

Adaptive index models for marker-based risk stratification

Lu Tian* Robert Tibshirani †

November 16, 2009

Abstract

We use the term *index predictor* to denote a score that consists of K binary rules such as “age > 60” or “blood pressure > 120 mm Hg”. The index predictor is the sum of the scores, yielding a value from 0 to K . Such scores are often used in clinical studies to stratify population risk: they are usually derived from subject area considerations. In this paper we propose a fast procedure for automatically constructing such indices based on a training dataset, for linear regression, logistic regression and Cox survival models. We also extend the procedure to create indices for detecting treatment-marker interactions. The methods are illustrated on a study with protein biomarkers as well as two microarray gene expression studies.

1 Introduction

When predicting a phenotype such as clinical response or survival time from a set of biomarkers, an “index predictor” is sometimes used. This consists of a set of binary rules such as “marker $x_k \geq c_k$ ” or “marker $x_k < c_k$ ” for each of K markers. For each observation we add up the binary scores yielding an index s taking values in $\{0, 1, \dots, K\}$. This has the advantage of simplicity: it is easy to state and interpret, and also can capture situations where

*Depts. of Health, Research & Policy, Stanford Univ, Stanford CA, 94305; lujian@stanford.edu

†Depts. of Health, Research & Policy, and Statistics, Stanford Univ, Stanford CA, 94305; tibsh@stat.stanford.edu

prognostic effects are shared by multiple markers. A popular example is the International Prognostic Index (IPI) used for risk classification in Non-Hodgkins lymphoma (TIN-HsLPF (1993)). The IPI consists of one point for each of:

- Age greater than 60 years
- Stage III or IV disease
- elevated serum LDH (> 1)
- ECOG/Zubrod performance status of 2, 3, or 4
- More than 1 extranodal site.

The resulting score lies in 0 to 5, with higher scores indicating greater risk. Sometimes the IPI score is further simplified into two or three categories as (low, high) or (low, medium, and high) for risk stratification.

An example is shown Figure 1. Shown are the survival curves from a set of patients with Non-Hodgkins lymphoma, for each of the levels of the IPI. There is clear separation in the groups.

In this paper we propose a method for adaptively constructing an index predictor from a set of training data. We also return to this example and demonstrate that our proposal can re-construct the IPI empirically from a set of training data.

This paper is organized as follows. In section 2 we introduce the adaptive index model and our algorithm for its estimation. We discuss an example that in which protein biomarkers are used to predict the presence of ovarian cancer. In Section 3 we discuss the AIM model for survival model, using Cox’s proportional hazards model and illustrate how the AIM procedure can re-discover the international prognostic index (IPI) discussed above. We extend the AIM procedure to look for interactions between markers and a binary treatment factor in Section 5. We also discuss the construction of *surrogate markers*. In Section 6 we investigate the performance of the AIM procedure with a large number of predictors, and propose the use of “pre-conditioning” to avoid overfitting. The degrees of freedom of the AIM procedure is studied both mathematically and numerically in Section 7. There are clear connections to other methods such as CART (Breiman et al. (1984)), PRIM (Friedman & Fisher (1999)), their more recent refinements in LeBlanc et al. (2005) and LeBlanc et al. (2002), boosted trees (see e.g. Friedman et al. (2000), Friedman (2001)), and the “logic regression”. (Ruczinski et al. (2003)). We discuss these and make some concluding remarks in Section 8.

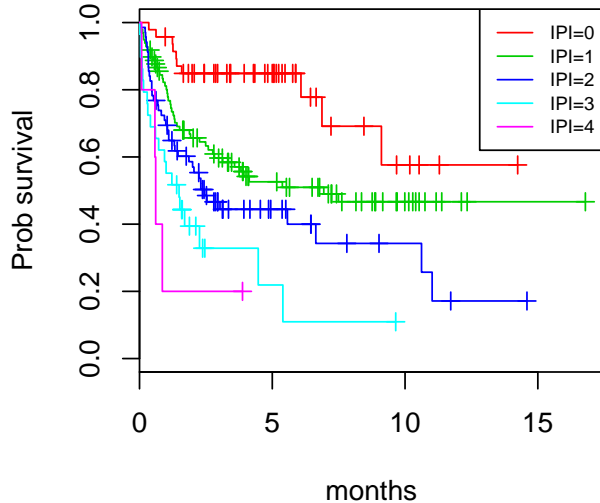


Figure 1: *Survival curves from a set of patients with Non-Hodgkins lymphoma, for each of the levels of the International Prognostic Index (IPI).*

2 Adaptive index models

Consider a supervised learning problem (regression or generalized regression) with data (x_i, y_i) , $i = 1, 2, \dots, N$. Here x_i is a p -vector of predictor variables and y_i in an outcome variable. The three major applications that we consider are the linear regression model, the logistic model for binary data where $y_i \in \{0, 1\}$ and Cox's proportional hazards model for survival data where $y_i = (T_i, \delta_i)$, where T_i is a right censored survival time and δ_i is the censoring indicator. Denote the log-likelihood or log partial log-likelihood by $\ell(\eta; \mathbf{x}, \mathbf{y})$, where η is the usual linear combination of predictors. For example η is the linear predictor in a regression model, the log-odds in the logistic model and the log-hazard in the proportional hazards model. We consider an *index model* in the form of

$$\eta = \beta_0 + \beta \cdot \sum_{k=1}^K I(\tilde{x}_k^* \leq c_k) \quad (1)$$

with $K \leq p$. The predictors \tilde{x}_k^* are from the set $\{\pm x_1, \pm x_2, \dots, \pm x_p\}$ and the corresponding cutpoints c_k are chosen in a forward stepwise manner to maximize the log-likelihood $\ell(\eta; \mathbf{x}, \mathbf{y})$. The result is a simple “index” predictor $s = \sum_{k=1}^K I(\tilde{x}_k^* \leq c_k)$ which is just a count ranging from 0 to K . By allowing \tilde{x}_k^* to equal $-x_j$, we effectively allow cuts of the complementary form $x_j \geq c_j$.

What makes our procedure attractive is the fact that as we change the cutpoint c_k , updating formulas can be derived for the score test for testing $\beta = 0$. Next we give the details of the updating scheme for linear and logistic model.

Our model is

$$\eta_i = \beta_0 + \beta \sum_{k=1}^K I(\tilde{x}_k^* \leq c_k).$$

$\eta_i = E(Y_i|x_i)$ in the linear model, while $\eta_i = \log[p_i/(1 - p_i)]$ in the logistic model, where $p_i = \text{Prob}(y = 1|x_i)$. Suppose that we have a score $s = \sum_{j=1}^{k-1} I(\tilde{x}_j^* \leq c_j)$ and want to decide whether to add a term $z = I(x^* \leq c)$. Hence we fit $\eta = \beta_0 + \beta(s + z)$ and test $\beta = 0$ in the regression model. Letting $\hat{\mu}_0 = \bar{y}$, the average of $\{y_i, i = 1, \dots, N\}$, we have the score vector and information matrices

$$U = (U_1, U_2) = \left(\sum_{i=1}^N (y_i - \hat{\mu}_0), \sum_{i=1}^N s_i (y_i - \hat{\mu}_0) \right)$$

$$\text{and} \quad I = \begin{pmatrix} n\hat{v}_0 & \hat{v}_0 \sum_{i=1}^N s_i \\ \hat{v}_0 \sum_{i=1}^N s_i & \hat{v}_0 \sum_{i=1}^N s_i^2 \end{pmatrix},$$

where \hat{v}_0 is the empirical variance of $\{y_i\}$ in both the linear and logistic models. The score test is $U_2/V^{1/2}$ where $V^{-1} = I_{11}/(I_{11}I_{22} - I_{12}I_{21})$.

Without the loss of generality, we assume that the observations are sorted according to $x^* = \{x_i^*, i = 1, \dots, N\}$, the predictor of interest, i.e., $x_1^* < x_2^* < \dots < x_N^*$. We have the following updating formulas as the cutpoint c moves from x_{i-1}^* to x_i^* :

$$U_2 = U_2 + (y_i - \hat{\mu}_0)$$

$$I_{12} \leftarrow I_{12} + \hat{v}_0$$

$$I_{22} \leftarrow I_{22} + \hat{v}_0(1 + 2s_i).$$

Thus we can scan through all possible cutpoints for a given predictor in just $O(n)$ operations. We summarize the algorithm below, called “AIM” for “Adaptive Index Models”.

AIM procedure

1. Begin with $k = 0, s = 0$.
2. For $k = 1, 2, \dots, K$ update $s \leftarrow s + I(x_j > c_j)$ or $s \leftarrow s + I(x_j \leq c_j)$ where (j, c_j) maximize the score test over the markers not yet entered.

In practice we set the maximum model size K to, say, 10 or 20 and estimate the best model size k by cross-validation.

