University of Pennsylvania
UPenn Biostatistics Working Papers

Year 2006 *Paper* 12

Group Additive Regression Models for Genomic Data Analysis

Yihui Luan* Hongzhe Li†

*

†

Group Additive Regression Models for Genomic Data Analysis

Yihui Luan and Hongzhe Li

Abstract

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 and gene-based genetic association analysis of type 1 diabetes. Results from analysis of two breast cancer data sets indicate that the pathways of Metalloendopeptidases (MMPs) and MMP inhibitors, as well as cell proliferation, cell growth and maintenance are important to breast cancer relapse and survival. Results from analysis of a set of nonsynonymous SNPs on chromosome 6 confirmed a few genes that are associated with type 1 diabetes.

Group Additive Regression Models for Genomic Data Analysis

Yihui Luan[1,2] **and Hongzhe Li**[1]

[1]Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104-6021, U.S.A.

[2]School of Mathematics and System Sciences, Shandong University Jinan, Shandong 250100 P. R. China

Address for correspondence:

Hongzhe Li, Ph.D.

Department of Biostatistics and Epidemiology

University of Pennsylvania School of Medicine

423 Guardian Drive-920 Blockley Hall

Philadelphia, PA 19104-6021

Telephone: (215) 573-5038; Fax: (215) 573-4865

E-mail: hli@ucceb.med.upenn.edu

Abstract

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 and gene-based genetic association analysis of type 1 diabetes. Results from analysis of two breast cancer data sets indicate that the pathways of Metalloendopeptidases (MMPs) and MMP inhibitors, as well as cell proliferation, cell growth and maintenance are important to breast cancer relapse and survival. Results from analysis of a set of nonsynonymous SNPs on chromosome 6 confirmed a few genes that are associated with type 1 diabetes.

Keywords: Linear models, Accelerated failure time models, Variable importance, Microarray, Single nucleotide polymorphisms, Boosting.

1 Introduction

New high-throughput technologies are generating various high-dimensional genomic data for investigating complex biological systems and complex 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 Gene Ontology (GO) (2000) or into known biological pathways such as the KEGG 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.

Methods from machine learning or statistical learning literature have gained great popularity in analysis of genomic data, especially microarray gene expression data. Among these, boosting (Freund, 1995; Freund and Schapire, 1996) and random forest (Breiman, 2001) are the two most successful and practical methods and have been demonstrated to perform well in building predictive models using high-dimensional genomic data (Dettling and Bühlmann, 2003; Li and Luan, 2005; Wei and Li, 2006). Friedman, Hastie and Tibshirani (2000) made an insightful connection between boosting and additive logistic regression modeling and showed that the boosting procedure is an optimization method for finding a classifier minimizing a particular exponential loss function in the framework of additive modeling. Friedman (2001) proposed a general gradient descent boosting (GDBoosting) framework which can be applied to various regression

models. Focusing on regression and L_2 loss functions, Bühlmann and Yu (2003) proposed a novel component-wise boosting procedure based on cubic smoothing splines or least squares and demonstrated its effectiveness in the presence of high-dimensional predictors. Recently, Bühlmann (2006) proposed a component-wise L_2 boosting procedure in a high-dimensional linear model setup and proved a consistent result on prediction; allowing the predictors to grow exponentially faster than the sample sizes under the sparsity conditions.

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 (2006) proposed a nonparametric pathway-based regression model and a modification of the GDBoosting procedure 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. In this paper, we propose a group 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 component-wise boosting procedure. We also present results from analysis of several real genomic data sets. Finally, we present a brief discussion of the results and methods.

2 Group Additive Regression Models

Suppose that there is a total of p genomic variables that can be divided into K groups whose activities might be related to the phenotype of interest. We allow some variables to belong to multiple groups. Here the groups can be different pathways or functional sets when the variables are the gene expression levels. The groups can also be genes where SNPs with genes are the respective variables. Let the vector $x = (x_1, x_2, \ldots, x_p)^T$ represent all the p variables, where the superscript T represents the transpose. Let the kth group have p_k variables, denoted by $\{x_{k1}, x_{k2}, \ldots, x_{kp_k}\}$, where $\{k1, k2, \ldots, kp_k\} \subset \{1, 2, \ldots, p\}$. Note that $p \leq \sum_{k=1}^{K} p_k$. Let $x_{(k)} = (x_{k1}, x_{k2}, \ldots, x_{kp_k})^T, k = 1, \ldots, K$ be the vector of the genomic data and y represent the phenotype.

