genes. Moreover, the underlying forward selection strategy represents a very greedy algorithm (Efron et al., 2004). Such greediness is likely to be detrimental to predictive performance in the microarray $p \gg n$ setting which, as previously mentioned, rewards simple methods. Indeed, in Section 3.4 we show that here gene harvesting does relatively poorly in predicting survival in the test dataset. One avenue toward simpler models is via regularization, as considered in the next section.

3 LARS-Lasso for Survival Phenotypes

Forward stepwise regression is an adaptive, greedy algorithm. Gene harvesting combines this with "basis expansion", the addition of (numerous, derived) covariates, as constituted by the cluster averaged expression profiles. In the microarray $p \gg n$ context, this combination proves extremely costly from the perspective of *effective number of parameters* or (adaptive) degrees-of-freedom. Methods for estimating these quantities based on simulation (Ye, 1998) or permutation (Tibshirani and Knight, 1999) have been proposed. Applying the latter to a microarray study featuring a continuous response, Segal et al., (2003) showed that a single gene harvesting step cost $\approx 0.5n$, while five steps cost $0.9n$. While the quantifications are certainly problem specific, the message is that microarray regression problems are preferably tackled using some form of regularization in order to contain these costs. This same conclusion has been reached for microarray classification problems (Dudoit et al., 2002; Tibshirani et al., 2002). We employ a recently devised regularized regression technique, LARS-Lasso, (Efron et al., 2004), for which the form of regularization, an L_1 penalty, leads to interpretable predictors.

3.1 LARS-Lasso for Cox Proportional Hazards Models

Consider a standard Cox regression model of the form

$$\begin{aligned}\lambda(t) &= \lambda_0(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \dots \beta_p X_p) \\ &= \lambda_0(t)\exp(\beta' X) \qquad (2)\end{aligned}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function, $\beta = (\beta_1, \dots, \beta_p)$ is the vector of the regression coefficients, and $X = (X_1, \dots, X_p)$ is the vector of gene expression levels with the corresponding Lymphochip values of $x_i = (x_{i1}, \dots, x_{ip})$ for the i^{th} patient.

The partial likelihood (Cox, 1972) is then

$$L(\beta) = \prod_{r \in D} \frac{\exp(\beta' x_r)}{\sum_{j \in R_r} \exp(\beta' x_j))} \qquad (3)$$

where D is the set of patient indices for those experiencing the event (here death) and R_r denotes the set of indices of patients at risk at time t_r-. Let $l(\beta) = logL(\beta)$.

In the case of continuous response Tibshirani (1996) demonstrated that penalized least squares, using an L_1 bound on the coefficients, offered advantages over the heretofore used L_2 bound, as employed by ridge regression and, more recently, support vector machines (Cristianini and Shawe-Taylor, 2000). In particular, in addition to improving on prediction accuracy through shrinkage the nature of the L_1 constraint is such that interpretation is enhanced by “zeroing out” many covariates. That is, the least absolute deviation (L_1) constraint simultaneously effects shrinkage and selection, spawning the acronym "lasso": least absolute shrinkage and selection operator.

For survival outcomes Tibshirani (1997) defined the lasso estimate β via constrained partial likelihood:

$$\hat{\beta}(s) = \text{argmax} l(\beta); \qquad \text{subject to } \sum_{j=1}^{p} |\beta_j| \leq s. \tag{4}$$

Here s is a tuning parameter that determines how many coefficients are non-zero. Accordingly, varying s provides a continuous form of subset regression and thereby overcomes the instability associated with conventional, discrete versions. For estimation, Tibshirani (1997) uses an iteratively reweighted least squares (IRLS) based linearization to reformulate this constrained optimization problem as an L_1 penalized linear regression model and then applies the same quadratic programming approach used for continuous outcomes. The linearization utilizes the following components: $\eta = \beta' X$; $\mu = \partial l / \partial \eta$; $A = -\partial^2 l / \partial \eta \eta^T$; $z = \eta + A^{-} \mu$. A one term Taylor series approximation for the negative log partial likelihood, $-l(\beta)$ is then given by $(z - \eta)^T A (z - \eta)$ and the iterative estimation scheme is

1. Fix s and initialize $\hat{\beta} = 0$.
2. Compute η, μ, A and z based on the current value of $\hat{\beta}$.
3. Minimize $(z - \hat{\beta}' X)^T A (z - \hat{\beta}' X)$ subject to $\sum \mid \hat{\beta}_j \mid \leq s$.
4. Iterate between steps 2 and 3 until $\hat{\beta}$ converges.

