Group additive regression models for genomic data analysis

YIHUI LUAN
Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104-6021, USA and School of Mathematics and System Sciences, Shandong University, Jinan, Shandong 250100, People’s Republic of China

HONGZHE LI*
Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104-6021, USA
hongzhe@mail.med.upenn.edu

SUMMARY
One important problem in genomic research is to identify genomic features such as gene expression data or DNA single nucleotide polymorphisms (SNPs) that are related to clinical phenotypes. Often these genomic data can be naturally divided into biologically meaningful groups such as genes belonging to the same pathways or SNPs within genes. In this paper, we propose group additive regression models and a group gradient descent boosting procedure for identifying groups of genomic features that are related to clinical phenotypes. Our simulation results show that by dividing the variables into appropriate groups, we can obtain better identification of the group features that are related to the phenotypes. In addition, the prediction mean square errors are also smaller than the component-wise boosting procedure. We demonstrate the application of the methods to pathway-based analysis of microarray gene expression data of breast cancer. Results from analysis of a breast cancer microarray gene expression data set indicate that the pathways of metalloendopeptidases (MMPs) and MMP inhibitors, as well as cell proliferation, cell growth, and maintenance are important to breast cancer–specific survival.

Keywords: AFT models; Boosting; Gradient descent boosting; Pathway.

1. INTRODUCTION

New high-throughput technologies are generating various high-dimensional genomic data for investigating complex biological systems and phenotypes. These data also provide an opportunity for identifying pathways and genes that are related to various clinical phenotypes. One great challenge in studying the relationship between genomic data and phenotypes is to deal with the high dimensionality of such data. The “curse of dimensionality” makes most traditional statistical methods unsuitable or inefficient for analyzing...
such genomic data. However, it is important to note that although the genomic data are high dimensional, they are intrinsically low dimensional, implying that we should expect a small number of genes that are related to the phenotype of interest. In addition, many high-dimensional genomic data can be naturally grouped into small sets based on current biological knowledge. For example, when analyzing microarray gene expression data, one can group genes into functionally similar sets as in The Gene Ontology Consortium (2000) or into known biological pathways such as the Kyoto encyclopedia of genes and genomes pathways (Kanehisa and Goto, 2002). The gene expression levels of these genes can be used to characterize the activity levels of the pathways, which may in turn affect the phenotypes. When one analyzes large-scale SNP data, one can group the SNPs within the intragenic and regulatory regions of a given gene into a group and perform gene-based association analysis (Neale and Sham, 2004). The SNPs within a gene can be used to characterize the functionality of this gene. The focus of this paper is to develop group additive regression (GAR) models for identifying these groups of the genomic variables related to complex phenotypes.

As a motivating example to our proposed methods, Miller and others (2005) reported a gene expression profiling study of 251 primary breast cancer tissues resected in Uppsala County, Sweden, from January 1, 1987, to December 31, 1989, using Affymetrix Chip HG-133A and HG-133B (GEO Accession No. GSE3494). Among these patients, 236 had follow-up information in terms of time and event of disease-specific survival. Many novel approaches have been developed in recent years to analyze such data sets in order to identify genes that are related to disease-specific survival and to build predictive models for future survival (see Li, 2006, for a review of such methods). Often the genes identified are then linked to known biological pathways or functional categories to further investigate the potential pathways. Different from these commonly used methods, our proposed approach starts with a set of known pathways or gene sets to identify directly the pathways that might be related to breast cancer–related survival. In particular, in our analysis of this data set (see Section 6), we merged the Affymetrix data with the cancer-related pathways provided in the SuperArray Web site (http://superarray.com/). Our goal is to identify the pathways that are related to survival time in breast cancer patients.

In order to take into account and to utilize the group structure of the genomic data and to utilize prior biological knowledge, Wei and Li (2007) proposed a nonparametric pathway–based regression model and a modification of the general gradient descent boosting (GDBoosting) procedure (Friedman, 2001) in order to identify pathways that are related to clinical phenotypes, where they used regression trees (Breiman and others, 1984) as base learners. Although trees are very flexible in modeling potential interactions among the variables, the resulting model from GDBoosting is a linear combination of many small trees, which can be difficult to interpret in terms of variable importance. Recently, B¨uhlmann (2006) proposed a component-wise gradient descent L2 boosting (C-GDBoosting) procedure in a high-dimensional linear model setup and proved an important consistency result on prediction, allowing the predictors to grow exponentially faster than the sample sizes under the sparsity conditions. In this paper, we extend B¨uhlmann’s C-GDBoosting procedure to a group-wise L2 gradient descent boosting (G-GDBoosting) procedure for identifying groups of variables that are related to the phenotypes in the framework of GAR models using least squares or regularized least squares as base learners. Such a procedure results in GAR models with explicit expressions of the estimators and a natural way of defining the importance of a group of variables to the phenotypes. Such importance scores can then be applied to rank the importance of the groups of variables in a regression modeling framework.