We assume that the phenotype, y, is related to the genomic data x by through the following GAR model,

$$y = \sum_{k=1}^{K} F_k(x_{(k)}) + \epsilon \tag{1}$$

where ϵ is the noise term, $F_k(x_{(k)})$ is the group effect as determined by the genomic data $x_{(k)}$ of the kth group. This model assumes additive effects of different groups on the phenotype y. One simple GAR model is to assume that $F_k(x_{(k)})$ is modeled as a linear model,

$$F_k(x_{(k)}) = \sum_{k=1}^{K} \beta_k^T x_{(k)} \tag{2}$$

where β_k is a vector of coefficients corresponding to the genomic data in the group k. Alternatively, in order to model the interactions of genomic data within a group, we can assume the following model for $F_k(x_{(k)})$,

$$F_k(x_{(k)}) = \sum_{l=1}^{p_k} \beta_{kl} x_{kl} + \sum_{l \neq l'} \beta_{kll'} x_{kl} x_{kl'} \tag{3}$$

where $\beta_{kll'}$ measures the interaction effect between two genomic features within the kth group. In analysis of real data sets, when K is large, we should expect sparsity of the models, i.e., we

should expect that many of $F_k(x_{(k)})$ should be zero. The question is how to identify the groups with $F_k(x_{(k)}) \neq 0$.

3 A Group Gradient Descent Boosting Procedure with Least Square as Weak Learners

Suppose that we have n *i.i.d.* samples. Let y_i represent the phenotype, $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_p^{(i)})^T$ represent the genomic measurements for the ith individual and let $x_{(k)}^{(i)} = (x_{k\cdot 1}^{(i)}, x_{k\cdot 2}^{(i)}, \ldots, x_{k\cdot p_k}^{(i)})^T$ be the genomic measurements in the kth group for the ith individual. The sample data set $\{y_i, x^{(i)}\}_{i=1}^n$ follows the GAR model,

$$y_i = F(x^{(i)}) + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\{\epsilon_i, i = 1, \ldots, n\}$ are noises and $F(x^{(i)}) = \sum_{k=1}^K \beta_k^T x_{(k)}^{(i)}$. We first introduce the following notation:

$$\begin{aligned}
X_{(k)}^T &= \left(x_{(k)}^{(1)}, \ldots, x_{(k)}^{(n)}\right), \text{ a matrix of } p_k \text{ by } n, k = 1, \ldots, K, \\
X &= \left(X_{(1)}, \ldots, X_{(K)}\right), \text{ a matrix of } n \text{ by } \sum_{k=1}^K p_k, \\
Y &= (y_1, \ldots, y_n)^T, \text{ an } n-\text{dimensional vector}, \\
\epsilon &= (\epsilon_1, \ldots, \epsilon_n)^T, \text{ an } n-\text{dimensional vector}, \\
H_{(k)} &= X_{(k)} \left(X_{(k)}^T X_{(k)}\right)^{-1} X_{(k)}^T, \text{ a square matrix of order } n, \ k = 1, \ldots, K, \\
B_{(k)} &= \left(X_{(k)}^T X_{(k)}\right)^{-1} X_{(k)}^T, \text{ a matrix of } p_k \text{ by } n, \ k = 1, \ldots, K.
\end{aligned}$$

It is easy to see that $H_{(k)}$ and $B_{(k)}$ are projection matrices and "hat" matrices determined by the variables in the kth group. Based on these notations, we present the following group gradient descent boosting (G-GDBoosting) algorithm:

A Group Gradient Descent Boosting Algorithm

Initialization. Set $m \leftarrow 0$, and $\hat{\beta}_k^{(m)} = \hat{\beta}_k^{(0)} = 0, k = 1, \ldots, K$. Consequently,

$$\begin{aligned} \hat{F}^{(m)}(x) &= \hat{F}^{(0)}(x) = \sum_{k=1}^{K} \left(\hat{\beta}_k^{(0)}\right)^T x_{(k)} = 0, \\ \hat{Y}^{(m)} &= \hat{Y}^{(0)} = \left(\hat{F}^{(0)}(x^{(1)}), \ldots, \hat{F}^{(0)}(x^{(n)})\right)^T = 0, \\ \hat{U}^{(m)} &= \hat{U}^{(0)} = Y - \hat{Y}^{(0)} = Y. \end{aligned}$$

Step 1. Select an index, $i_m \in \{1, \ldots, K\}$, such that the i_mth group mostly explains the current residual $\hat{U}^{(m)}$ by linear regression. More specifically, i_m is chosen by the following equation,

$$i_m = \text{argmin}_{1 \le k \le K} \|\hat{U}^{(m)} - H_{(k)}\hat{U}^{(m)}\|^2,$$

where $\|\cdot\|$ represents the conventional Euclidean L_2 norm.

Step 2. Update the current estimates of coefficient vectors $\{\hat{\beta}_1^{(m)}, \ldots, \hat{\beta}_K^{(m)}\}$,

$$\hat{\beta}_k^{(m+1)} = \hat{\beta}_k^{(m)},\ k \neq i_m,\ \ \hat{\beta}_{i_m}^{(m+1)} = \hat{\beta}_{i_m}^{(m)} + \rho B_{i_m}\hat{U}^{(m)},$$

where ρ is the learning rate.

Step 3. Continue the updates,

$$\begin{aligned} \hat{F}^{(m+1)}(x) &= \hat{F}^{(m)}(x) + \rho B_{i_m}\hat{U}^{(m)}x_{(i_m)} = \sum_{k=1}^{K} \left(\hat{\beta}_k^{(m+1)}\right)^T x_{(k)}, \\ \hat{Y}^{(m+1)} &= \left(\hat{F}^{(m+1)}(x^{(1)}), \ldots, \hat{F}^{(m+1)}(x^{(n)})\right)^T, \\ \hat{U}^{(m+1)} &= Y - \hat{Y}^{(m+1)}. \end{aligned}$$

Step 4. Set $m \leftarrow m + 1$, go to **Step 1** unless $m + 1 = M$. The final estimate is

$$\hat{F}^{(M)}(x) = \sum_{k=1}^{K} \left(\hat{\beta}_k^{(M)}\right)^T x_{(k)}.$$

This algorithm is a special case of the general gradient descent boosting procedure of Friedman (2001). The key difference is in Step 1, where instead of all the variables being used to build

the base learners as in Friedman (2001), we build the base learners using only the variables within the groups and select the group that fits the residuals the best. At each boosting step, at most one 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 one variable, the algorithm becomes the component-wise boosting algorithm proposed in Bühlmann (2006). Since the base learners are linear, it can be easily verified that the following recursive formula holds for $\hat{Y}^{(m)}$ and $\hat{\beta}_k^{(m)}$ at the mth boosting step:

$$\begin{aligned} \hat{Y}^{(m)} &= \hat{A}_m Y, \\ \hat{\beta}_k^{(m)} &= \hat{D}_k^{(m)} Y, \end{aligned}$$

for $k = 1, \ldots, K,\ m = 0, \ldots, M$, where $\{\hat{A}_m, \hat{D}_k^{(m)}, k = 1, \ldots, K\}_{m=0}^{M}$ are given by the following recursive formula,

$$\begin{aligned} \hat{A}_0 &= 0,\ \ \hat{A}_m = I - (I - \rho H_{i_0}) \cdots \left(I - \rho H_{i_{m-1}}\right),\ m = 1, \ldots, M, \\ \hat{D}_k^{(0)} &= 0,\ \ k = 1, \ldots, K, \\ \hat{D}_k^{(m)} &= \hat{D}_k^{(m-1)},\ k \neq i_{m-1}, k = 1, \ldots, K, \\ \hat{D}_k^{(m)} &= \hat{D}_k^{(m-1)} + \rho B_{i_{m-1}} \left(I - \hat{A}_{m-1}\right), k = i_{m-1},\ m = 1, \ldots, M, \end{aligned}$$