The quadratic programming approach used for the minimization in step 3 starts from the least squares solution. Hence it does not handle the $p > n$ case and so is inapplicable to microarray applications. This limitation, along with efficiency concerns, motivated Osborne *et al.* (2000) to regard the lasso as a convex programming problem and to devise an algorithm based on homotopy methods. While the objectives of handling $p > n$ and improving efficiency were realized, the algorithm remained problematic for microarray settings as described in Segal et al., (2003). A highly efficient algorithm, least angle regression (LARS), that includes lasso as a special case, was recently devised for the continuous outcome setting by Efron et al., (2004). Remarkably, LARS has the same computational cost as a standard (unconstrained) multiple least squares fit. A software package, `lars()` for R and Splus is available from http://www-stat.stanford.edu/~hastie/Papers/LARS/. We loosely treat LARS and lasso as interchangeable here, results being equivalent for models of reasonable size and refer to the procedure as LARS-Lasso.

LARS-Lasso has been extended to survival outcomes by Gui and Li (2004). They employ a slightly modified LARS-Lasso algorithm to solve the step 3 minimization. The modification consists of mapping the minimization into the original, unweighted problem via Cholesky decomposition of A. Gui and Li apply the resultant methodology to the DLBCL Lymphochip data as analyzed by Rosenwald et al., (2002). They obtain interpretable results with good predictive performance (described subsequently) in concordance with LARS-Lasso goals. However, embedding the modified LARS-Lasso estimation scheme within the IRLS strategy for handling survival endpoints serves to undo much of the computational efficiency that LARS-Lasso formerly delivered. For example, run times were 8 hours for a single solution and 2 days for ten fold cross-validation of the tuning parameter s on a Dell desktop with Intel Pentium 4 3.2GHz and 2.00GB of RAM (Jiang Gui, personal communication). Next, we describe an estimation strategy that restores the computational efficiency of LARS-Lasso.

3.2 Residual Finesse

The computational burden associated with Gui and Li's (2004) adaptation of LARS-Lasso to survival outcomes derives from embedding the (modified) LARS-Lasso algorithm within the iterative reweighted least squares algorithm, which in turn is required to cope with the complexities of the Cox proportional hazards model. However, at the expense of some approximation, these can be eliminated using what we term a "residual finesse". This strategy involves substituting suitably chosen residuals for the survival endpoint, enabling inheritance of simple algorithms applicable to continuous outcomes and bypassing difficulties deriving from censoring. Residual finesse has been employed to adapt additive (Cox) models (Grambsch, Therneau and Fleming, 1995; Segal et al., 1995), multiple adaptive regression splines (MARS) (LeBlanc and Crowley, 1999) and regression trees (LeBlanc and Crowley, 1992; Keleş and Segal, 2002) to censored survival outcomes. Perhaps anticipating such a strategy Gui and Li (2004) pose two questions/objections: (i) which residuals should be used amongst the competing types available for survival data, and (ii) there are no guarantees that optimizing resultant residual sums of squares (RSS) improves fit to the survival data (Therneau and Grambsch, 2000). In response to (i) we next make a case for use of *deviance residuals*, and defer consideration of (ii) to the Discussion.

For the Rosenwald et al., (2002) lympochip study, we are in the simple setting where there are no time-dependent covariates, at most one event per patient, and each patient is under observation from time $t = 0$ (time of treatment). The martingale residual for the i^{th} patient then takes a simple nonintegral form and can be interpreted as the difference between observed and expected number of events. This expected number is

$$\hat{E}_i = \hat{E}_i(\beta) = \exp(\hat{\beta}' x_i)\hat{\Lambda}_0(t_i; \hat{\beta}) \tag{5}$$

and the Martingale residual is $\hat{M}_i = \delta_i - \hat{E}_i$. Deviance residuals are derived from the highly skewed martingale residuals by a normalizing transformation analogous to deviance residuals for Poisson regression:

$$\begin{aligned} \hat{D}_i &= \text{sign}(\hat{M}_i) \times \sqrt{-\hat{M}_i - \delta_i \log((\delta_i - \hat{M}_i)/\delta_i} \\ &\approx \frac{\delta_i - \hat{E}_i}{\sqrt{\hat{E}_i}} \end{aligned} \tag{6}$$

where the approximation follows from a one-term Taylor expansion (Therneau and Grambsch, 2000).

Returning to the LARS-Lasso Cox proportional hazards estimation scheme of Gui and Li, we can show that $(\text{diag}(A))_i \approx \hat{E}_i$. Further, these diagonal elements of A dominate the off-diagonals of A by an order of magnitude (Tibshirani, 1997). Also, we have that $(z - \hat{\beta}' X) =$

$A^{-}M$. Accordingly, the minimization (step 3) can be approximated as

$$
\begin{aligned}
(z-\hat{\beta}'X)^{T}A(z-\hat{\beta}'X) &= (A^{-}M)^{T}A(A^{-}M) \\
&\approx \sum_{i=1}^{n}(\delta_i-\hat{E}_i)^2/\hat{E}_i \\
&\approx RSS(\hat{D}) \qquad (7)
\end{aligned}
$$

the residual sum-of-squares of (first order Taylor approximation) deviance residuals.