Figure 2 shows the run time in seconds of the AIM procedure for logistic regression for various combinations of n, p and the maximum model size K . In the left and middle panels we run the algorithm until $K = 20$ terms have been added. In the right panel we have fixed p at 100. We see that the algorithm is remarkably fast, and scales roughly linearly in n, p and K .

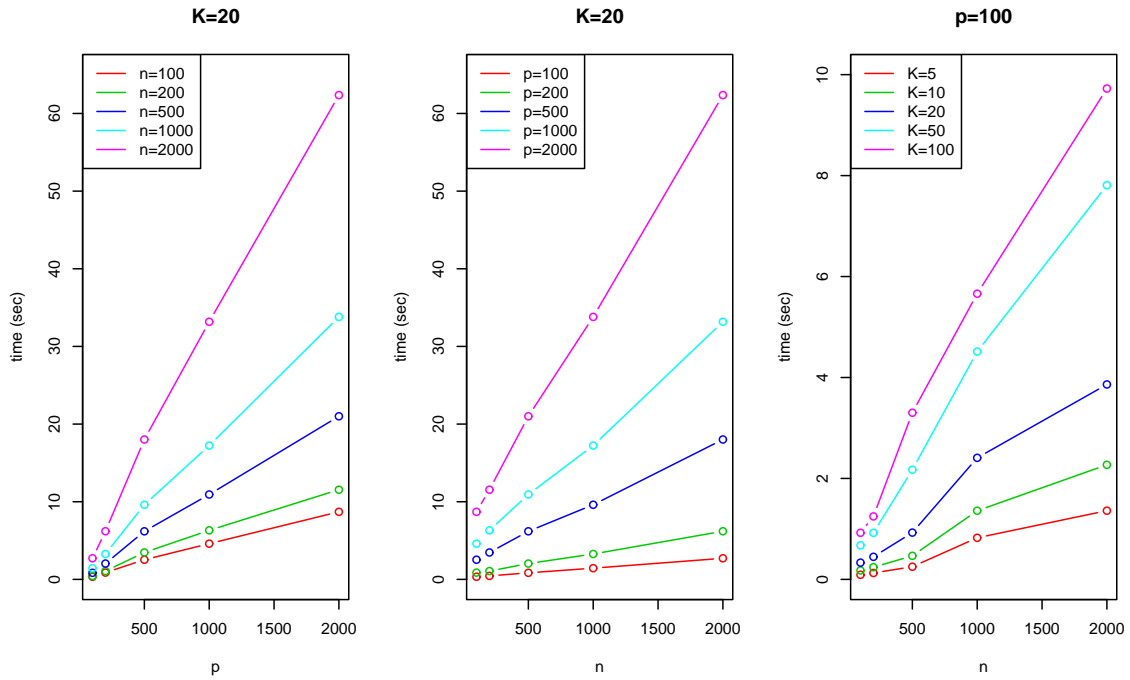


Figure 2: *Timings (sec) for the AIM procedure, for various values of n, p , and the number of terms K .*

In some cases it can be helpful to iteratively re-adjust the split points via a *backfitting procedure* post-AIM analysis or pre-process the data with method such as supervised principal components analysis pre-AIM analysis. We illustrate this in Section 6. We also note that this model is related to *boosted trees* (see e.g. Friedman et al. (2000), Friedman (2001)) in the case where the trees are stumps (single split trees). In the AIM model we further constrain all of the stumps to share the same multiplier.

2.1 Ovarian cancer data

The data for this example is taken from Fredriksson et al. (2008). It consists of 20 blood protein biomarkers in each of two groups: healthy and patients with ovarian cancer. There are 20 patients in each group. We applied the AIM procedure with a maximum of 10 biomarkers. Tenfold cross-validation was used to assess the model prediction, producing the curves in Figure 3. In the figure we have used two different methods for making predictions. Given the (integer) score s , the “logit” method fits a logistic regression in the training set to s and uses the resulting model to make predictions in the validation set. The “cutpoint method” finds the cutpoint c that produces the fewest errors in predicting y (or $1 - y$) as $I(s \leq c)$ in the training data, and then uses this cutpoint to make predictions in the validation set. Both methods yield error rates of about 10-15% with 2-3 markers, and perform better than nearest shrunken centroids procedure that was used in the original paper. Also shown are the error rates for standard forward stepwise logistic regression. The AIM score s has the form of $I(x_7 \leq 11.8) + I(x_{16} > 9.0) + I(x_9 \leq 12.0)$. The optimal cutpoint is $s \leq 2$, so the prediction rule classifies to the cancer group if all three biomarkers fall in their “red” regions. Figure 4 shows that the 3 markers over the training set, with red points indicating that the corresponding condition (such as “ $x_7 \leq 11.8$ ”) is satisfied. Figure 5 shows a schematic of the final model.

Figure 6 shows the result of applying CART to these data, using a minimum node size of five on the left and three on the right. The cross-validated errors for the two trees were 10% and 35% respectively, with the first being about the same as that for the AIM fit with 3 markers. The values below each node are the numbers of observations in the training set in each of the two classes. We see that in each case the CART nodes are pure or almost pure. In contrast, the AIM procedure produces a model with a score equal to 0, 1, 2, 3 or 4. The corresponding counts are (1, 0), (15, 0), (4, 3) and (0, 17).

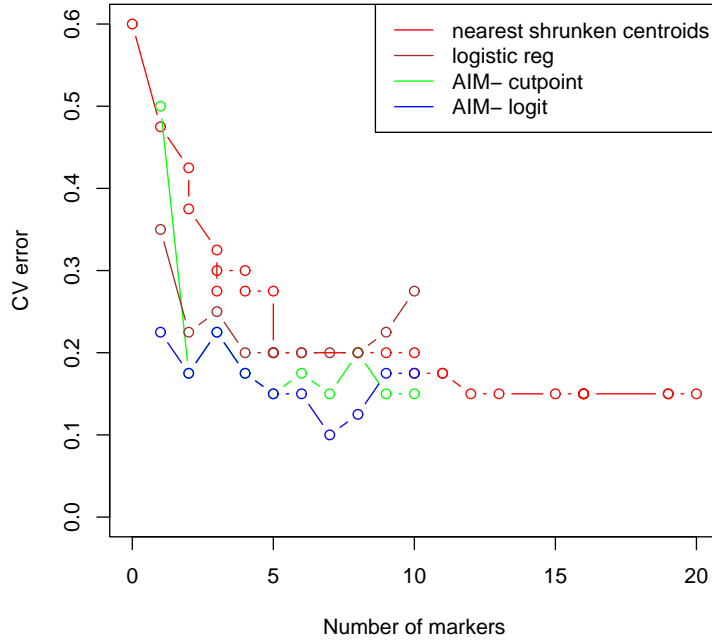


Figure 3: *Protein biomarkers data: Cross-validated error curves for AIM and nearest shrunken centroids. the standard error of each curve is about 1%.*

Hence AIM has (potentially) found an intermediate group with a score of two and approximately equal numbers of disease and non-disease patients.

3 AIM for Survival data

In the following section, we present the algorithm for scanning through all possible cutpoints for a given predictor in survival analysis. Here we have an outcome $y_i = (T_i, \delta_i)$. and predictors x_1, \dots, x_p . Our model is the proportional hazards model

$$h(t|\mathbf{x}) = h_0(t) \exp(\eta) \tag{2}$$

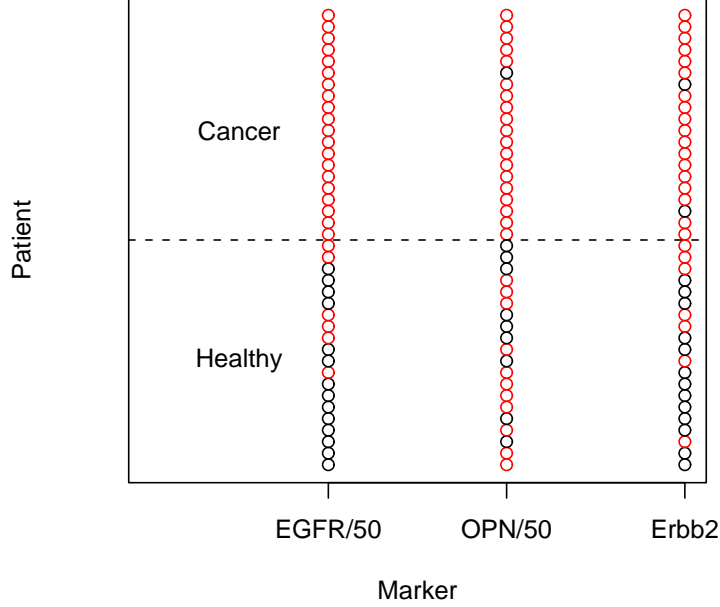


Figure 4: *Protein biomarkers data. Focussing on the model with 3 biomarkers, a red point indicates that the biomarker satisfies its corresponding split condition in that sample.*

where $h(t|\mathbf{x})$ is the hazard function and $\eta = \sum_{j=1}^K I(\tilde{x}_j^* \leq c_j)$ for $\tilde{x}_j^* \in \{\pm x_1, \dots, \pm x_p\}$.