where I is the identity matrix of order n. Based on this recursive formula, for a chosen indices $\{i_0, i_1, \ldots, i_{m-1}\}$, we have the following expression for the hat matrix,

$$\hat{D}_k^{(m)} = \sum_{\{l:\ 0 \leq l \leq m-1, i_l = k\}} \rho B_k \left(I - \hat{A}_l\right),\ k = 1, \ldots, K,\ m = 1, \ldots, M.$$

This hat matrix is used to define the effective degrees of freedom of the associated boosting procedure used in our AIC definition (see Section 3.2).

3.1 Group gradient descent boosting with penalized least squares as base learners

The G-GDBoosting procedure with least squares as weak learners involves the inverse of matrix $\left(X_{(k)}^T X_{(k)}\right)$. If the number of variables in some groups is larger than the sample size, or the variables within one group are highly correlated, $\left(X_{(k)}^T X_{(k)}\right)$ 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 re-define the matrices $H_{(k)}$ and $B_{(k)}$ used in the G-GDBoosting procedure as the following:

$$\begin{aligned} H_{(k)}^{(\lambda)} &= X_{(k)} \left(X_{(k)}^T X_{(k)} + \lambda I\right)^{-1} X_{(k)}^T, \text{ a square matrix of order } n,\ k = 1, \ldots, K, \\ B_{(k)}^{(\lambda)} &= \left(X_{(k)}^T X_{(k)} + \lambda I\right)^{-1} X_{(k)}^T, \text{ a matrix of } p_k \text{ by } n,\ k = 1, \ldots, K, \end{aligned}$$

where I is an identity matrix and λ 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

Boosting needs to stop at a suitable number of iterations to avoid overfitting. One approach is to use cross-validation to find the best step number m that yields the best prediction result. Alternatively, the trace of the boosting hat matrix $\hat{A}_m$ can be interpreted as the degree of freedom of the resulting estimator and a corrected AIC (Hurvich *et al.*, 1998) score function of m can be defined as

$$\begin{aligned} \text{AIC}(m) &= \log(\hat{\sigma}^2) + \frac{1 + \text{trace}(\hat{A}_m)/n}{1 - (\text{trace}(\hat{A}_m) + 2)/n}, \\ \hat{\sigma}^2 &= n^{-1} \|Y - \hat{Y}^{(m)}\|^2 = n^{-1} \|\hat{U}^{(m)}\|^2. \end{aligned}$$

Given a large positive integer M_0, an estimate of the stopping iteration step number can be chosen as

$$\hat{M} = \text{argmin}_{\{1 \leq m \leq M_0\}} \text{AIC}(m).$$

In the case when the penalized least squares are used as base learners, the AIC score function depends on two arguments: m and λ in L_2 regularization. Consequently, the estimate of the best stopping iteration step becomes

$$\hat{M} = \text{argmin}_{\{1 \leq m \leq M_0,\, 0 \leq \lambda\}} \text{AIC}(m, \lambda).$$

3.3 Selection of relevant groups

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 relevant or important to the phenotypes. We present in the following a quantitative measurement of the importance of the selected groups to the phenotype.

Based on the simple closed form estimates, $\hat{\beta}_k^{(M)} = \hat{D}_k^{(M)} Y$, the covariance of the estimates may be approximated as

$$cov(\hat{\beta}_k^{(M)}) \approx \hat{\sigma}^2 \hat{D}_k^{(M)} (\hat{D}_k^{(M)})^T,$$

where $\hat{\sigma}^2$ is an estimate of the error variance. Based on this approximate covariance estimate, a sensible way of defining the importance of the group k is by the following importance score for the kth group,

$$\hat{\tau}_k = \frac{1}{p_k} \hat{\sigma}^{-2} \left(\hat{\beta}_k^{(M)} \right)^T \left(\hat{D}_k^{(M)} (\hat{D}_k^{(M)})^T \right)^{-1} \hat{\beta}_k^{(M)}, \tag{4}$$