The rest of the paper is organized as follows. We first introduce the GAR models. We then present the G-GDBoosting procedure for fitting the GAR models. We present simulation studies to evaluate the methods and to compare the results with the C-GDBoosting procedure. We also present results from analysis of a microarray gene expression data set of breast cancer. Finally, we present a brief discussion of the results and methods.
2. GAR MODELS

We consider the regression settings when the covariates can be naturally divided into $K$ biologically relevant groups, the activities of which might be related to the phenotype of interest. Here, the groups can be different pathways or functional sets where the variables are the gene expression levels. The groups can also be genes where the SNPs within genes are the respective variables. We allow some variables to belong to multiple groups, for example, the same genes may belong to multiple pathways. Let the vector $X_k = (X_{k1}, X_{k2}, \ldots, X_{kp_k})'$ represent the $p_k$ variables in the $k$th group. Finally, let $X = \bigcup \{X_k\}$ be the genomic data and $Y$ be the continuous response variable.

We assume that the response $Y$ is related to the genomic data $X$ through the following GAR model:

$$Y = \sum_{k=1}^{K} F_k(X_k) + \epsilon,$$

where $\epsilon$ is the noise term and $F_k(X)$ is the group effect as determined by the genomic data $X_k$ of the $k$th group. This model assumes additive effects of different groups on the response $Y$. In order to include the intercept in the model, we include 1 group with 1 as the only variable. There are several ways of specifying the function $F_k(X_k)$. One simple GAR model is to assume that $F_k(X_k)$ is modeled as a linear model,

$$F_k(X_k) = \beta_k' X_k = \sum_{l=1}^{p_k} \beta_{kl} X_{kl},$$

where $\beta_k' = \{\beta_{k1}, \ldots, \beta_{kp_k}\}$ is a vector of coefficients corresponding to the genomic data in the group $k$. If $X_k$ is the vector of gene expression data of the $p_k$ genes in the $k$th pathway, $F_k(X_k)$ can be interpreted as the pathway activity that is defined as a linear combination of the expression data as in (2.2). Combining (2.1) and (2.2) leads to an ordinary linear model. However, it should be emphasized here that the focus of our paper is to identify those groups with $\|\beta_k\|_2 \neq 0$ when $K$ is large.

Alternatively, in order to model the interactions of genomic data within a group, we can assume the following model for $F_k(X_k)$:

$$F_k(X_k) = \sum_{l=1}^{p_k} \beta_{kl} X_{kl} + \sum_{l\neq l' \geq 2} \beta_{kll'} X_{kl} X_{kl'},$$

where $\beta_{kll'}$ measures the interaction effect between 2 genomic features within the $k$th group. In this model, we allow genes to interact within a pathway but not across the pathways.

In analysis of real data sets, when $K$ is large, we should expect sparsity of the models, that is, we should expect that many of $F_k(X_k)$ in model (2.1) should be zero. In Section 3, we introduce a G-GDBOOSTing procedure for fitting the GAR models with specification of the function $F_k(X_K)$ given by (2.2) or (2.3) and for identifying the groups with $F_K(X_k) \neq 0$.

3. A G-GDBOOSTING PROCEDURE WITH LEAST SQUARES AS WEAK LEARNERS

Suppose that we have $n$ i.i.d. samples. Let $Y_i$ be the response value and $X_{i,k} = (X_{i,k1}, \ldots, X_{i,kp_k})'$ be genomic measurements in the $k$th group for the $i$th individual. Let $X_i = \{X_{i,1}, \ldots, X_{i,K}\}$ be the
genomic data measured for the \( i \)th individual on the \( K \) groups. We first assume that the sample data set \( \{Y_i, X_i,k, k = 1, \ldots, K\}_{i=1}^{n} \) follows the linear GAR model,

\[
Y_i = \sum_{k=1}^{K} \beta_k^{(m)} X_{i,k} + \epsilon_i, \quad i = 1, \ldots, n,
\]

where \( \{\epsilon_i, i = 1, \ldots, n\} \) are the noises.

Before we introduce the G-GDBoosting algorithm, we define the following notation:

\[
X_k' = (X_{1,k}, \ldots, X_{n,k}), \text{ a matrix of } p_k \text{ by } n, k = 1, \ldots, K,
\]

\[
X = (X_1, \ldots, X_K), \text{ a matrix of } n \text{ by } \sum_{k=1}^{K} p_k,
\]

\[
H_k = X_k(X_k'X_k)^{-1}X_k', \text{ a square matrix of order } n, k = 1, \ldots, K,
\]

\[
B_k = (X_k'X_k)^{-1}X_k', \text{ a matrix of } p_k \text{ by } n, k = 1, \ldots, K,
\]

\[
Y = (Y_1, \ldots, Y_n)', \text{ a } n\text{-dimensional vector.}
\]