So, in order to effect the L_1 constrained minimization of step 3 we initially compute null deviance residuals and using these as outcomes for the computationally efficient standard (continuous response) LARS-Lasso algorithm. Null deviance residuals can be obtained simply in Splus or R by specifying residual type and zero iterations in the call to `coxph()`.

3.3 Time-Dependent ROC Curves

Typically, in the microarray - DLBCL survival literature, the evaluation of how well a predictive model performs is done as follows: (i) risk scores based on the fitted model are computed for patients in a (withheld) test dataset, (ii) strata (usually two) are created based on thresholding these scores, and (iii) log-rank testing of between-strata survival differences is performed. The greater the achieved significance the more predictive the model is deemed. Limitations of this approach include not just the arbitrariness of the imposed stratification but, more importantly, the familiar shortcoming of p-values not necessarily capturing effect size / explained variation.

A more refined approach is afforded by use of time-dependent ROC curves, proposed by Heagerty et al., (2000) and used in the present context by Gui and Li (2004). Denote the predictive model as $f(X)$ and define time-dependent sensitivity and specificity functions at cutoff point c as

$$
\begin{aligned}
\text{sensitivity}(c,t|f(X)) &= Pr\{f(X)>c|\delta(t)=1\} \qquad (8) \\
\text{specificity}(c,t|f(X)) &= Pr\{f(X)\leq c|\delta(t)=0\} \qquad (9)
\end{aligned}
$$

where $\delta(t)$ is the event indicator at time t. The corresponding time-dependent ROC curve at time t, $\text{ROC}(t|f(X))$, is then just the plot of $\text{sensitivity}(c, t|f(X))$ vs $1 - \text{specificity}(c, t|f(X))$ with cutoff point c varying. We use the analogous area under the time-dependent $\text{ROC}(t|f(X))$ curve, $\text{AUC}(t|f(X))$, as a summary in the next sections. In order to estimate the conditional probabilities in (8) and (9), accounting for possible censoring, we follow Heagerty et al., in employing a nearest neighbor estimator for the bivariate distribution function (Akritas, 1994).

3.4 Results

Figure 1 shows coefficient trajectories as obtained by applying LARS-Lasso to the training data with null deviance residuals as outcome on expression data as covariates. The individual trajectories correspond to individual genes and can best be interpreted when viewed as functions of increasing values of the L_1 constraint s. When $s = 0$ there are no terms (genes) in the model. As s increases successively more genes enter. Note that while the magnitude of the coefficients tend to increase with increasing s, there are instances of non-monotonicity. This is due to between-gene correlation. From both an interpretative and predictive standpoint the critical issue is determining an appropriate value of s.

We approached determination of s using the built-in cross-validation routines provided by `lars()`. However, results were unstable, showing considerable variation according to changes in number of cross-validation folds. Gui and Li (2004) report that a model with 4 genes minimizes cross-validated partial likelihood, but they do not provide standard errors. For comparison purposes we focus on similarly sized models. However, we emphasize that such determinations are inherently unstable in the $p \gg n$ microarray setting as detailed in Segal et al., (2003) and further discussed in Section 4.

The top 4 genes selected, and the order in which they are chosen, using L_1 penalized partial likelihood (Gui and Li, 2004) coincide with the top 4 selections (and order) that result from use of our deviance residual finesse. Indeed, the agreement extends appreciably further, providing some indication that the approximations employed in arriving at the deviance residual finesse

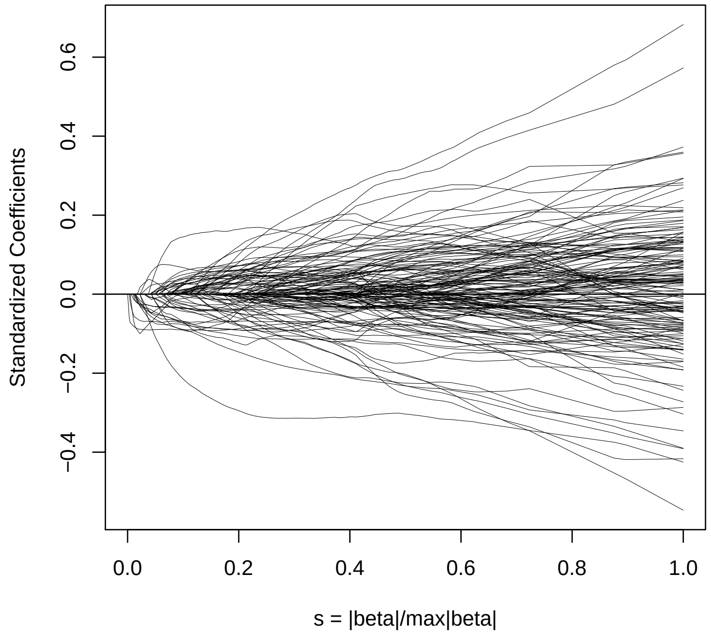

Figure 1: Gene coefficient profiles obtained by applying LARS-Lasso to null deviance residuals.

are reasonable. These top 4 genes are itemized (in order of selection) in Table 2. Strikingly, they represent 3 of the 4 signatures identified by Rosenwald et al., (2002). If we extend to the top 10 genes then 7/10 (coincident with Gui and Li's results) are representative of the signatures, with the remaining 3 having putative biological relevance.