We construct the score test as follows. Suppose that we want to decide whether to add a term $z = (x^* \leq c)$ to the existing score s . Let $w = s + z$. The score test statistics is $U/I^{1/2}$ where

$$U = \sum_{i=1}^N \delta_i \left\{ w_i - \frac{\sum_{l \in R_i} w_l}{n_i} \right\} \text{ and } I = \sum_{i=1}^N \delta_i \left[\frac{1}{n_i} \sum_{l \in R_i} w_l^2 - \left\{ \frac{1}{n_i} \sum_{l \in R_i} w_l \right\}^2 \right],$$

where R_i is risk set and n_i is the size of risk set at time T_i . Without loss of generality, we assume that $-\infty = x_0^* < x_1^* < \dots < x_N^*$. Hence if we move c

Marker 7 < 11.76			
20/26	0/14		
Marker 16 >= 8.99			
1/10	19/30		
Marker 9 < 12.03			
18/25	2/15		
Overall score			
0 0/1	1 0/15	2 3/7	3 17/17

Figure 5: *Protein biomarkers data. Schematic of the three splits and final model in bottom panel. The numbers such as 20/26 indicate that there are 20 out of 26 patients in the training set with cancer in the corresponding stratum.*

from x_{i-1}^* to x_i^* , the test statistics can be updated as

$$U \leftarrow U + \delta_i - \sum_{T_k \leq T_i} \frac{\delta_k}{n_k}$$

and

$$I \leftarrow I + \sum_{T_k \leq T_i} \frac{\delta_k(n_k - 1)}{n_k^2} + 2s_i \sum_{T_k \leq T_i} \frac{\delta_k}{n_k} - 2 \sum_{T_k \leq T_i} \sum_{T_k \leq T_j} \frac{\delta_k \{s_j + I(x_j^* \leq x_{i-1}^*)\}}{n_k}$$

with the initial values

$$U = \sum_{k=1}^N \delta_k \left\{ s_k - \frac{\sum_{l \in R_k} s_l}{n_k} \right\},$$

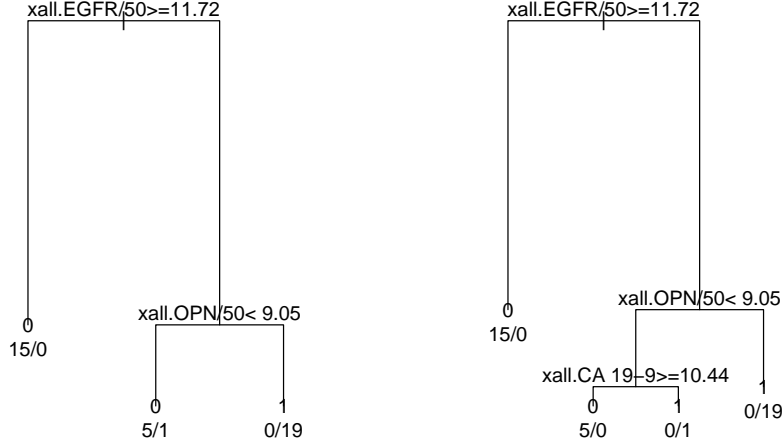


Figure 6: *Protein biomarkers data: CART trees with three terminal nodes (left) and four terminal nodes (right).*

and

$$I = \sum_{k=1}^N \delta_k \left[\frac{1}{n_k} \sum_{l \in R_k} s_l^2 - \left\{ \frac{1}{n_k} \sum_{l \in R_k} s_l \right\}^2 \right].$$

3.1 Lenz data and IPI

Here we analyze data on Non-Hodgkin’s lymphoma from Lenz et al. (2008). There are 248 patients, with the outcome being overall survival time and a large set of gene expression measurements from microarrays. The patients received either CHOP or RCHOP treatments. Later we analyze the interaction between gene expression and efficacy of the treatment. Here we explore whether the AIM procedure can re-construct the widely used international prognostic index (IPI). The details of the IPI are given in the Introduction.

We divided the data into approximately equal-sized training and test sets, and input the five predictors (age, stage, LDH, ECOG status, number of sites) into the AIM procedure. We then computed the Cox score statistic

for resulting index over the test set. This process was repeated 20 times, giving an average Cox score of 3.24(.18). The actual IPI had an average score of 3.37(.18). The cutpoints for stage were > 1 versus 1, 19 out of 20 times, as opposed to the standard definition of (1,2) versus (3,4). The numbers of extranodal sites split as > 1 , > 2 and > 3 are 11, 6 and 3 times respectively. ECOG split as 0,1 versus > 1 18 times out of 20. The distribution of split points for age and LDH are shown in Figure 7 with the corresponding standard IPI split points shown by the red lines. The cutpoints for age are approximately centered at the standard cutpoint of 60, but those for LDH are considerably above the standard cutpoint of 1.0.

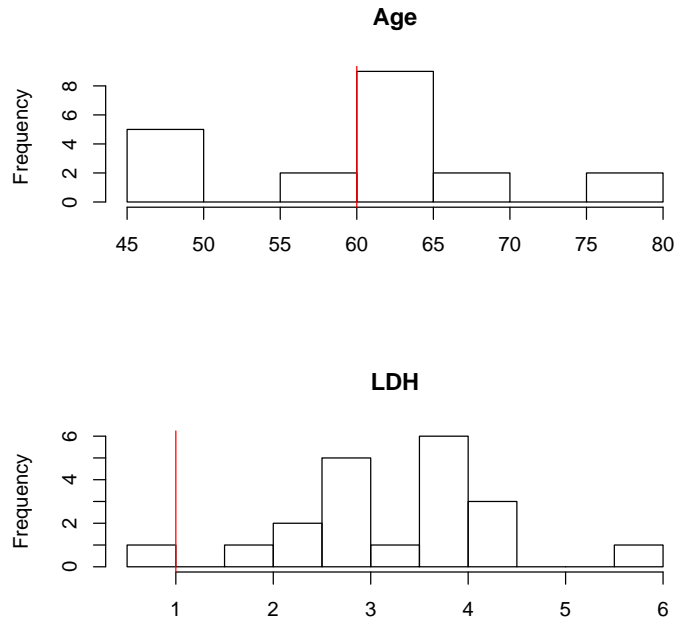


Figure 7: *Lymphoma data: distribution of split points from age and LDH over random splits of the data.*

Table 1 shows the results when we artificially add standard Gaussian noise markers (independent of the outcome) to the training and test sets. In each case we used 3-fold cross-validation to choose the model size for AIM.

Number of noise markers	Average Cox score	Proportion of markers from original five
0	3.2(.20)	1.0
2	2.9(.18)	.71
5	2.3(.13)	.50
10	2.5(.16)	.33

Table 1: *Results for artificial IPI experiment.*

We see that the procedure still maintains good performance even when a substantial number of noise markers is added.

4 Simulation study

In this section we carry out a small simulation study comparing the performance of AIM to logistic regression. We generate data in two settings, one in which the logistic regression is true and the other in which the AIM model is true. There are $n = 200$ samples and $p = 10$ predictors, all independent standard Gaussian variates. In each case only the first few predictors relate to the binary outcome. In the logistic model, $\beta = (1, 1, 2, 2, 3, 0, 0, \dots)$, and $\text{Prob}(Y = 1|\mathbf{x}) = \{1 + \exp(-\beta\mathbf{x})\}^{-1}$. In the AIM model, $s = \sum_{i=1}^3 I(x_i > 0)$ and $\text{Prob}(Y = 1|\mathbf{x}) = \{1 + \exp(3 - 2s)\}^{-1}$. Figure 8 shows the results of 10 simulations from these two models. It is a sobering reminder that both logistic regression and AIM make modeling assumptions, and can perform poorly when their underlying assumptions do not hold. In addition, in the bottom panel AIM shows a tendency to overfit after only a few terms have been added. Also included in both panels are the results when AIM is allowed to include up to 5 splits per marker. In the first setting it performs better than the vanilla AIM procedure, as it tries to approximate the linear effect with multiple binary splits.