where a large value of $\hat{\tau}_k$ would suggest that $\beta_k \neq 0$, or the kth group is associated with the phenotype. If the matrix $\left(\hat{D}_k^{(M)} (\hat{D}_k^{(M)})^T \right)$ is singular, we modify the definition of the importance

score as the following: we first write

$$\left(\hat{D}_k^{(M)}(\hat{D}_k^{(M)})^T\right) = U\text{diag}\{\lambda_1, \ldots, \lambda_{p_k}\}U^T,$$

where $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_{p_k}$ are eigenvalues of $\left(\hat{D}_k^{(M)}(\hat{D}_k^{(M)})^T\right)$ and U is an orthogonal matrix. In the singular case, assume the rank of $\left(\hat{D}_k^{(M)}(\hat{D}_k^{(M)})^T\right)$ is $l(l < p_k)$, then we use the following matrix

$$U\text{diag}\left\{\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_l}, 0, \ldots, 0\right\}U^T,$$

to replace $\left(\hat{D}_k^{(M)}(\hat{D}_k^{(M)})^T\right)^{-1}$ in the formula (4).

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 $c_1, \cdots, c_n$ 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 $\delta_i = I\{y_i \leq c_i\}$, as well as the associated covariate x_i. The observed data are therefore $\{(t_i, \delta_i, x_i) : i = 1, \cdots, n\}$. The general accelerated failure time (AFT) model (Wei, 1992) can be written as

$$g(t_i) = F(x_i) + \epsilon_i,\ i = 1, \cdots, n$$

where $g(\cdot)$ is the known transformation function (e.g., log transformation), and $F(x_i)$ is defined as in equation (2) or (3). To estimate the function $F(x_i)$, one can define a weighted loss function by the inverse probability of censoring (Robins and Rotnitzky, 1992; van der Laan and Robins, 2002) as

$$l(F) = \sum_{i=1}^{n}\left[(g(t_i) - F(x_i))^2\frac{\delta_i}{\hat{S}(t_i)}\right] = \sum_{\{1\leq i\leq n,\ \delta_i=1\}}\left[(g(t_i) - F(x_i))^2\frac{1}{\hat{S}(t_i)}\right], \quad (5)$$

where $\hat{S}(\cdot)$ is the Kaplan-Meier estimate of the survival function for the censoring variable c. More specifically, if the different elements of set $\{t_i : \ 1 \leq i \leq n, \delta_i = 0\}$ are ordered as

$$t_{(1)} < t_{(2)} < \ldots, < t_{(q)},$$

then for each $u \geq 0$, $\hat{S}(u)$ is defined by

$$\hat{S}(u) = \prod_{\{k:\ t_{(k)} \leq u\}} \left(1 - \frac{d_k}{n_k}\right),$$

where d_k is the number of observed censoring at time $t_{(k)}$, or the cardinality of set $\{i : \ 1 \leq i \leq n, t_i = t_{(k)}, \delta_i = 0\}$, n_k is the number of individuals at risk of censoring at time $t_{(k)}$, or the cardinality of set $\{i : \ 1 \leq i \leq n, t_i \geq t_{(k)}\}$; when the set $\{k : \ t_{(k)} \leq u\}$ is empty, define $\hat{S}(u) = 1$.

Based on this weighted loss function (5), we can simply modify the previous algorithm by replacing $\hat{U}_i^{(m)}$ with

$$\hat{U}_i^{(m)} = -\frac{\partial l(F(x))}{\partial F(x)}|_{F(x)=F_m(x_i)},$$

and the least square fit of X_k to $\hat{U}^{(m)}$ with a weighted least square fit. In addition, in order to obtain the closed form estimate of β_k and the corrected AIC, we need to replace $H_{(k)}$ with

$$H_{(k)} = X_{(k)}((WX_{(k)})^T(WX_{(k)}))^{-1}(WX_{(k)})^T W,$$

where W is the $n \times n$ diagonal matrix with diagonal elements $W_{ii} = \sqrt{w_i}, i = 1, \cdots, n$ and $w_i = \delta_i / \hat{S}(t_i)$.