Table 2: Top four genes selected by L_1 penalized proportional hazards and the deviance residual based approximation thereof.

GenBank ID	Signature	Description
AA805575	Germ	thyroxine-binding globulin precursor
LC_29222	Lymph	
X00452	MHC	major histocompatibility complex, class II, DQ alpha 1
X59812	Lymph	cytochrome P450, subfamily XXVIIA polypeptide 1

Next, we illustrate the standard approach to evaluating the predictive performance of models using the test data. Following Li and Gui (2004) and Gui and Li (2004) we *a priori* elect to stratify into two (high and low) risk groups by using a risk score cutpoint of zero. Results for

6 competing methods are given in Table 3.

Table 3: Stratification results for six competing predictive models.

Method	#{high risk}	log-rank p-value	R^2
Gui, Li – $L1$ penalized PH	36	0.0004	0.138
LARS-Lasso – Deviance residuals	33	0.0013	0.113
Lossos et al – 6 gene model	42	0.0136	0.074
Rosenwald et al – Signatures	32	0.0005	0.128
Harvesting – $\alpha = 0$	33	0.20	0.02
Bair, Tibshirani – SuperPC	34	0.0213	0.061

The gene harvesting entry corresponds to a model with four terms. These correspond to two singleton genes, one cluster of size two and one cluster of size three. One of the singletons is thyroxine-binding globulin precursor (AA805575) which was also chosen by the L_1 penalized approaches and is included in the germinal-center B-cell signature. The selected harvesting model did not favor inclusion of larger clusters ($\alpha = 0$). The Bair and Tibshirani (2004) model is based on 23 genes selected by thresholding univariate Cox regression test statistics (akin to Lossos et al., 2004) obtained from fitting individual genes to survival on the training data. A multivariate predictor is then obtained by combining these genes using principal components analysis. More than half (13/23) the genes belong to the MHC class II signature/cluster.

R^2 values are those provided by the `coxph()` R function and are based on partial likelihood differences as described in Nagelkerke (1991). We consider these further in the Discussion.

To overcome the arbitrariness of stratification into low and high risk groups we next obtain time-dependent ROC curves for the same six models. The corresponding time-dependent areas under the ROC curves are displayed in Figure 2.

Both the stratified and time-dependent ROC curve analyses are consistent and support the following findings.

1. The deviance residual based approximation to direct L_1 penalization seems empirically reasonable. This is no surprise given the previously noted correspondence in gene selections. But, the agreement evidenced here is additionally affirming since the associated coefficients

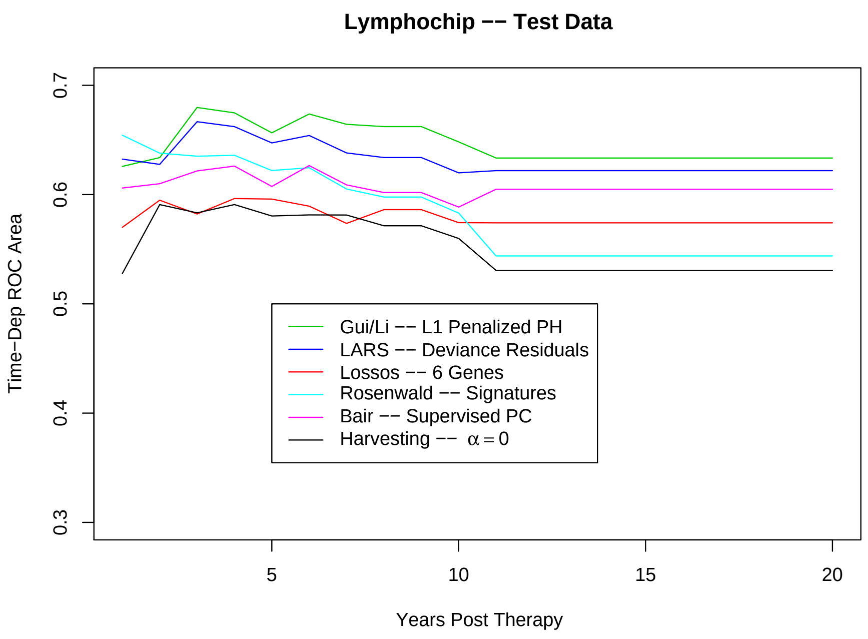

Figure 2: Areas under time-dependent ROC curves for six competing models.

are estimated under different criteria (Cox partial likelihood, squared error loss).

2. The best results are obtained by both L_1 penalization schemes and the signatures approach. Again, this is not surprising in view of the overlap in gene selections. Further, as already illustrated with the MHC Class II signature, the genes constituting the signatures themselves are strongly correlated and so the corresponding average expression profile is similar to the profile of any individual member gene.