5 Treatment interactions

In this section we show how to derive an index that explicitly models the interaction between a set of markers and a binary treatment variable. We will present the algorithm for scanning the cutoff points under linear, logistic

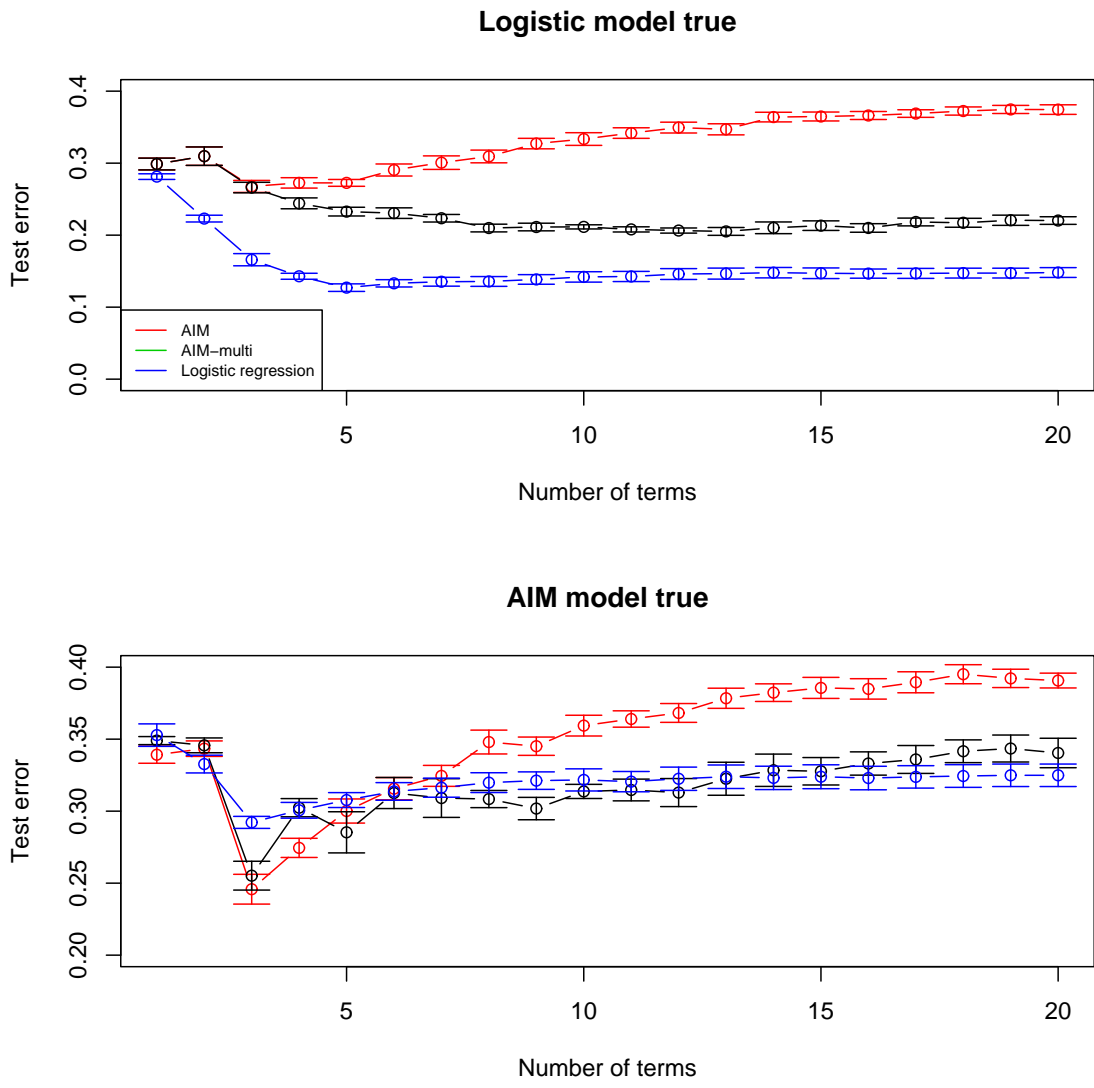


Figure 8: *Simulated data results with $N = 200$, $p = 10$. Mean and ± 1 se of the test error, for AIM with single and multiple splits per marker, and logistic regression. In the left panel the data were generated from a logistic regression model; in the right panel, the data were generated from the AIM model.*

and Cox models. With the efficient scanning algorithms, the AIM procedure given in section 2 can be readily used to construct the “IPI”-like index for the interaction of interest. In the rest of the section, we assume that s is the current score and we want to construct the new score in the form of $w = s + I(x^* < c)$, where the predictor x^* is ordered, i.e., $-\infty = x_0^* < x_1^* < \dots < x_n^*$.

To determine the cutoff point, we may perform a score test for testing $H_0 : \gamma = 0$, under the assumptions that

$$E(Y|r, w) = \beta'z + \gamma w \times r$$

and

$$\text{Prob}(Y = 1|r, w) = \frac{e^{\beta'z + \gamma w \times r}}{1 + e^{\beta'z + \gamma w \times r}},$$

for linear and logistic models, respectively, where $z = (1, r)'$ and r is the treatment indicator. We will use logistic models with binary responses to illustrate the algorithm. The method for continuous responses is similar. The score test statistics in logistic regression is $U/I^{1/2}$, where

$$U = \sum_{i=1}^N w_i r_i (y_i - \hat{p}_i), \quad I = I_{11} - I_{21} \hat{\Sigma} I_{12},$$

$$I_{11} = \sum_{i=1}^N w_i^2 r_i \hat{p}_i (1 - \hat{p}_i),$$

$$I_{21} = I'_{12} = \sum_{i=1}^N w_i r_i \begin{pmatrix} 1 \\ r_i \end{pmatrix} \hat{p}_i (1 - \hat{p}_i) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \sum_{i=1}^N w_i r_i \hat{p}_i (1 - \hat{p}_i),$$

$$\hat{\Sigma} = \left\{ \sum_{i=1}^N z_i z_i' \hat{p}_i (1 - \hat{p}_i) \right\}^{-1},$$

and \hat{p}_i is the empirical mean of responses given the treatment r_i . Let $\hat{\sigma}$ be the sum of all the entries in the two by two matrix $\hat{\Sigma}$. I can be simplified as

$$\sum_{i=1}^N w_i^2 r_i \hat{p}_i (1 - \hat{p}_i) - \left\{ \sum_{i=1}^N w_i r_i \hat{p}_i (1 - \hat{p}_i) \right\}^2 \hat{\sigma}.$$

When the cutoff point c moves from x_{i-1}^* to x_i^* , we have

$$U \leftarrow U + r_i (y_i - \hat{p}_i)$$

and

$$I \leftarrow I_1 - I_2^2 \hat{\sigma},$$

where

$$I_1 \leftarrow I_1 + (2s_i + 1)r_i \hat{p}_i(1 - \hat{p}_i)$$

and

$$I_2 \leftarrow I_2 + r_i \hat{p}_i(1 - \hat{p}_i).$$

For the survival responses, we have $\{(T_i, \delta_i, r_i, w_i), i = 1, \dots, N\}$. We consider the score test under the Cox model

$$h(t|r, w) = h_0(t)e^{\beta r + \gamma w \times r}.$$

The test statistics is $U/I^{1/2}$, where

$$\begin{aligned} U &= \sum_{i=1}^N \delta_i \left\{ w_i r_i - \frac{\sum_{l \in R_i} e^{\hat{\beta} r_l} w_l r_l}{\sum_{l \in R_i} e^{\hat{\beta} r_l}} \right\}, \\ I &= \sum_{i=1}^N \delta_i \left[\frac{\sum_{l \in R_i} e^{\hat{\beta} r_l} w_l^2 r_l^2}{\sum_{l \in R_i} e^{\hat{\beta} r_l}} - \left\{ \frac{\sum_{l \in R_i} e^{\hat{\beta} r_l} w_l r_l}{\sum_{l \in R_i} e^{\hat{\beta} r_l}} \right\}^2 \right] \\ &\quad - \left(\sum_{i=1}^N \delta_i \left[\frac{\sum_{l \in R_i} e^{\hat{\beta} r_l} w_l r_l^2}{\sum_{l \in R_i} e^{\hat{\beta} r_l}} - \frac{\sum_{l \in R_i} e^{\hat{\beta} r_l} w_l r_l \sum_{l \in R_i} e^{\hat{\beta} r_l} r_l}{\{\sum_{l \in R_i} e^{\hat{\beta} r_l}\}^2} \right] \right)^2 V_0, \\ V_0 &= \left(\sum_{i=1}^N \delta_i \left[\frac{\sum_{l \in R_i} e^{\hat{\beta} r_l} r_l^2}{\sum_{l \in R_i} e^{\hat{\beta} r_l}} - \left\{ \frac{\sum_{l \in R_i} e^{\hat{\beta} r_l} r_l}{\sum_{l \in R_i} e^{\hat{\beta} r_l}} \right\}^2 \right] \right)^{-1}, \end{aligned}$$

and $\hat{\beta}$ is the maximum partial likelihood estimator for β_0 under the null model: $h(t|r, w) = h_0(t)e^{\beta_0 r}$. We introduce the following notations to present the algorithm for scanning all cut-off points:

$$I_1 = \sum_{k=1}^N \delta_k \frac{\sum_{l \in R_k} e^{\hat{\beta} r_l} w_l^2 r_l^2}{\sum_{l \in R_k} e^{\hat{\beta} r_l}} \quad I_2 = \sum_{k=1}^N \delta_k \left\{ \frac{\sum_{l \in R_k} e^{\hat{\beta} r_l} w_l r_l}{\sum_{l \in R_k} e^{\hat{\beta} r_l}} \right\}^2, \quad I_3 = \sum_{k=1}^N \delta_k \frac{\sum_{l \in R_k} e^{\hat{\beta} r_l} w_l r_l^2}{\sum_{l \in R_k} e^{\hat{\beta} r_l}}$$

and

$$I_4 = \sum_{k=1}^N \delta_k \frac{\sum_{l \in R_k} e^{\hat{\beta} r_l} w_l r_l \sum_{l \in R_k} e^{\hat{\beta} r_l} r_l}{\{\sum_{l \in R_k} e^{\hat{\beta} r_l}\}^2}.$$