It is easy to see that \( H_k \) is the “hat” matrix determined by the variables in the \( k \)th group. Our proposed G-GDBoosting procedure is a special case of the GDBoosting procedure of Friedman (2001) and is similar in spirit to the C-GDBoosting procedure of Bühlmann (2006). The proposed procedure involves iterative fitting of the residuals using the covariates in each of the \( K \) groups and, at each step, choosing the group that provides the best fit to the residuals as measured by the residuals sum of squares. The key of the G-GDBoosting procedure is to select a group index, \( i_m \in \{1, \ldots, K\} \) at the \( m \)th boosting step, such that the \( i_m \)th group mostly explains the current residual \( \hat{U}^{(m)} = (\hat{U}_1^{(m)}, \ldots, \hat{U}_n^{(m)})' \) by linear regression, where \( \hat{U}_i^{(m)} = Y_i - \sum_{k=1}^{K} \hat{F}_k^{(m)}(X_{i,k}) \) and \( \hat{F}_k^{(m)}(X_k) \) is the estimate of the function \( F_k(X_k) \) at the \( m \)th boosting step. More specifically, \( i_m \) is chosen such that

\[
i_m = \arg\min_{1 \leq k \leq K} \| \hat{U}^{(m)} - H_k \hat{U}^{(m)} \|_2^2,
\]

where \( \| \cdot \|_2 \) represents the conventional Euclidean L2-norm. To account for groups with different sizes, we can use the Akaike information criterion (AIC) (Akaike, 1973) to choose \( i_m \) such that

\[
i_m = \arg\min_{1 \leq k \leq K} n \log \left( \hat{\sigma}_k^{2(m)} + 1 \right) + 2(p_k + 1),
\]

where \( \hat{\sigma}_k^{2(m)} \) is the maximum likelihood estimator of the error variance in the linear regression model with \( \hat{U}^{(m)} \) as response and \( X_k \) as predictors. Details of the algorithm are given in supplementary materials available at Bioinformatics online. The key difference from Friedman’s GDBoosting procedure is that instead of using all the variables to build the base learners, we build the base learners using only the variables within the groups. At each boosting step, at most 1 new group is added to the model; the algorithm can, therefore, be used to select the relevant groups. In addition, if all the groups include only 1 variable, the algorithm becomes the C-GDBoosting procedure proposed in Bühlmann (2006).

Since the base learners are linear, similar to Bühlmann (2006), it can be easily verified that the following recursive formula holds for \( \hat{Y}^{(m)} \) and \( \hat{\beta}_k^{(m)} \) at the \( m \)th boosting step:

\[
\hat{Y}^{(m)} = \hat{A}_m Y, \quad \hat{\beta}_k^{(m)} = \hat{D}_k^{(m)} Y, \quad k = 1, \ldots, K,
\]
where \( \{ \hat{A}_m, \hat{D}_k^{(m)} \}, k = 1, \ldots, K \}_{m=0}^{M} \) are given by the following formula:

\[
\hat{A}_0 = 0, \quad \hat{A}_m = I - (I - \rho H_l) \cdots (I - \rho H_{l-1}),
\]

\[
\hat{D}_k^{(m)} = \sum_{\{l: 0 \leq l \leq m-1, \ i_1 = k \}} \rho B_{k_1}(I - \hat{A}_l), \quad k = 1, \ldots, K, \ m = 1, \ldots, M,
\]

where \( I \) is the identity matrix of order \( n \) and \( \rho \) is the learning rate (e.g. \( \rho = 0.05 \)).

### 3.1 G-GDBOOSTING with penalized least squares as base learners

The G-GDBOOSTING procedure with least squares as weak learners involves the inverse of matrix \((X_k'X_k)\). If the number of variables in some groups is larger than the sample size, or the variables within 1 group are highly correlated, \((X_k'X_k)\) can be singular or near singular, so the previous algorithm cannot be applied directly. To mediate this problem, we propose to apply a ridge regression or penalized least square regression in place of the ordinary least regressions as base learners in the proposed G-GDBOOSTING procedure. More specifically, we redefine the matrices \( H_k \) and \( B_k \) used in the G-GDBOOSTING procedure as follows:

\[
H_k^{(\lambda)} = X_k(X_k'X_k + \lambda I)^{-1}X_k', \quad \text{a square matrix of order } n, \ k = 1, \ldots, K,
\]

\[
B_k^{(\lambda)} = (X_k'X_k + \lambda I)^{-1}X_k', \quad \text{a matrix of } p_k \text{ by } n, \ k = 1, \ldots, K,
\]

where \( I \) is an identity matrix and \( \lambda \) is a tuning parameter for \( L_2 \)-penalized estimation. The G-GDBOOSTING algorithm remains the same as that presented in a previous section with \( H_k \) and \( B_k \) being replaced by \( H_k^{(\lambda)} \) and \( B_k^{(\lambda)} \).

### 3.2 Criteria for stopping the boosting iterations and selection of relevant groups

Boosting needs to stop at a suitable number of iterations to avoid overfitting. Following Bühlmann (2006), we choose the number of the boosting steps \( m \) which minimizes a corrected AIC (Hurvich and others, 1998), where the trace of the boosting hat matrix \( \hat{A}_m \) is used as the degree of freedom of the resulting estimator. In the case when the penalized least squares are used as base learners, the AIC score function depends on 2 arguments: \( m \) and \( \lambda \) in \( L_2 \)-regularization. Consequently, the estimate of the best stopping iteration step becomes \( \hat{M} = \arg \min_{1 \leq m \leq M_0, 0 \leq \lambda} \text{AIC}(m, \lambda) \).