Given that this set of methods yield the best results it is purposeful to further compare amongst them. Arguably, the deviance residual based LARS-Lasso approach is the easiest to implement and affords a more direct, less ad hoc approach than the use of signatures. That the signature approach does so well is a tribute to the perspicacity with which the signatures themselves were chosen. However, the fact the test data were used as part of this selection process is a concern. It is difficult to assess how much of a concern since the process itself is so subjective.

3. The relatively lesser performance of the Lossos et al., (2004) model and the Bair and Tibshirani (2004) model comes despite (or because) these methods use more genes (6 and 23 respec-

tively) than the above methods. This is presumably attributable to the gene selection process which, in both instances, is based on thresholding *univariate* significance levels. That Bair and Tibshirani (2004) observe diminished predictive performance when such univariate-based preselection is not employed testifies to the need for some form of regularization/selection. The results here suggest that gains can be achieved by pursuing this in a multivariate fashion as per LARS-Lasso. Of further note are the comparison studies undertaken by Bair and Tibshirani that, for the DLBCL data as analyzed here, show superior performance of their supervised principal components technique versus a suite of competitors. By extension, the L_1 penalized approaches perform better still.

4. The worst performance belongs to gene harvesting. Quite simply we ascribe this to greediness. As mentioned previously, gene harvesting, with its basis expansion and unpenalized forward selection strategy, is extravagant when cost is measured via effective number of parameters. Of further interest is the performance disparity between gene harvesting and the Rosenwald et al., (2002) signature approach. To the extent that the latter represents a model reachable via the former, the superior results for signatures again illustrate the perspicacity of signature selection and/or the re-use of test data.

5. These performance rankings should not be over-interpreted since the overall differences as measured by the time-dependent ROC areas (Figure 2) are modest. Moreover, even the best predictive performance when (properly) assessed in this fashion is modest – recall that an ROC area of 0.5 corresponds to predicting a binary outcome by a coin toss. This perspective corrects the overly optimistic viewpoint conveyed by log-rank testing of derived high/low risk groups (Table 3) and gives pause to use of the associated models / genes as the basis for treatment decisions or prognosis as has been advocated.

3.5 Cross Study Comparisons

To further assess the generality of the models specified by use of the LARS-Lasso strategy we investigated how well methods performed when applied to gene expression data obtained under alternate platforms, for which linked, post-therapy, DLBCL survival was also available. In particular, we applied the LARS-Lasso model as developed using Lymphochip to data obtained using Affymetrix Hu6800 arrays (Shipp et al., 2002). We then undertook the reverse operation, developing a model using Affymetrix data and evaluating against the Lymphochip study. Effecting these comparisons requires processing to convert (absolute) intensity measures as yielded by single channel platforms such as Affymetrix to the (relative) log-ratio expression values provided by a competitive hybridization scheme (two channel arrays in general and Lymphochip in particular). Since there is no definitive way to do this we (separately) applied two processing schemes that have been used in this setting (Wright et al., 2003; Lossos et al., 2004). Additionally, it is necessary to match genes (where possible) between platforms. We used the NetAffx resource (Liu et al., 2003) and Lymphochip GenBank accessions for this purpose.

Again, we summarize results via the area under time-dependent ROC curves. The left panel of Figure 3 presents results where the evaluation data is the same Lymphochip test set as used above. We have reproduced the curve corresponding to the six gene model of Lossos et al., (2004). Recall that this model was developed by selecting from genes implicated in studies using both Lymphochip and Affymetrix platforms. We have modified the model obtained using L_1 penalized proportional hazards from that graphed in Figure 2 and summarized in Table 3. The modification consists of eliminating the LC_29222 gene since this gene is not represented on the Affymetrix Hu6800 chip. Finally, we have used the Lymphochip test data to appraise a model developed using L_1 penalized proportional hazards on the Affymetrix data. The converse set of curves obtained when using these same models but evaluated on Affymetrix data are given in the right panel of Figure 3. The analysis leading to these curves employed the conversion approach of Lossos et al., (2004), but similar results and conclusions pertain if the conversion approach of Wright et al., (2003) is used. Since there were only 66 patients in this study for the purposes of these comparisons we have not partitioned into training and test sets. Results for the 6 gene L_1 penalized model developed on this same data are therefore optimistic. Some conclusions regarding the generalizability across datasets in particular, and the utility of microarray expression profiling for predicting DLBCL survival more broadly, can be drawn by contrasting the two panels.