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 component-wise L_2 boosting of Bühlmann (2006).

5.1 Description of the models for simulating the data

For all the following simulations, we simulate the $x^{(i)}$ from a uniform distribution $[-0.5, 0.5]$, and the error ϵ_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 four models with different degrees of complexity.

Model 1: For the first model, we assume that there is a total of 25 groups of genes, each including four genes. We assume that the function in model (1) is generated based on the following GAR model,

$$F(x) = \sum_{k=1}^{25} \alpha_k^T x_{(k)},\ x_{(k)} = (x_{4(k-1)+1}, \ldots, x_{4k})^T \in \mathrm{R}^4,\ k = 1, \ldots, 25,$$

where $\{\alpha_k, k = 1, \ldots, 25\}$ have the following values

$$(\alpha_1, \alpha_2, \alpha_3) = \begin{pmatrix} 1 & -0.5 & 0.8 \\ 1.2 & 1.3 & -1.4 \\ -2 & 1.5 & -1.6 \\ 3 & 2.6 & 2.7 \end{pmatrix},\ \alpha_k = 0,\ k = 4, \ldots, 25.$$

This model implies that only the first three groups are related to the outcome y.

Model 2: The second model is similar to Model 1, except that each group has 10 instead of only 4 genes, i.e., $x = (x_1, \ldots, x_{250})^T \in \mathrm{R}^{250}$, and we assume

$$F(x) = \sum_{k=1}^{25} \beta_k^T x_{(k)},\ x_{(k)} = (x_{10(k-1)+1}, \ldots, x_{10k})^T \in \mathrm{R}^{10},\ k = 1, \ldots, 25,$$

where $\{\beta_k, k = 1, \ldots, 25\}$ have the following values,

$$(\beta_1, \beta_2, \beta_3) = \begin{pmatrix} \alpha_1 & \alpha_2 & \alpha_3 \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \end{pmatrix}, \ \beta_k = 0, \ k = 4, \ldots, 25.$$

Again, only the first three pathways are relevant. Different from Model 1, for the first three groups, we assume that only 4 out of 10 genes are relevant to the outcome y.

Model 3: This model mimics the phenotype heterogeneity, where we assume that half of the samples are generated from the following GAR model

$$F(x) = \beta_1^T x_{(1)} + \beta_2^T x_{(2)} + \beta_3^T x_{(3)}$$

and another half of the samples are generated from the following GAR model,

$$F(x) = \beta_1^T x_{(4)} + \beta_2^T x_{(5)} + \beta_3^T x_{(6)}$$

where $\{\beta_k, k = 1, 2, 3\}$ and $\{x_{(k)}, k = 1, \ldots, 25\}$ are the same as in Model 2. In this model, six groups are relevant to the phenotype y.

Model 4: 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 $x = (x_1, \ldots, x_{100})^T \in \mathrm{R}^{100}$. In our analysis, we divide the variables into 25 groups, each including four variables, $\{x_{(k-1)4+1}, \cdots, x_{(k-1)4+4}\}$ for the kth group. Different from the previous three models, Model 4 does not have a simple linear form.

5.2 Identification of the relevant groups

In the following simulations, a sample size of 100 was used for Model 1 and Model 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 four models. 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 four 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 three 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 three groups always have the highest importance scores. When noise variance increases to $\sigma^2 = 4$, the first three groups are simultaneously selected as the top three groups in 82% of the replications. Similarly, for Model 2, the first three 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 three groups with the highest importance scores are selected, at least 2 and 3 out of the first three 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 six 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 three groups have higher importance scores than the others. Group 3 has smaller importance scores due to the fact that there are only two genes in this group that are related to the outcome. If the top three 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. We report here the results for models with noise variance $\sigma^2 = 4$; however, similar results are observed for $\sigma^2 = 1$. Figure 2 shows the box plots of the group importance scores over 200 simulations for the four different simulated models considered when the gene groups in the analysis did not include the relevant groups. Clearly, no groups have shown higher scores than others, indicating that no groups are more important to the phenotypes than the others. This indicates that the importance scores can indeed be used for measuring the importance of the groups to the phenotypes.