After the model is fit, an important issue to address is to evaluate the importance of the groups selected. As noted in a previous section, at each of the boosting steps, the G-GDBOOSTING algorithm either updates the groups selected or adds a new group to the model. The groups that are selected by the algorithm should in general be important or relevant to the responses. In order to rank the selected groups based on their relative importance to the response, we present a quantitative measurement of the importance of the selected groups. For group \( k \) that was selected during the G-GDBOOSTING procedure, although the simple closed-form estimate of the coefficient can be written as \( \hat{\beta}_k^{(M)} = \hat{D}_k^{(M)}Y \), the covariance of the estimates is not easy to derive due to the fact that one has to account for the uncertainty in group selection. However, conditioning on the sequence of the groups that were selected, the covariance may be approximated as

\[
\text{cov}(\hat{\beta}_k^{(M)}) \approx \hat{\sigma}^2 \hat{D}_k^{(M)} \hat{D}_k^{(M)'}.
\]

where \( \sigma^2 \) is an estimate of the error variance. Such an approximation to the variance estimate has also been used for lasso estimates by Tibshirani (1995). Also pointed out by Tibshirani (1995), a difficulty with this covariance formula is that it gives an estimated variance of 0 for predictors that were not selected.
However, since our goal is to rank the groups that were selected, a sensible way of defining the importance of the group \( k \) among those selected is by the following importance score for the \( k \)th group:

\[
\hat{e}_k = \frac{1}{p_k} \hat{\sigma}^{-2} (\hat{\rho}_k(M)') (\hat{D}_k(M) (\hat{D}_k(M))^{-1}) \hat{\rho}_k(M),
\]

where a large value of \( \hat{e}_k \) would suggest that \( \beta_k \neq 0 \), or the \( k \)th group is associated with the response. If the matrix \( \hat{D}_k(M) (\hat{D}_k(M))' \) is singular, we modify the definition of the importance score as the following: we first write \( \hat{D}_k(M) (\hat{D}_k(M))' = U \text{diag}(\lambda_1, \ldots, \lambda_{p_k}) U' \), where \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{p_k} \) are eigenvalues of \( \hat{D}_k(M) (\hat{D}_k(M))' \) and \( U \) is an orthogonal matrix. In the singular case, assume that the rank of \( \hat{D}_k(M) (\hat{D}_k(M))' \) is \( l \) \((l < p_k)\), then we use \( U \text{diag}(1/\lambda_1, \ldots, 1/\lambda_l, 0, \ldots, 0) U' \) to replace \((\hat{D}_k(M) (\hat{D}_k(M))')^{-1} \) in the formula (3.1), where \( \text{diag}(...) \) denotes the diagonal matrix.

4. Group additive accelerated failure time models

We now consider the case when there is right censoring on some of the observations \( Y_i \). Suppose that we have a random censoring time \( C \), which is independent of the survival time \( Y \) and the covariates \( X \). Let \( C_1, \ldots, C_n \) be i.i.d. realizations of \( C \). What is observed is an event time \( t_i = \min(Y_i, C_i) \) and a censoring indicator \( \delta_i = 1 \{Y_i \leq C_i\} \), as well as the associated covariates \( X_{i,k} = (X_{i,k1}, \ldots, X_{i,kp_k})' \) in the \( k \)th group. In order to include the intercept in the model, we include 1 group with 1 as the only variable. Let \( X_i \) denote the covariate vector for the \( i \)th individual. The observed data are therefore \( \{(t_i, \delta_i, X_{i,k}, k = 1, \ldots, K) : i = 1, \ldots, n\} \). The general group additive accelerated failure time (AFT) model (Wei, 1992) can be written as

\[
g(t_i) = F(X_{i,}) + \epsilon_i = \sum_{k=1}^{K} F_k(X_{i,k}) + \epsilon_i, \quad i = 1, \ldots, n,
\]

where \( g(\cdot) \) is the known transformation function (e.g. log transformation) and \( F_k(\cdot) \) is defined as in (2.2) or (2.3). To estimate the function \( F(\cdot) \), one can define a weighted loss function by the inverse probability of censoring (van der Laan and Robins, 2002; Horthorn and others, 2006) as

\[
l(F) = \sum_{i=1}^{n} \left[ (g(t_i) - F(X_{i,}))^2 \frac{\delta_i}{\hat{S}(t_i)} \right] = \sum_{i=1}^{n} w_i (g(t_i) - F(X_{i,}))^2,
\]

where \( \hat{S}(\cdot) \) is the Kaplan–Meier estimate of the survival function for the censoring variable \( C \) and \( w_i = \delta_i / \hat{S}(t_i) \). Based on this weighted loss function (4.1), Horthorn and others (2006) presented a general GDBOosting procedure for the AFT models, which simply modifies the previous algorithm by replacing the least square fit of \( X_k \) to the current residual \( U^{(m)} \) with a weighted least square fit with weight vector \( w = (w_1, \ldots, w_n)' \). In addition, in order to obtain the closed-form estimate of \( \beta_k \) and the corrected AIC, we need to replace \( H_k \) with

\[
H_k = W X_k ((W X_k)' (W X_k))^{-1} (W X_k)',
\]

where \( W \) is the \( n \times n \) diagonal matrix with diagonal elements \( W_{ii} = \sqrt{w_i}, i = 1, \ldots, n \) (Horthorn and others, 2006).

5. Simulation studies

In this section, we present simulation studies to demonstrate the effectiveness of the proposed G-GDBOosting procedure for fitting the GAR models. In all the examples, the learning rate is fixed at
\( \rho = 0.05 \). In addition, we also compare the results with those obtained using the C-GDBoosting of Bühlmann (2006).