Thus when c changes from x_{i-1}^* to x_i^* ,

$$U \leftarrow U + \delta_i r_i - e^{\hat{\beta} r_i} r_i \sum_{T_k \leq T_i} \frac{\delta_k}{\sum_{l \in R_k} e^{\hat{\beta} r_l}} \quad \text{and} \quad I \leftarrow (I_1 - I_2) - (I_3 - I_4)^2 V_0,$$

where

$$\begin{aligned} I_1 &\leftarrow I_1 + (2s_i + 1) r_i^2 e^{\beta r_i} \sum_{T_k \leq T_i} \frac{\delta_k}{\sum_{l \in R_k} e^{\hat{\beta} r_l}}, \\ I_2 &\leftarrow I_2 + 2r_i e^{\beta r_i} \sum_{T_k \leq T_i} \sum_{T_k \leq T_j} \frac{\delta_k e^{\hat{\beta} r_j} \{s_j + I(x_j^* \leq x_{i-1}^*)\} r_j}{\{\sum_{l \in R_k} e^{\hat{\beta} r_l}\}^2} + r_i^2 e^{2\beta r_i} \sum_{T_k \leq T_i} \frac{\delta_k}{\{\sum_{l \in R_k} e^{\hat{\beta} r_l}\}^2}, \\ I_3 &\leftarrow I_3 + r_i^2 e^{\hat{\beta} r_i} \sum_{T_k \leq T_i} \frac{\delta_k}{\sum_{l \in R_k} e^{\hat{\beta} r_l}}, \quad \text{and} \quad I_4 \leftarrow I_4 + r_i e^{\hat{\beta} r_i} \sum_{T_k \leq T_i} \frac{\delta_k \sum_{l \in R_k} e^{\hat{\beta} r_l} r_l}{\{\sum_{l \in R_k} e^{\hat{\beta} r_l}\}^2}. \end{aligned}$$

5.1 Lymphoma data

We applied the AIM procedure to look for interactions between treatment (CHOP or RCHOP) and gene expression. The genes were first clustered into 149 “metagenes”. We split the 414 patients randomly into training and test sets of equal size. Using $K = 3$ markers, AIM produced a score from 0 to 4 with patient counts (0,49, 121, 4) and (2, 60, 169, 8) in the training and test sets respectively. Figure 9 shows the survival curves in the test set, stratifying by score (0,1) versus (2,3) in the middle and right panels. The procedure has identified a subset of patients that may not benefit from RCHOP treatment as compared to CHOP.

5.2 Surrogate predictors

When building a model by searching among a sizable number of predictors, there are often alternative models and predictors that fit the data nearly as well as the chosen model. More specifically, in the AIM fit, for any given marker and the corresponding splitting there may be other markers and split points that produce nearly the same risk stratification. The *surrogate markers* may be of interest to the scientist, and could also be used when applying AIM fit to data in which some of the marker values are not observed for part of the samples.

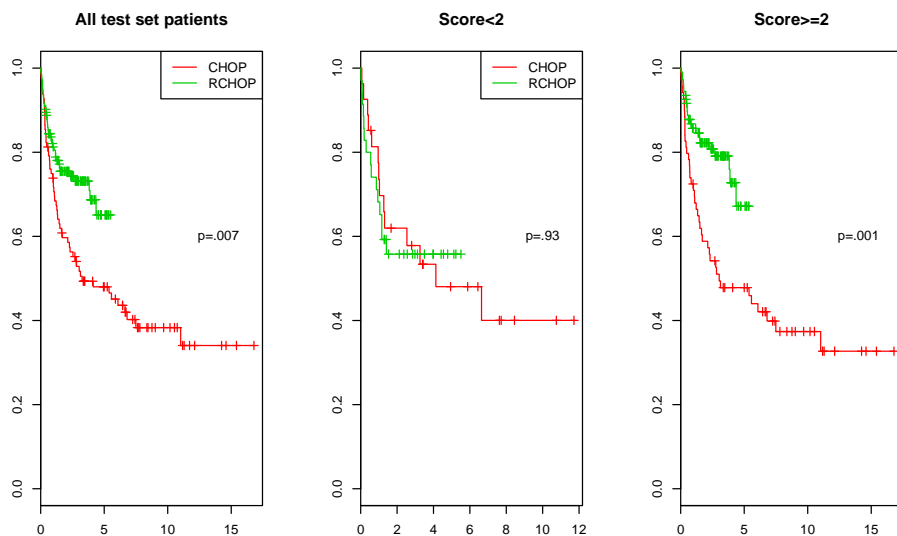


Figure 9: *Lenz data: shown the survival curves in the test set, for all patients in the left panel, and stratifying by score (0,1) versus (2,3) in the middle and right panels. The AIM interaction procedure has identified a subset of patients that may not benefit from RCHOP treatment as compared to CHOP.*

We can apply the one term logistic regression AIM procedure with the primary marker split as the outcome to find the surrogate marker. Figure 10 shows the best five surrogates for the three primary markers for the interaction model of Figure 9. The vertical axis shows the misclassification error when using the surrogate marker to predict the primary marker split. We see that each primary marker has at least one surrogate yielding an error rate of about 10%. Replacing each of the primary markers with their best surrogate markers produces a p-values of 0.01 (right panel of Figure 9). Thus the surrogates are predictive but perhaps not as strongly predictive as the primary markers.

6 Many predictors and pre-conditioning

Zhao et al. (2005) collected gene expression data on 14,814 genes from 177

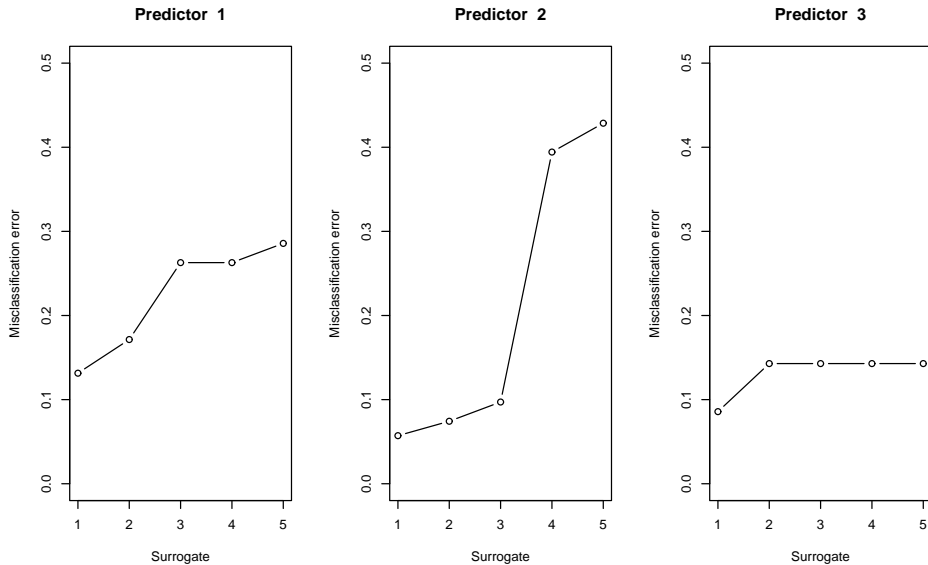


Figure 10: *Lenz data*: shown are the best five surrogates for the three primary markers for the interaction model of Figure 9. The vertical axis shows the misclassification error when using the surrogate marker to predict the primary marker split.

kidney patients. Survival times (possibly censored) were also measured for each patient,

In the original paper, the data were split into 88 samples to form the training set and the remaining 89 samples formed the test set. Here we consider 10 random splits with a size of $(88, 89)$ and report results over this 10 splits.

For computational speed, we chose the 1000 genes having largest variance across the training samples, and applied AIM to just those genes. The red curve in the left panel of Figure 11 shows the mean \pm one standard error of the test set Cox score achieved by the AIM procedure. We see that it doesn't reach 2.0, corresponding to a significance level of 0.05. This is not surprising; since there are so many predictors, the AIM procedure— like many forward stepwise methods— can avert.