5.3 Comparison to component-wise boosting

As a comparison, we present the variable importance scores based on the component-wise gradient descent boosting procedure of Bühlmann (2006) in Figure 3. This procedure is a special case of the G-GDBoosting procedure where each group includes only one variable. Comparing to Figure 1, we observe that signals as measured by the variable importance scores are not as clear as those obtained for groups of variables using the proposed G-GDBoosting procedure. More specifically, for Model 1, if the top 12 genes with the largest importance scores are selected, the probability that at least 6, 8, 10 and 12 of the 12 relevant genes are selected among the top 12 genes are 99.5%, 87.5%, 33% and 0% respectively for the low noise variance models and 72.5%, 13.5%, 0% and 0% respectively for the high noise variances. For Model 2, if the top 12 genes with the largest importance scores are selected, the probability that at least 6, 8, 10 and 12 of the 12 relevant genes are simultaneously selected are 96%, 61%, 4% and 0% respectively in low noise variance models and become 28%, 2%, 0% and 0% respectively in high noise variance models. For Model 3, if the top 24 genes with the largest importance scores are selected, the probabilities

that at least 10, 12 and 16 out of the 24 relevant genes are simultaneously selected are 73%, 33% and 0% respectively in low noise cases, but decrease to 7%, 3% and 0% respectively in high noise case. Finally, for Model 4, if the top 10 genes with the largest importance scores are selected, the probabilities that 4, 6 and 8 of the 10 relevant genes are simultaneously selected are 26%, 2% and 0% respectively for high noise variance models. These numbers, when compared with those based on the G-GDBoosting method, clearly indicate that the component-wise boosting method does not perform as well in selecting relevant variables. However, the chance of selecting the relevant groups of genes is higher using the proposed G-GDBoosting procedure, further demonstrating the advantage of using the group information when selecting the relevant groups of variables.

5.4 Prediction errors

In order to investigate the prediction performance of the G-GDBoosting procedure and to compare the results with the component-wise boosting, for each model, we generated two sets of training samples of the same sample size as in the previous section, one for low noise variance and one for high noise variance. We then generated 500 new samples as the testing sets for Models 1 and 2 and 1000 testing samples for Model 3. Let $\hat{F}(x)$ be the estimated function in the GAR model (1) based on the training set; we then computed the 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 testing samples. Table 1 presents the MSEs for the simulated Models 1-3 and two noise variances using both component-wise and group-wise boosting procedures. We observe that the G-GDBoosting procedure results in smaller mean MSEs and also smaller variances of the MSEs, further indicating that the G-GDBoosting method indeed provides better prediction than the component-wise GDBoosting.

6 Applications to Real Data Sets

In this section, we present applications of the proposed methods to three real data sets, including two breast cancer microarray gene expression data sets and one type 1 diabetes SNP data set.

6.1 Application to two breast cancer microarray gene expression data sets

Miller *et al.* (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). The authors identified an expression signature for p53 that can be used for predicting the mutation status, transcriptional effects, and patient survival. Among these patients, 236 of them had follow-up information in terms of time and event of disease-specific survival. The same 245 genes in 33 cancer-related sub-pathways used in the previous example (see Table 2 for the pathways and the number of genes in each pathway) were used in our analysis of this data set.

We applied the proposed group gradient descent boosting procedure with L_2 penalized least squares as weak learners for the AFT model 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 three pathways were the only pathways selected during the boosting procedure.

We also applied the method to another breast cancer gene expression data set as reported in Sotiriou *et al.* (2006) (GEO Accession No. GSE2990), including gene expression data from 189 invasive breast carcinomas. Among these 189 patients, 88 of them are from the data set of Miller *et al.* (2005), and 101 are patients from the John Radcliffe Hospital (Oxford, UK). Treating the relapse-free survival time as the outcome in the AFT model, the G-GDBoosting

procedure also identified the MMP pathway and the cell growth and maintenance pathway as the two most important pathways related to breast cancer relapse. These two pathways were the only pathways that were selected during the boosting procedure.