5.1 Description of the models for simulating the data

For all the following simulations, we simulate the elements of \( X_{i,k} \) from a uniform distribution \([-0.5, 0.5]\) and the error \( \epsilon_i \) from a normal distribution \( N(0, \sigma^2) \). We consider both low-noise variance \( \sigma^2 = 1 \) and high-noise variance \( \sigma^2 = 4 \) and repeat all the simulations 200 times. We consider the following 4 models with different degrees of complexity.

For the first model (Model 1), we assume that there is a total of 25 groups of genes, each including 4 genes. We assume that only the first 3 groups are related to the response \( Y \), and the corresponding coefficients are \( \beta_1 = (1.0, 1.2, -2.0, 3.0)' \), \( \beta_2 = (-0.5, 1.3, 1.5, 2.6)' \), and \( \beta_3 = (0.8, -1.4, -1.6, 2.7)' \). We then generated \( Y \) based on model (2.1) with corresponding functions specified as in (2.2).

The second model (Model 2) is similar to Model 1, except that each group has 10 instead of only 4 genes for a total of 250 genes. Again we assume that only the first 3 groups are related to \( Y \) and their corresponding coefficients are the same as in Model 1.

The third model (Model 3) mimics the phenotype heterogeneity, where we assume that half of the samples are generated from the GAR model with \( F(X) = \beta'_1 X_1 + \beta'_2 X_2 + \beta'_3 X_3 \) and another half of the samples are generated from the GAR with \( F(X) = \beta'_1 X_4 + \beta'_2 X_5 + \beta'_3 X_6 \), where \( \{ \beta'_k, k = 1, 2, 3 \} \) are the same as in Model 1.

Finally, for Model 4, we generate binary data \( Y \in \{-1, 1\} \) from the following model:

\[
\log \left( \frac{P(Y = 1|X)}{P(Y = -1|X)} \right) = 0.5 \sum_{j=1}^{10} X_j \left( 1 + \sum_{k=1}^{6} (-1)^k X_k \right)
\]

and assume that \( X = (X_1, \ldots, X_{100})' \). In this model, there is no natural grouping of the \( X \) variables. However, in our analysis, we divided the variables into 25 groups, each including 4 variables, \( \{X_{(k-1)4+1}, \ldots, X_{(k-1)4+4}\} \) for the \( k \)th group.

5.2 Identification of the relevant groups

In the following simulations, a sample size of 100 was used for Models 1 and 2, 200 was used for Model 3, and 300 was used for Model 4. Different sample sizes were used due to varying complexity of the 4 models. For each model, 200 replications were simulated. For the simulated data sets, AIC was used to select the number of the boosting steps with the maximum number of steps set to 1000–4000 for different models. For all the data sets, the optimal step sizes selected by AIC were smaller than these maximum numbers of steps.

Figure 1 shows the boxplots of the important scores based on the G-GDBoosting procedure for each of the gene groups based on 200 replications for high-noise variance with \( \sigma^2 = 4 \) (similar plots are observed for low-noise cases with \( \sigma^2 = 1 \)). For each of the 4 models, it is clear that the relevant gene groups have higher importance scores than irrelevant gene groups. For Model 1, it is clear that the first 3 groups have much higher importance scores than the others (see upper left plot). In fact, in the low-noise case, in all the 200 replications, the first 3 groups always have the highest importance scores. When noise variance increases to \( \sigma^2 = 4 \), the first 3 groups are simultaneously selected as the top 3 groups in 82% of the replications. Similarly, for Model 2, the first 3 groups have much higher importance scores than the others in both low- and high-noise cases (see upper right plot of Figure 1). If the top 3 groups with the highest
importance scores are selected, at least 2 and 3 out of the first 3 groups are simultaneously selected with probabilities of 100% and 99.5% in the low-noise situation and 92.5% and 47.5% in the high-noise case.

For Model 3, there are 6 gene groups or 24 genes that are related to the response. The bottom left plot of Figure 1 shows that the first 6 groups have higher importance scores than the other 19 irrelevant groups. If the top 6 groups with the largest importance scores are selected, at least 4, 5, and 6 groups out of the first 6 are simultaneously selected with probabilities of 94%, 67%, and 9% in the low-noise case and 53%, 16%, and 2% in the high-noise situation.

For Model 4, we generated 300 samples and repeated simulations 200 times. The lower right plot of Figure 1 clearly shows that the first 3 groups have higher importance scores than the others. Group 3 has smaller importance scores due to the fact that there are only 2 genes in this group that are related to the outcome. If the top 3 groups with the highest importance scores are chosen, the probabilities that 1, 2, and 3 relevant groups are simultaneously selected are 85%, 37.5%, and 3%, respectively.

As a comparison, we also examined the behavior of the importance scores when the relevant groups are not included in the analysis. As expected, no groups have shown higher scores than the others, indicating that no groups are more important to the phenotypes than the others (see supplementary materials available at Biostatistics online for details).