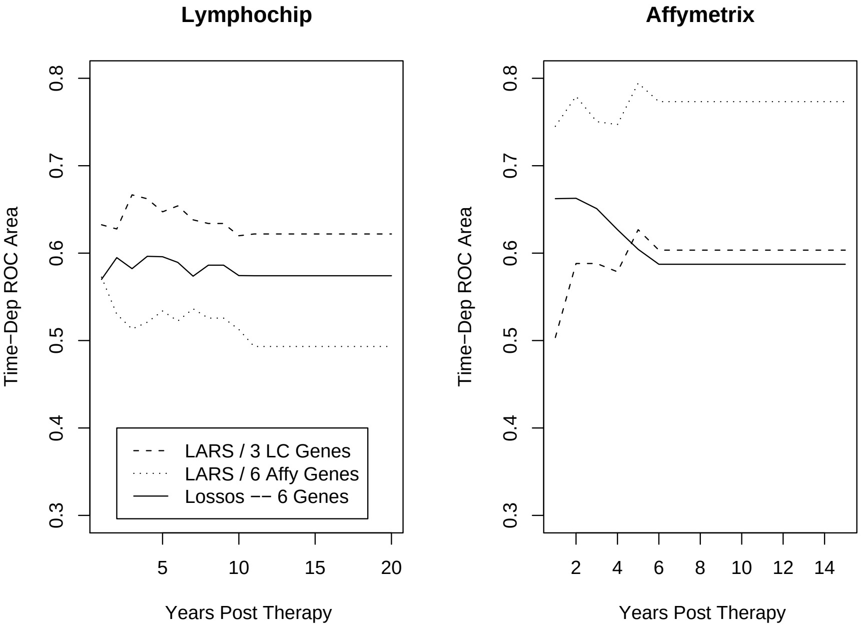

Figure 3: Areas under time-dependent ROC curves for three models compared across platforms.

The intermediary performance of the Lossos et al., (2004) model can be ascribed to the manner in which it was developed. While appeal was made to both microarray platforms in distilling a list of 36 candidate genes for construction of a multivariate predictor, the model itself was based on direct RT-PCR measurement of corresponding gene expression. Accordingly, it does not perform as well as models developed under the respective platforms that reemploy microarray-based expression measurements. Analogously, the Lympochip based model is best with Lympochip data and *vice versa* for the Affymetrix based model. But, what is additionally notable is poor performance of all models away from the settings in which they were developed. Lossos et al., (2004) also reapply their model to these Lymphochip and Affymetrix datasets and contend that their predictor validates. Their stated objective was "to devise a model that was technically simple and applicable for routine clinical use." When validity is assessed via time-dependent ROC curves, instead of risk group stratification, it appears that this goal has not been realized, more accurate predictions being required for clinical prognosis.

4 Discussion

In our reanalysis of microarray studies of gene expression with linked DLBCL survival phenotypes, a number of interesting findings emerged from both a methodologic and substantive standpoint, summarized in Sections 3.4 and 3.5 respectively. These results are perhaps surprising to the extent that (i) the performance of gene harvesting, which ostensibly generalizes the signature based approach of Rosenwald et al., (2002), is notably inferior, and (ii) there is striking agreement between the algorithmic, L_1 penalized proportional hazards methods and the subjective, signature-based approach. For (i), the explanation lies in the greediness of the gene harvesting selection process which is averted by the explicit hand-picking of signatures. That this hand-picking does so well, as indicated by (ii) and the results given in Table 3 and Figure 2, is likely attributable to a combination of investigator insight and/or signature extraction using test as well as training data. In any event, the use of L_1 penalized methods for microarray studies affords a good technique that balances prediction and interpretation. For linked continuous phenotypes this approach is made computationally viable through use of the LARS algorithm, with the deviance residual finesse enabling extension to survival phenotypes.

As always, there are several avenues for further work. In order to couple the sparse variable selection attributes of L_1 penalties with clustering/signature constructs, whereby sets of strongly correlated variables are either jointly included or excluded, Zou and Hastie (2003) developed the "elastic net" which combines L_1 and L_2 penalties for continuous phenotypes. Extensions to survival outcome analogous to that devised for LARS-Lasso warrant development and exploration.

As indicated in Section 3.4 model selection via cross-validation is highly unstable in this (microarray) setting, the $p \gg n$ data configuration accentuating this well known attribute of cross-validation. Indeed, Efron (2004) shows that covariance penalty based model selection schemes (e.g., AIC) can be viewed as Rao-Blackwellized versions of cross-validation and therefore are conferred with greater stability. This leads Efron to advocate use of covariance penalty methods despite these approaches requiring additional model assumptions. However, another required ingredient is an estimate of error variance, σ^2. Typically, it is recommended that this be obtained from a "big" model. While some guidelines have been advanced for specifying such big models these pertain to the $n > p$ situation and breakdown for microarray problems. Trying to mimic specifications based on degrees of freedom or effective numbers of parameters (see Section 3) is circular since these also require an estimate of σ^2. In the particular case of LARS-Lasso we could base model size on the onset of non-monotonicity in the coefficient traces (*cf* Figure 1) as discussed in Section 3.4.