As an alternative, we first computed the supervised principal component predictor \hat{y} (Bair et al. 2006) using a feature score cutoff of 1.5. Then we ap-

plied the regression version of AIM to the outcome \hat{y} . This strategy is called *pie-conditioning* (Paul et al. 2008) and can often improve the performance of forward stepwise strategies.

The green curve in the left panel of Figure 11 shows that the preconditioned AIM curve procedure achieves a much higher significance level on the test set, and with just a few predictors, it yields an index as predictive as the SPC score itself (blue) which involves hundreds of predictors.

The right panel shows the corresponding results when backfitting is done as well. The vanilla AIM procedure performs slightly better, while the preconditioned SPC version performs slightly worse.

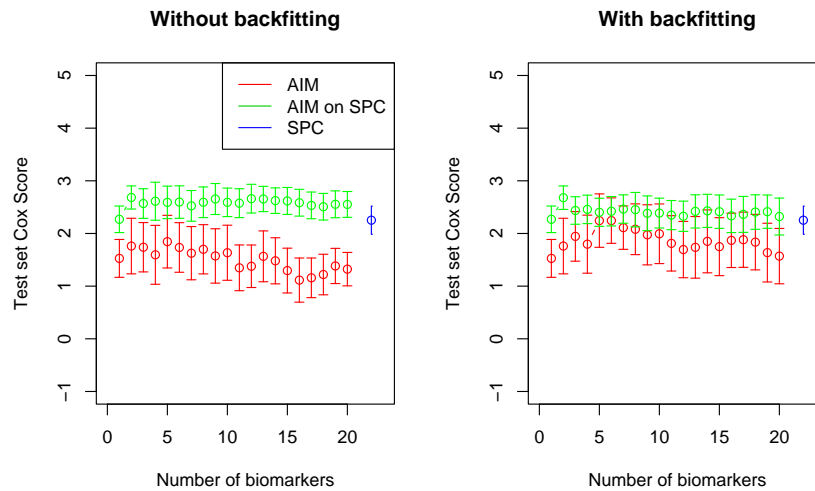


Figure 11: *Results for Kidney cancer microarray data. The left panel uses the vanilla forward stepwise procedure while the right panel does readjusts the cutpoints by backfitting after each term is added. The red curve shows the mean \pm one standard error of the test set Cox score achieved by the AIM procedure, while the green curve are the results for AIM applied to the supervised principal component predictor. The score for the supervised principal component predictor itself is shown in blue.*

7 Degrees of freedom of the AIM fit

In this section, we consider the question: how many degrees of freedom are used in fitting an AIM model? As the model is fit adaptively, this is a complicated issue. One popular notion of degrees of freedom is the expected drop in deviance compare to the null model:

$$\text{df}(\hat{\mu}_k) \equiv \text{E}[\text{dev}(\mathbf{y}, \hat{\mu}_0) - \text{dev}(\mathbf{y}, \hat{\mu}_k)] \quad (3)$$

where $\hat{\mu}_0$ is the null fit, $\hat{\mu}_k$ is the AIM fit with k terms and dev is the deviance. A discussion of this definition appears for example in Chapter 7 of Hastie et al. (2008). For a regression model with k fixed predictors including intercept, this equals to $k - 1$. For an AIM fit with k markers, however, this will greatly exceed k for a number of reasons: a) the split points are found adaptively, and b) the markers are chosen by an (adaptive) stepwise procedure. In Figure 12 we investigate this numerically. Setting $N = 100, p = 10$, we generate independent standard Gaussian predictors and binary outcomes independent of the predictors. In the left panel we show fit just x_1 with AIM, ordinary logistic regression and CART. The logistic regression averages about 1 degree of freedom as it should, while CART and AIM average about 2 and 4 degree of freedom, respectively. On the right panel we see the result for a k -term model fit by standard forward stepwise logistic regression, AIM and CART. All exceed k , as we would expect. (For CART the number of terms refers to the number of splits). The AIM procedure uses more than 3 degrees of freedom per term near the beginning of the sequence and fewer than 3 near the end. CART uses more degrees of freedom than AIM as the model grows larger

Figure 13 shows the results of the same experiment, except that each marker was generated randomly from the set of values $\{1, 2, 3, 4, 5\}$. The degrees of freedom used by AIM has decreased substantially as compared to Figure 12 (since there are fewer possible split points), but so has that for logistic regression. The latter uses only about 0.5 degrees of freedom for fitting a single predictor, something that surprised us.

To investigate the degrees of freedom of the AIM procedure theoretically, we take the approach proposed in Owen, A., (1991). The first objective is to estimate the degrees of freedom of AIM with a single continuous independent predictor $x_1 = \{x_{1i}, i = 1, \dots, N\}$ under the linear model. Recall that AIM selects the cut off value based on the random process $S_1(c) = n^{-1/2} \sum_{i=1}^N (y_i - \hat{\mu}_0)I(x_{1i} \leq c)$ indexed by c . If y and x_1 are independent, by the functional

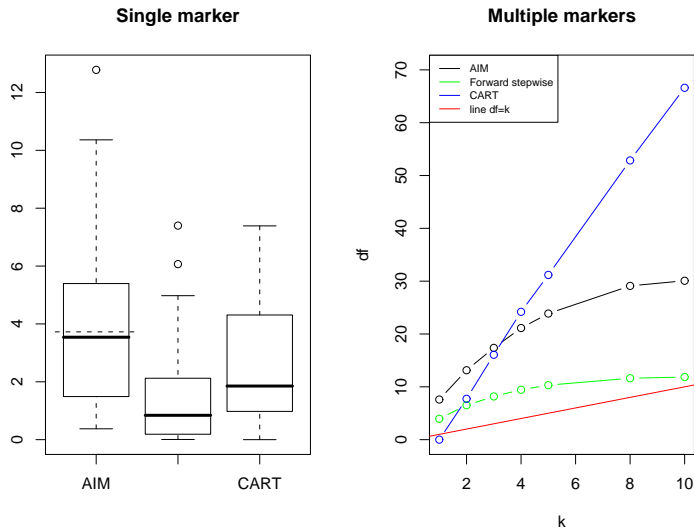


Figure 12: *Estimated degrees of freedom of the AIM, logistic regression and CART, for a single marker (left panel) and multiple markers (right panel). The marker values were standard Gaussian.*

central limit theorem, the process $S_1(c)$ converges to a scaled Brownian bridge process $\sigma_0 W\{F_1(c)\}$ in distribution over the interval $\{c : F_1(c) \in [\epsilon, 1 - \epsilon]\}$ for any positive ϵ as $n \rightarrow \infty$, where $W(\cdot)$ is the standard Brownian bridge process, $F_1(c)$ is the cumulative distribution function of the predictor and $\sigma_0^2 = \text{var}(Y)$. Since $\hat{v}_0(c)$, the consistent estimator for the variance of $S_1(c)$, converges to $\sigma_0^2 F_1(c)\{1 - F_1(c)\}$, the score test statistics evaluated at the cut off value selected by AIM is

$$\sup_{F_1(c) \in [\epsilon, 1 - \epsilon]} \frac{S_1(c)^2}{\hat{v}_0(c)} \approx M_1(\epsilon) = \sup_{t \in [\epsilon, 1 - \epsilon]} \frac{W(t)^2}{t(1 - t)},$$

for large n . It is known that $M_1(\epsilon) \rightarrow +\infty$ in probability as $\epsilon \rightarrow 0$ (Mason & Schuenemeyer 1983). However, for any fixed $\epsilon < 0.5$

$$\text{Prob}\{M_1(\epsilon) \geq m\} \approx A_\epsilon(m) = \sqrt{2m} \exp(-m/2) \log(1/\epsilon - 1) / \sqrt{\pi}$$

for m in the upper tail of the distribution (Nair 1984). In Figure 14, we plot the probability $\text{Prob}(M_1(0.1) \geq m)$, $A_{0.1}(m)$ and the survival function

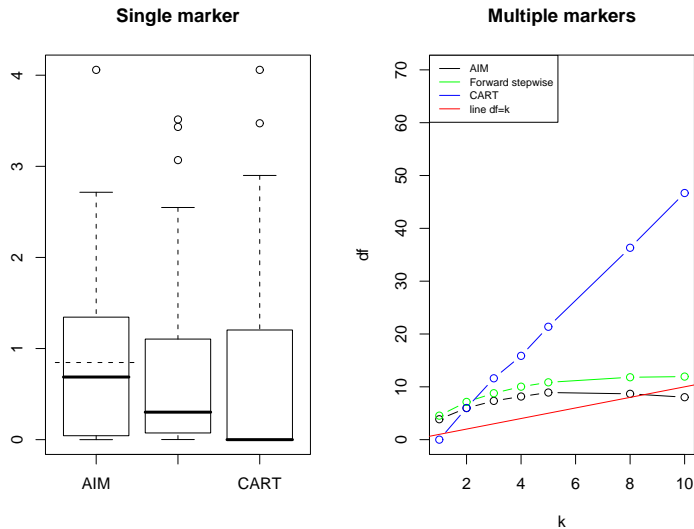


Figure 13: *Estimated degrees of freedom of the AIM and logistic regression fit, for a single marker (left panel) and multiple markers (right panel). Marker values generated from $\{1, 2, 3, 4, 5\}$.*

of χ_4^2 at the tail of the distribution of $M_1(0.1)$, where $\text{Prob}(M_1(0.1) \geq m)$ is estimated based on 50,000 Monte-Carlo simulations. It can be seen that both $A_{0.1}(m)$ and the survival function of χ_4^2 approximate $\text{Prob}(M_1(0.1) \geq m)$ very well. Furthermore, the same Monte Carlo simulation also demonstrates that the mean of $M_1(0.1)$ is 4.3, which is similar to the mean of χ_4^2 . Therefore, the score test statistics from the AIM model built from a single continuous covariate is approximately χ_4^2 and the degree of freedom of AIM model is about 4 if $\epsilon = 0.1$.