5.3 Comparison to C-GDBoosting

As a comparison, we also applied the C-GDBoosting procedure of Bühlmann (2006) to the simulated data sets, where the AIC criterion was used for selecting the number of the boosting steps with the maximum
Table 1. Simulation results to compare the predictive performance as measured by the empirical prediction MSEs between the C-GDBoosting (C-GDB) and the G-GDBoosting (G-GDB) methods. Mean, median, and variance of MSEs were obtained based on 100 model simulations. In each model simulation, the empirical MSE was estimated based on 500 testing samples.

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C-GDB</td>
<td>G-GDB</td>
<td>C-GDB</td>
</tr>
<tr>
<td>$\sigma^2 = 1$</td>
<td>Mean 0.70</td>
<td>0.34</td>
<td>1.49</td>
</tr>
<tr>
<td></td>
<td>Median 0.66</td>
<td>0.33</td>
<td>1.42</td>
</tr>
<tr>
<td></td>
<td>Variance 0.06</td>
<td>0.02</td>
<td>0.14</td>
</tr>
<tr>
<td>$\sigma^2 = 4$</td>
<td>Mean 2.17</td>
<td>1.09</td>
<td>4.34</td>
</tr>
<tr>
<td></td>
<td>Median 2.09</td>
<td>1.04</td>
<td>4.32</td>
</tr>
<tr>
<td></td>
<td>Variance 0.42</td>
<td>0.15</td>
<td>0.83</td>
</tr>
</tbody>
</table>

number of steps set to 1000–4000 for different models. For all the data sets, the optimal step sizes selected by AIC were smaller than these maximum numbers of steps. Since the proposed G-GDBoosting procedure aims to identify groups of variables that are related to the response and the C-GDBoosting aims to identify single variables that are related to the response, our comparisons are focused on predictive performance. For each model, we generated 2 sets of training samples of the same sample size as in Section 5.2, 1 for low-noise variance and 1 for high-noise variance. We then generated 500 new samples as the testing sets for all 3 models. Let $\hat{F}(X)$ be the estimated function in the GAR model (2.1) based on the training set; we then computed the empirical prediction mean square error (MSE) as

$$\frac{1}{m} \sum_{i=1}^{m} |F(X_i) - \hat{F}(X_i)|^2,$$

where $\{Y_i, X_i\}_{i=1}^m$ are the $m = 500$ testing samples. We repeated the whole procedure 100 times. Table 1 presents a summary of the MSEs over 100 replications for the simulated Models 1–3 and 2 noise variances using both C-GDBoosting and G-GDBoosting procedures. We observe that the G-GDBoosting procedure results in smaller MSEs and also smaller variances of the MSEs, further indicating that the G-GDBoosting method indeed provides better prediction than the C-GDBoosting procedure when the group information is utilized.