We return to the issue of minimizing residual based R^2's not equating to improving fit for censored survival data, as raised by Therneau and Grambsch (2000); see Section 3.2. A key consideration is how fit is measured in this context. The R^2 measure provided by Therneau's R function `coxph`, and featured in Table 3, is based on information theoretic ideas (Nagelkerke, 1991). However, O'Quigley and Xu (2001) show that an equivalent criterion can be expressed as a weighted R^2 involving Schoenfeld residuals. Heagerty and Zheng (2003) describe analogies between the underpinnings of O'Quigley and Xu's formulation and time-dependent ROC curves.

We view the superior performance of the L_1 penalized approaches noted in Section 3.4, as being illustrative of the "principle" enunciated by Rosset et al., (2004): "Working in high dimension and regularizing is statistically preferable to a two-step procedure of first reducing the dimension, then fitting a model in the reduced space." Zhu and Hastie (2004) use L_2 penalized logistic regression to pursue classification in the context of microarray cancer studies with categorical outcomes. Because the L_2 penalty does not possess the built-in selection property of L_1 penalties, they need to employ feature (gene) elimination strategies. We are presently investigating LARS-Lasso style L_1 penalized algorithms for such categorical outcome types and, more generally the entire class of generalized linear models, based on analogous deviance residual finesses to that used here for survival outcomes.

References

Akritas, M.G. (1994). "Nearest Neighbor Estimation of a Bivariate Distribution Under Random Censoring," *Annals of Statistics*, 22,1299-1327.

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

Bair, E., and Tibshirani, R. (2004), "Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data," *Public Library of Science: Biology*, 2, 511-522.

Cristianini, N., and Shawe-Taylor, J. (2000), *An Introduction to Support Vector Machines*, Cambridge: Cambridge University Press.

Donoho, D.L., and Johnstone, I.M. (1994), "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, 81, 425-455.

Dudoit, S., Fridlyand, J., and Speed, T.P. (2002), "Comparison of Discrimination Methods for the Classification of Tumors using gene expression data," *Journal of the American Statistical Association*, 97, 77-87.

Dudoit, S., and Fridlyand, J. (2002), "A Prediction-based Resampling Method to Estimate the Number of Clusters in a Dataset," *Genome Biology*, 3, 0036.1-0036.21.

Efron, B., Hastie, T.J., Johnstone, I., and Tibshirani, R.J. (2004), "Least Angle Regression," *Annals of Statistics*, 32, 407-451.

Efron, B. (2004), "The Estimation of Prediction Error: Covariance Penalties and Cross-Validation," *Journal of the American Statistical Association*, 99, 619-642.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1988), "Cluster Analysis and Display of Genome-Wide Expression Patterns," *Proceedings of the National Academy of Sciences*, 95, 14863-14868.

Friedman, J.H. (1991), "Multivariate Adaptive Regression Splines," *Annals of Statistics*, 19, 1-67.

Golub, T.R., Slonim, D.K., Tamayo, P., et al. (1999), "Molecular Classification of Cancer:

Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, 286, 531-537.

Grambsch, P.M., Therneau, T.M., and Fleming, T.R. (1995), "Diagnostic Plots to Reveal Functional Form for Covariates in Multiplicative Intensity Models," *Biometrics*, 51, 1469-1482.

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," http://repositories.cdlib.org/cbmb/L1Cox/

Hastie, T., Tibshirani, R., Botstein, D., and Brown, P. (2001), "Supervised Harvesting of Expression Trees," *Genome Biology*, 2, 0003.1-0003.12.

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

Heagerty, P.J., Lumley, T. and Pepe, M. (2000), "Time Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker," *Biometrics*, 56, 337-344.

Heagerty, P.J., and Zheng, Y. (2003), "Survival Model Predictive Accuracy and ROC Curves," Technical Report, Department of Biostatistics, University of Washington.

Keleş, S., and Segal, M.R. (2002), "Residual-Based Tree-Structured Survival Analysis," *Statistics in Medicine*, 21, 313-326.

LeBlanc, M., and Crowley, J. (1992), "Relative Risk Regression Trees for Censored Survival Data," *Biometrics*, 48, 411-425.

LeBlanc, M., and Crowley, J. (1999), "Adaptive Regression Splines in the Cox Model," *Biometrics*, 55, 204-213.

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

Liu, G., Loraine, A.E., Shigeta, R., Cline, M., Cheng, J., Valmeekam, V., Sun, S., Kulp, D., and Siani-Rose, M.A. (2003), “NetAffx: Affymetrix Probesets and Annotations,” *Nucleic Acids Research*, 31, 82-86.

Lossos, I.S., Czerwinski, D.K., Alizadeh, A.A., Wechser, M.A., Tibshirani, R., Botstein, D., and Levy, R. (2004), “Prediction of Survival in Diffuse Large-B-Cell Lymphoma Based on the Expression of Six Genes,” *New England Journal of Medicine*, 350, 1828-1837.

Nagelkerke, N.J.D. (1991), “A Note on a General Definition of the Coefficient of Determination,” *Biometrika*, 78, 691-692.