If there are p independent continuous predictors x_1, \dots, x_p unrelated to the response, AIM selects the first predictor according to

$$\sup_{F_i(c) \in [\epsilon, 1-\epsilon]} \frac{S_i(c)^2}{\hat{v}_i(c)} \approx M_i(\epsilon) = \sup_{t \in [\epsilon, 1-\epsilon]} \frac{W_i(t)^2}{t(1-t)}, i = 1, \dots, p$$

where $F_i(c)$ is the cumulative distribution function of the i -th predictor, $S_i(c) = n^{-1/2} \sum_{j=1}^N (y_j - \hat{\mu}_0) I(x_{ij} \leq c)$, and $M_i(\epsilon), i = 1, \dots, p$ are supremum values of p independent standardized Brownian bridge processes over the

interval $[\epsilon, 1 - \epsilon]$, respectively. Thus, the degree of freedom of the first model is the expectation of $\max\{M_1(\epsilon), \dots, M_p(\epsilon)\}$, which is approximately the expectation of the maximum of p independent random variables following χ_4^2 , when $\epsilon = 0.1$.

After the first step, AIM selects cut off point of a new predictor x_k by maximizing

$$\begin{aligned} & \left[\frac{n^{-1/2} \sum_{i=1}^N (y_i - \hat{\mu}_0) \{s_i + I(x_{ki} \leq c)\}}{\{\sigma_0^2 + \sigma_k^2(c)\}^{1/2}} \right]^2 \\ &= \frac{S_0^2}{\sigma_0^2} + \frac{S_k(c)^2}{\sigma_k^2(c)} + \frac{\{S_0\sigma_0^{-2} - S_k(c)\sigma_k(c)^{-2}\}^2}{\sigma_0^{-2} + \sigma_k(c)^{-2}} \end{aligned} \quad (4)$$

where s_i is the current score, σ_0^2 is the variance for $S_0 = n^{-1/2} \sum_{i=1}^n (y_i - \hat{\mu}_0)s_i$ and $\sigma_k^2(c)$ is the variance of $S_k(c) = n^{-1/2} \sum_{i=1}^N (y_i - \hat{\mu}_0)I(x_{ki} \leq c)$. This is approximately equivalent to maximizing

$$\frac{S_k(c)^2}{v_k^2(c)}$$

while restricting that $S_k(c)$ has the same sign as that of S_0 and the maximized score test statistics is approximately

$$\frac{S_0^2}{\sigma_0^2} + \sup_{F_k(c) \in [\epsilon, 1-\epsilon]} \frac{S_k(c)^2}{\sigma_k^2(c)},$$

since typically the last term in (4) is relatively small due to the fact that S_0 and $S_k(c)$ have the same sign. In other words, the degree of freedom of using the additional predictor x_k is approximately

$$\left[\sup_{F_k(c) \in [\epsilon, 1-\epsilon]} \frac{S_k(c)}{\hat{v}_k(c)} \right]^2 \approx M_k^+(\epsilon) = \left[\sup_{t \in [\epsilon, 1-\epsilon]} \frac{W_k(t)}{\sqrt{t(1-t)}} \right]^2. \quad (5)$$

For any fixed ϵ

$$\text{Prob}\{M_k^+(\epsilon) \geq m\} \approx A_\epsilon^+(m) = \sqrt{m} \exp(-m/2) \log(1/\epsilon - 1) / \sqrt{2\pi}$$

for m in the upper tail of the distribution of $M_k^+(\epsilon)$ (Nair, 1984). In Figure 15, we plot $A_{0.1}^+(m)$, the survival function of χ_3^2 and the probability

$\text{Prob}(M_k^+(0.1) \geq m)$, which is estimated based on 50,000 Monte-Carlo simulations, on the tail of the distribution of $M_1^+(0.1)$. It can be seen that both $A_{0.1}^+(m)$ and the survival function of χ_3^2 approximate $\text{Prob}(M_k^+(0.1) \geq m)$ very well. Furthermore, the same Monte Carlo simulation also demonstrated that the mean of $M_k^+(0.1)$ is 2.6, which is slightly smaller than the mean of χ_3^2 . Therefore, the increased degree of freedom of using one additional predictor x_k in AIM is approximately 3, if $\epsilon = 0.1$. Due to the sequential nature of AIM procedure, we may conclude that degrees of freedom in AIM model with 1, 2, \dots , $p - 1$ and p terms can be approximately by

$$d_1 = E\{\tilde{\xi}_{(p:p)}\}, d_2 = d_1 + E\{\tilde{\Delta}_{(p-1):p}\}, \dots, \text{ and } d_p = \sum_{i=1}^{p-1} d_i + E\{\tilde{\Delta}_{(1:p)}\}, \quad (6)$$

where $\tilde{\xi}_{p:p}$ is the p -th order statistics from p independent random variables following χ_4^2 , and $\tilde{\Delta}_{k:p}$ is the k -th order statistics of p independent random variables following χ_3^2 . Therefore each term in the AIM fit uses about 3 degrees of freedom on average, which confirms our observation in Figure (12). When p independent predictors are correlated, we expect that the degrees of freedom of AIM procedure are smaller than their counterparts when all predictors are independent. Therefore, $\{d_1, \dots, d_p\}$ in (6) provides a general conservative bound for degrees of freedom in AIM. Figure 16 plots the true degree of freedom empirically estimated from 4,000 Monte-Carlo simulations versus $\{d_1, \dots, d_p\}$. The two curves agreed fairly well especially for the initial part, which is the most important region in practice. The two curves diverge slightly when the number of terms closes to $p = 50$. It is likely due to the cumulative effect of series of approximation.

8 Discussion

We have presented a method for adaptive construction of index predictor for regression, classification and survival analysis.

There is some related work in the literature. The seminal tree-based CART methodology of Breiman et al. (1984) uses binary splits to produce a decision tree. The terminal nodes (leaves) of the tree are boxes in the feature space. Related to this is the patient rule induction method (PRIM) of Friedman & Fisher (1999) which also constructs boxes in feature space, but they are not connected by a binary tree. LeBlanc et al. (2005) and LeBlanc

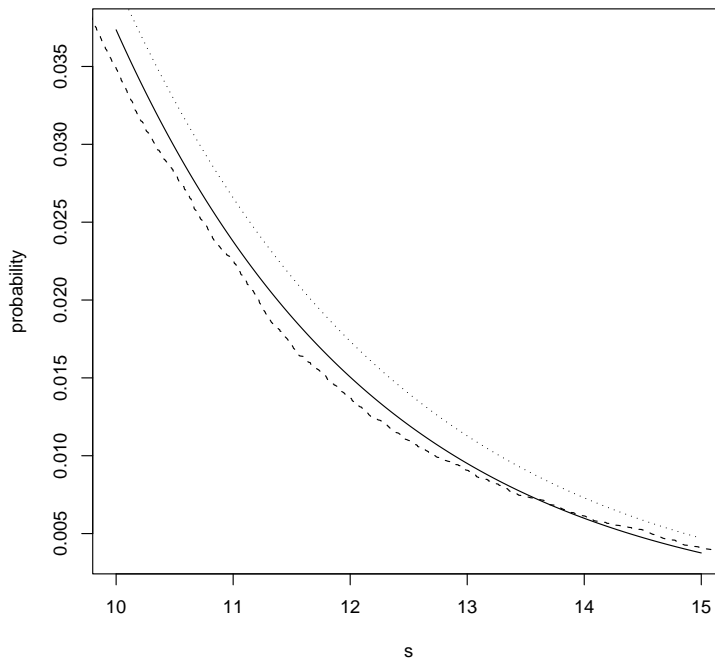


Figure 14: *Solid line: $A(m) = \sqrt{2m} \exp(-m/2) \log(1/\epsilon - 1) / \sqrt{\pi}$, dotted line: the survival function of χ_4^2 , dashed line: the empirical survival probability of $M_1(0.1)$ based on 50,000 simulations*

et al. (2002) refine these methods for survival outcomes, and adapt them for clinical trial data. Ruczinski et al. (2003) introduce “logic regression” consisting of a set of “and” and “or” rules applied to binary predictors. A simulated annealing procedure is used to estimate the rules.