6. Application to a Breast Cancer Microarray Gene Expression Data Set

In this section, we present a breast cancer microarray gene expression data set to demonstrate the application of the proposed method in identifying pathways that are related to breast cancer survival. We analyzed a microarray gene expression data set presented in Miller and others (2005), where the gene expression levels were profiled on 251 primary breast cancer tissues resected in Uppsala County, Sweden, from January 1, 1987, to December 31, 1989, using Affymetrix Chip HG-133A and HG-133B (GEO Accession No. GSE3494). Among these patients, 236 had follow-up information in terms of time and event of disease-specific survival. In our analysis, we merged the Affymetrix data with the cancer-related pathways provided by SuperArray (http://superarray.com/) and identified 245 genes in 33 cancer-related sub-pathways (see Table 2 for the pathways and the number of genes in each pathway; note that some genes belong to multiple pathways). Our goal is to identify the pathways that are related to survival time in breast cancer patients. The numbers of genes in each pathway that are significant for different $\alpha$-levels based on the univariate Cox regression are also presented in this table. At the $\alpha$-level of 0.001, 16 genes...
Table 2. Thirty-three pathways and their descriptions considered in breast cancer data analysis. Under the column No. of genes, the numbers represent the number of genes in the SuperArray database, the number of genes with expression data, and the number of genes that are significant based on univariate Cox regression analysis for $\alpha$-levels of 0.05, 0.01, and 0.001.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>No. of Genes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18 17 4 1 1</td>
<td>Anti-apoptosis</td>
</tr>
<tr>
<td>2</td>
<td>4 3 1 0 0</td>
<td>VHL Caspase activation</td>
</tr>
<tr>
<td>3</td>
<td>3 3 1 0 0</td>
<td>DNA damage response</td>
</tr>
<tr>
<td>4</td>
<td>24 22 7 4 1</td>
<td>Factors involved in other aspects of apoptosis</td>
</tr>
<tr>
<td>5</td>
<td>8 6 3 1 0</td>
<td>Induction of apoptosis</td>
</tr>
<tr>
<td>6</td>
<td>10 9 3 1 0</td>
<td>Induction of apoptosis by signals</td>
</tr>
<tr>
<td>7</td>
<td>6 5 4 2 1</td>
<td>Regulation of apoptosis</td>
</tr>
<tr>
<td>8</td>
<td>3 2 2 2 0</td>
<td>Apoptosis others</td>
</tr>
<tr>
<td>9</td>
<td>13 13 5 1 1</td>
<td>Cell cycle arrest</td>
</tr>
<tr>
<td>10</td>
<td>4 4 1 1 1</td>
<td>Cell cycle checkpoint</td>
</tr>
<tr>
<td>11</td>
<td>29 29 10 6 2</td>
<td>Factors involved in other aspects of cell cycle</td>
</tr>
<tr>
<td>12</td>
<td>81 75 22 10 5</td>
<td>Regulation of cell cycle</td>
</tr>
<tr>
<td>13</td>
<td>6 6 1 0 0</td>
<td>Cell differentiation/cell fate determination</td>
</tr>
<tr>
<td>14</td>
<td>63 62 12 3 1</td>
<td>Cell growth and/or maintenance</td>
</tr>
<tr>
<td>15</td>
<td>41 38 13 8 1</td>
<td>Cell proliferation</td>
</tr>
<tr>
<td>16</td>
<td>11 11 4 1 0</td>
<td>Growth factors</td>
</tr>
<tr>
<td>17</td>
<td>46 45 16 5 1</td>
<td>Regulation of cell proliferation/differentiation</td>
</tr>
<tr>
<td>18</td>
<td>10 10 2 0 0</td>
<td>Cell migration and motility</td>
</tr>
<tr>
<td>19</td>
<td>2 2 0 0 0</td>
<td>Cell–cell adhesion</td>
</tr>
<tr>
<td>20</td>
<td>6 5 0 0 0</td>
<td>Cell–matrix adhesion</td>
</tr>
<tr>
<td>21</td>
<td>10 10 4 2 1</td>
<td>MMPs/MMP inhibitors</td>
</tr>
<tr>
<td>22</td>
<td>13 13 2 0 0</td>
<td>Cell surface receptor-linked signal transduction</td>
</tr>
<tr>
<td>23</td>
<td>9 9 1 0 0</td>
<td>Frizzled and frizzled-2 signaling pathways</td>
</tr>
<tr>
<td>24</td>
<td>17 17 1 0 0</td>
<td>G-protein–coupled receptor signaling pathway</td>
</tr>
<tr>
<td>25</td>
<td>2 2 0 0 0</td>
<td>Insulin receptor signaling pathway</td>
</tr>
<tr>
<td>26</td>
<td>4 4 0 0 0</td>
<td>Integrin-mediated signaling pathway</td>
</tr>
<tr>
<td>27</td>
<td>29 29 4 0 0</td>
<td>Intracellular signaling cascade</td>
</tr>
<tr>
<td>28</td>
<td>6 6 2 1 0</td>
<td>Janus Kinase–signal transducer and activator of transcription cascade</td>
</tr>
<tr>
<td>29</td>
<td>2 2 1 0 0</td>
<td>Notch signaling pathway</td>
</tr>
<tr>
<td>30</td>
<td>3 3 1 0 0</td>
<td>RAS protein signal transduction</td>
</tr>
<tr>
<td>31</td>
<td>4 4 1 1 1</td>
<td>Rho protein signal transduction</td>
</tr>
<tr>
<td>32</td>
<td>13 11 1 1 1</td>
<td>Small GTPase-mediated signal transduction</td>
</tr>
<tr>
<td>33</td>
<td>16 16 2 1 0</td>
<td>Wnt receptor signaling pathway</td>
</tr>
</tbody>
</table>

in different pathways are significantly associated with survival from breast cancer, of which 5 are in the pathway related to regulation of cell cycle.

We applied the proposed G-GDBoosting procedure with $L_2$-penalized least squares as weak learners for the AFT model. The final model was selected based on the AIC criteria for $\lambda$ ranging from 0.5 to 50 and $M$ ranging from 200 to 3000 and identified that the pathways related to metalloendopeptidases (MMPs) and MMP inhibitors, as well as regulation of cell cycle, cell growth, and maintenance are important to breast cancer–specific survival. In fact, these 3 pathways were the only pathways selected during the boosting procedure. These pathways were also identified by Wei and Li (2007) using a regression-tree-based boosting procedure. Involvement of these pathways in breast cancer progression has been reported.
in the literature. The group of proteins of MMPs are enzymes capable of degrading extracellular factors that surround a cell’s environment. MMPs can directly cleave the matrix molecules that cells reside on, process growth factors to an active form, and mediate cleavage of cell-bound proteins that are exposed on the outside of the cell. Certain normal physiological processes require the action of these proteinases; however, dysregulation of MMPs is often seen in many diseases, including breast cancer. In breast cancer and other cancers, MMP dysregulation enhances tumor blood supply and their activity is necessary for many steps involved in metastatic spread (Scorilas and others, 2001; Nakopoulou and others, 2003).