O’Quigley J., and Xu R. (2001), “Explained Variation in Proportional Hazards Regression,” in *Handbook of Statistics in Clinical Oncology*, ed. J. Crowley, New York: Marcel Dekker, pp. 397-410.

Parmigiani G., Garrett, E.S., Irizarry, R.A., and Zeger, S.L. (2003), *The Analysis of Gene Expression Data*, New York: Springer.

Rosenwald, A., Wright, G., Chan, W.C., et al., (2002), “The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-Cell Lymphoma,” *New England Journal of Medicine*, 346, 1937-1947.

Rosset S., Zhu J., and Hastie, T. (2004). “Boosting as a Regularized Path to a Maximum Margin Classifier,” *Journal of Machine Learning Research*, 5, 941973.

Segal, M.R., James, I.R., French, M.A.H., and Mallal, S. (1995), “Statistical Issues in the Evaluation of Markers of HIV Progression,” *International Statistical Review*, 63, 179-197.

Segal, M.R., Dahlquist, K.D., and Conklin, B.R. (2003), “Regression Approaches for Microarray Data Analysis,” *Journal of Computational Biology*, 10, 961-980.

Shaffer, A.L., Rosenwald, A., Hurt, E.M., et al. (2001), “Signatures of the Immune Response,” *Immunity* , 15, 375-385.

Shipp, M.A., Ross, K.N., Tamayo, P., et al. (2002), "Diffuse Large B-Cell Lymphoma Outcome Prediction by Gene-Expression Profiling, and Supervised Machine Learning," *Nature Medicine*, 8, 68-74.

Smyth, G.K., Yang, Y.H., and Speed, T.P. (2003), "Statistical Issues in Microarray Data Analysis," In: Brownstein, M.J., and Khodursky, A.B. (eds.), *Functional Genomics: Methods and Protocols*, *Methods in Molecular Biology*, 224, 111-136. Totowa, NJ: Humana Press.

Speed, T.P. (ed.) (2003), *Statistical Analysis of Gene Expression Microarray Data*, New York: Chapman & Hall/CRC.

Therneau, T.M., and Grambsch, P.M. (2000), *Modeling Survival Data: Extending the Cox Model*, New York: Springer.

Tibshirani, R.J. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267-288.

Tibshirani, R.J. (1997), "The Lasso Method for Variable Selection in the Cox Model," *Statistics in Medicine*, 16, 385-395.

Tibshirani, R.J., and Knight, K. (1999), "The Covariance Inflation Criterion for Adaptive Model Selection," *Journal of the Royal Statistical Society, Series B*, 61, 529-546.

Tibshirani, R. J., Walther, G., and Hastie, T.J. (2001), "Estimating the Number of Clusters in a Dataset via the Gap Statistic," *Journal of the Royal Statistical Society*, Ser. B, 63, 411-423.

Tibshirani, R.J., Hastie, T.J., Narasimhan, B., and Chu, G. (2002), "Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression," *Proceedings of the National Academy of Sciences*, 99, 6567-6572.

Tseng, G.C., and Wong, W.H. (2004), "Tight Clustering: A Resampling-Based Approach for Identifying Stable and Tight Patterns in Data," *Biometrics*. (To Appear in December)

Tusher, V.G., Tibshirani, R.J., and Chu, G. (2001), "Significance Analysis of Microarrays Applied to the Ionizing Radiation Response," *Proceedings of the National Academy of Sciences*, 98, 5116-21.

West, M., Blanchette, C., Dressman, H., et al., (2001), "Predicting the Clinical Status of Human Breast Cancer Using Gene Expression Profiles," *Proceedings of the National Academy of Sciences*, 98, 11462-11467.

Wright, G., Tan, B., Rosenwald, A., Hurt, E.H., Wiestner, A., and Staudt, L.M. (2003), " A Gene Expression-Based Method to Diagnose Clinically Distinct Subgroups of Diffuse Large-B-Cell Lymphoma," *Proceedings of the National Academy of Sciences*, 100, 9991-9996.

Ye, J. (1998), "On Measuring and Correcting the Effects of Data Mining and Model Selection," *Journal of the American Statistical Association*, 93, 120-131.

Yeung K.Y., Fraley C., Murua A., Raftery, A.E. and Ruzzo, W.L. (2001), "Model-based Clustering and Data Transformations for Gene Expression Data," *Bioinformatics*, 17, 977-987.

Zhu, J., and Hastie, T. (2004), "Classification of Gene Microarrays by Penalized Logistic Regression," *Biostatistics*, 5, 427-444.

Zou, H., and Hastie, T. (2003), "Regularization and Variable Selection via the Elastic Net," Technical Report, Department of Statistics, Stanford University.

Acknowledgments: Assistance received from Jiang Gui, Hongzhe Li, Trevor Hastie, Rob Tibshirani, Eric Bair and Chuck McCulloch with respect to data / software provision and helpful discussions is greatly appreciated.