The AIM methodology is different from, and simpler and less ambitious than all of these methods. Like CART and PRIM, it makes binary splits on quantitative features. But while those methods combine the rules with “and”, AIM adds them to form a single score. This makes more efficient use of the data in situations where there is a “dose-response” effect involving a set of predictors. For example, given five biomarkers x_1, x_2, \dots, x_5 , if the outcome risk is proportional $\sum_{j=1}^5 I(x_j > c_j)$, then CART or PRIM would have

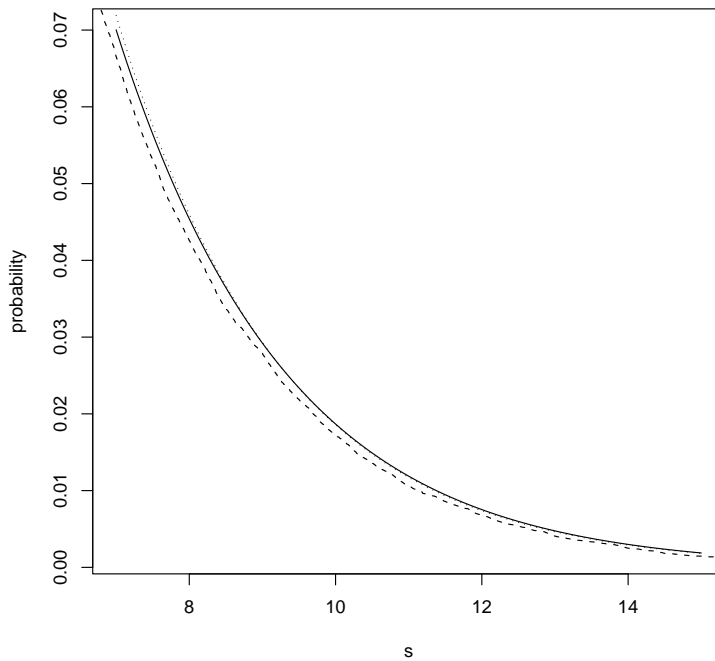


Figure 15: *Solid line: $A_\epsilon^+(m) = \sqrt{m} \exp(-m/2) \log(1/\epsilon - 1) / \sqrt{2\pi}$, dotted line: the survival function of χ_3^2 ; dashed line: the empirical survival probability of $M_k^+(0.1)$ based on 50,000 simulations*

difficulty capturing this additive structure with a small set of terminal nodes. The single score also facilitates the interpretation of the model produced by AIM fit.

R language software for AIM will be made freely available.

Acknowledgments

The second author acknowledges support from National Science Foundation Grant DMS-9971405 and National Institutes of Health Contract N01-HV-28183.

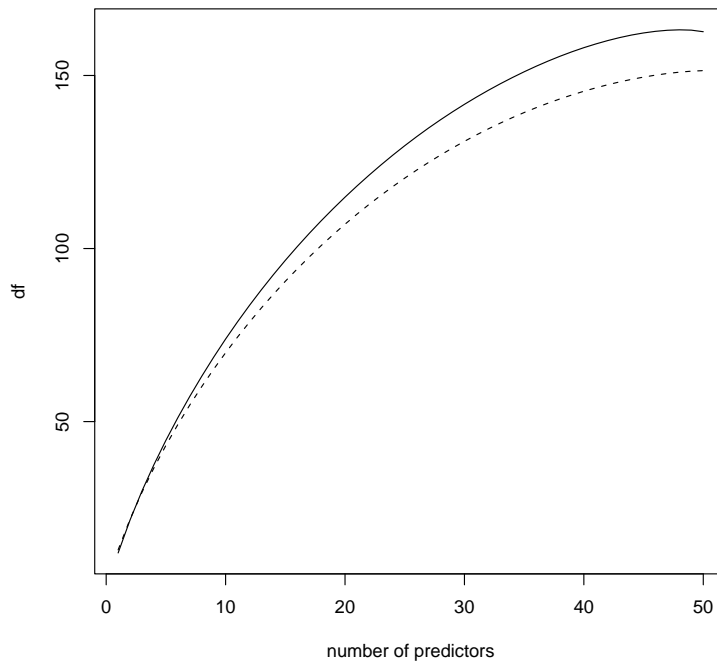


Figure 16: *Solid line: degree of freedom estimated based on 4000 Monte Carlo simulations, Dotted line: theoretically estimated degree of freedom based on (6)*

References

- Bair, E., Hastie, T., Paul, D. & Tibshirani, R. (2006), ‘Prediction by supervised principal components’, *Journal of the American Statistical Association* **101**, 119–137.
- Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984), *Classification and Regression Trees*, Wadsworth, New York.
- Fredriksson, S., Horecka, J., Brustugun, O. T., chlingemann Joerg, Koong, A. C., Tibshirani, R. & Davis, R. W. (2008), ‘Multiplexed proximity lig-

- ation assays to profile putative plasma biomarkers relevant to pancreatic and ovarian cancer’, *Clinical chemistry* **54**, 582–9.
- Friedman, J. (2001), ‘Greedy function approximation: the gradient boosting machine’, *Annals of Statistics* **29**, 1189–1232.
- Friedman, J. & Fisher, N. (1999), ‘Bump hunting in high dimensional data’, *Statistics and Computing* **9**, 123–143.
- Friedman, J., Hastie, T. & Tibshirani, R. (2000), ‘Additive logistic regression: a statistical view of boosting (with discussion)’, *Annals of Statistics* **28**, 337–307.
- Hastie, T., Tibshirani, R. & Friedman, J. (2008), *The Elements of Statistical Learning; Data Mining, Inference and Prediction (2nd edition)*, Springer Verlag, New York.
- LeBlanc, M., Jacobson, J. & Crowley, J. (2002), ‘Partitioning and peeling for constructing patient outcome groups’, *Stat. Methods Med. Res.* **11**, 1–28.
- LeBlanc, M., Moon, J. & Crowley, J. (2005), ‘Adaptive risk group refinement’, *Biometrics* **61**, 370–378.
- Lenz, G. Wright, G., Dave, S., Xiao, W., Powell, J., Zhao, H., Xu, W., Tan, B., Goldschmidt, N., Iqbal, J., Vose, J., Bast, M., Fu, K., Weisenburger, D., Greiner, T., Armitage, J., Kyle, A., May, L., Gascoyne, R., Connors, J., Troen, G., Holte, H., Kvaloy, S., Dierickx, D., Verhoef, G., Delabie, J., Smeland, E., Jares, P., Martinez, A., Lopez-Guillermo, A., Montserrat, E., Campo, E., Braziel, R., Miller, T., Rimsza, L., Cook, J., Pohlman, B., Sweetenham, J., Tubbs, R., Fisher, R., Hartmann, E., Rosenwald, A., Ott, G., Muller-Hermelink, H.-K., Wrench, D., Lister, T., Jaffe, E., Wilson, W., Chan, W., Staudt, L. & the Lymphoma/Leukemia Molecular Profiling Project (2008), ‘Stromal gene signatures in large-b-cell lymphomas’, *New England Journal of Medicine* **359**, 2313–2323.
- Mason, D. & Schuenemeyer, J. (1983), ‘A modified kolmogorov-smirnov test sensitive to tail alternative’, *Annals of Statistics* **11**, 933–946.

- Nair, V. N. (1984), ‘On the behavior of some estimators from probability plots’, *J. Amer. Statist. Assoc.*, **79**, 823–831.
- Owen, A., (1991), ‘Discussion of Multivariate Adaptive Regression Splines by Friedman’, *Annals of Statistics* **19**(1), 102–112.
- Paul, D., Bair, E., Hastie, T. & Tibshirani, R. (2008), “‘pre-conditioning” for feature selection and regression in high-dimensional problems’, *Annals of Statistics* **36**(4), 1595–1618.
- Ruczinski, I., Kooperberg, C. & LeBlanc, M. (2003), ‘Logic regression’, *Journal of Computational and Graphical Statistics* **12**, 475–511.
- TIN-HsLPP, P. (1993), ‘A predictive model for aggressive non-Hodgkin’s lymphoma. The international Non-Hodgkin’s Lymphoma Prognostic Factors Project.’, *N Engl J Med* **329**, 987–994.
- Zhao, H., Tibshirani, R. & Brooks, J. (2005), ‘Gene expression profiling predicts survival in conventional renal cell carcinoma’, *PLOS Medicine* pp. 511–533.