Figure 2 shows the predicted log-survival time using C-GDBoosting (Hothorn and others, 2006) and G-GDBoosting procedures for censored survival time, indicating that the G-GDBoosting procedure
provides better agreement between observed and predicted log-survival times. However, the figure also indicates that the relationship between the gene expression in these 33 cancer-related pathways and the survival time is relatively weak. To further compare these 2 procedures, we conducted bootstrap comparisons of overall mean prediction using a benchmark experiment following Hothorn and others (2006). In particular, for the learning sample, 100 bootstrap samples were drawn and the performance measures of the 2 procedures, that is, the empirical risk defined in terms of the inverse probability weights, were evaluated on the same sample of out-of-bootstrap observations. The models were fitted by both the G-GDBoosting procedure and the C-GDBoosting procedure on the bootstrap samples. The means out-of-bootstrap errors were 1.26 (SE = 0.36) for the G-GDBoosting and 1.33 (SE = 0.33) for the C-GDBoosting, and the mean of the fitted bootstrap sample errors was 0.49 (SE = 0.1) for the G-GDBoosting and 0.93 (SE = 0.22) for the C-GDBoosting. The results indicated some improvement in fitting and prediction of the G-GDBoosting procedure over the C-GDBoosting. Finally, out of the 100 bootstrap samples, there were only 5 pathways that were selected by the G-GDBoosting procedure, among which the regulation of cell cycle pathway and the pathway related to cell growth and maintenance was selected for all 100 bootstrap samples, the MMP pathway was selected 49 times, and the pathway related to cell proliferation was selected in 6 of the bootstrap samples, indicating the relative consistency of the pathways selected by the G-GDBoosting procedure.

7. Conclusions and Discussion

In this paper, we have proposed GAR models and a G-GDBoosting algorithm for identifying groups of variables that are related to the phenotypes of interest. As demonstrated in our applications to analysis of microarray gene expression data, these methods can be used for identifying groups of genes such as pathways that might be related to the phenotypes. As the large body of biological information on various aspects of the biological systems and pathways is available through databases or metadata, it is important to utilize the information in modeling genomic data, especially in identifying genes and their interactions and pathways that might be related to the phenotypes. The models proposed have a natural biological interpretation as pathway activities when gene expression data are used or genetic effects when SNPs data are used and can be applied to both continuous phenotypes and censored survival phenotypes. Different from the traditional regression analysis, the proposed methods naturally incorporate biological pathways or gene structures information. In addition, our methods consider multiple groups simultaneously. Our simulation studies indicate that when the variables can be appropriately grouped, our G-GDBoosting procedure results in smaller predictive MSEs than the C-GDBoosting.

It is worth comparing our methods with some recent work on utilizing the group structures of the data. An approach close to the proposed work is the average gene expressions for regression method by Park and others (2007) where they proposed a 2-step procedure that combines hierarchical clustering and lasso. By averaging the genes within the clusters obtained from hierarchical clustering, they define supergenes and use them to fit regression models. This is essential to treat the clusters of genes as groups and use simple averages as weak learners. However, instead of using boosting to select the groups, they use lasso (Tibshirani, 1995). Yuan and Lin (2006) recently proposed a group-lasso method in order to select features as a group among the predefined sets of variables rather than selecting a single term at a time as in the original lasso method (Tibshirani, 1995). Our G-GDBoosting procedure can be regarded as an alternative way of selecting groups of variables. Besides the applications presented in this paper, the G-GDBoosting procedure can also be applied to other problems as presented in Yuan and Lin (2006).

There are several issues that deserve further study. First, it is important to study the sensitivity of the proposed methods to the misspecification of the groups information and misspecification of the model. The first type of misspecification is that the genes included in the groups do not really belong to the
groups such as the pathways. However, this should not create a big problem since as long as the pathways include the relevant genes, they are likely to be selected by the proposed methods. Alternatively, one can use a standard variable selection procedure such as the forward selection/backward elimination to select the relevant variables in each group during each of the boosting step. Another type of misspecification is that the related genes are not included in the respected groups. The third type of misspecification is that the relevant groups are not included in the model. However, it should be noted that all types of regression analysis have such potential misspecification of the models, which can miss identifying the relevant groups. It would be interesting to compare the performance of the G-GDBoosting and the C-GDBoosting procedures under different types of group misspecification. Second, the ensemble methods have been proposed mainly for predictive purposes; however, as demonstrated by Friedman (2001) and also by our simulations, these methods can also be used for identifying groups of variables that are relevant to the phenotypes. Although the relative importance scores used in this paper seem to perform well for identifying relevant variables, much future research needs to be done to rigorously investigate the problem of defining variable importance in the setting of ensemble methods. For example, important future research should assess the statistical significance of such importance scores by deriving valid variance estimates of the estimated parameters or by using bootstrap or permutations.

In summary, we have proposed a GAR framework for identifying pathways and genes that are related to clinical phenotypes. The methods can be applied to both microarray gene expression data in the context of pathway-based or gene set–based analysis and SNP data in the context of gene-based association studies (see supplementary materials available at Biostatistics online for one such application). The methods presented in this paper are especially attractive in analysis of genome-wide association studies, where we can group the SNPs into the respective genes and genes into the respective pathways. We are currently exploring such applications. The R codes for implementing the proposed methods as well as documentation can be downloaded from the authors’ Web site (http://www.cceb.med.upenn.edu/~hli/G-GDB.htm).

ACKNOWLEDGMENTS

This research was supported by the National Institutes of Health grant ES009911 (Hongzhe Li and Yihui Luan) and the National Science Foundation of China grant 10671110 (Yihui Luan). We thank Mr Edmund Weisberg, MS at Penn Center for Clinical Epidemiology and Biostatistics, for editorial assistance and 2 reviewers for many helpful comments. Conflict of Interest: None declared.

REFERENCES


[Received August 28, 2006; revision January 17, 2007; accepted for publication April 3, 2007]