In summary, both data sets identified pathways related to MMPs and MMP inhibitors as well as cell growth and maintenance as important pathways that are related to cancer-specific survival. Miller's data also suggest that the pathway related to cell cycle regulation may also be related to breast cancer-specific survival. These pathways were also identified by Wei and Li (2006) 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 *et al.*, 2001; Nakopoulou *et al.*, 2003; Pellikainen *et al.*, 2004).

6.2 Application to type 1 diabetes SNP data set

We also applied the G-GDBoosting procedure to a type 1 diabetes data set reported by Clayton *et al.* (2005), where they analyzed 6322 nonsynonymous SNPs (nsSNPs) in 816 cases of type 1 diabetes and 877 population-based controls from Great Britain. The nsSNPs are those SNPs leading to an amino acid change in protein product, some are deleterious and some are neutral. Our analysis focused on the nsSNPs on chromosome 6, since there are several known type 1 diabetes-related genes and loci. On chromosome 6, we have 644 nsSNPs that belong to 286

genes. We used the additive coding for the genotypes. Figure 4 presents the relative importance scores (normalized to the maximum of the scores) of the genes along chromosome 6. We found that several genes in the HLA region are important to the development of type 1 diabetes, in which the TAP2 (transport 2, ATP binding cassette) has the highest relative importance scores. Other known T1 diabetes -related genes such as MICA, TNF and BAT2 also have relatively higher importance scores. In addition, one gene close to the IDDM15 region also shows relatively high importance scores. These results indicate that the G-boosting method can indeed identify genes known to be associated with the risk of type 1 diabetes.

7 Conclusions and Discussion

In this paper we have proposed group additive regression models and a group gradient descent boosting 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 mean square errors than the component-wise gradient descent boosting.

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 *et al.* (2006) where they proposed a two-step procedure that combines hierarchical clustering and Lasso. By averaging the genes within the clusters obtained from hierarchical clustering, they define super-genes and use them to fit regression models. This is essentially 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 group-Lars and group-Lasso methods 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 Lars (Efron *et al.*, 2004) and Lasso methods (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 these wrongly included genes should not be selected by the proposed methods. 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. Second, the ensemble methods have been proposed mainly for predictive purposes; however, as demonstrated by Breiman (2001) and 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 using bootstrap or permutations.

In summary, we have proposed a group additive regression 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. 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.

Acknowledgments

This research was supported by NIH grant ES009911, a grant from the Pennsylvania Department of Health and the National Natural Science Foundation of China grant 10441004 (YL). We thank Mr. Edmund Weisberg, MS at Penn CCEB for editorial assistance.

References

Breiman L (2001): Random forests. *Machine Learning*, 45:5-32.

Breiman L, Friedman JH, Olshen RA, Stone C (1984): *Classification and Regression Trees*. Wadsworth.

Bühlmann P and Yu B (2003): Boosting with the L_2 loss: regression and classification. *Journal of the American Statistical Association*, 98:324-339.

Bühlmann P (2006): Boosting for high-dimensional linear models. *Annals of Statistics*, 34: 559-583.

Clayton DG, Walker NM, Smyth DJ, Pask, R, Cooper JD, Maier LM, Smink, LJ, Lam AC, Ovington NR, Stevens HE, Nutland S, Howson JMM, Faham M, Moorhead M, Jones HB, Falkowski M, Hardenbol P, Willis TD and Todd JA (2005): Population structure, differential bias, and genomic control in a large scale, case-control association study. *Nature Genetics*, 37: 1243-1246.

Dettling M and Bühlmann P (2003): Boosting for tumor classification with gene expression data. *Bioinformatics* 19, 1061-1069.

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

Freund Y (1995): Boosting a weak learning algorithm by majority. *Information and Computation*, 121: 256-285.

Freund Y and Schapire R (1996): Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference*, 148-156.

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

Friedman J, Hastie T and Tibshirani R (2000): Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28: 337-407.

The Gene Ontology Consortium (2000): Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25: 25-29.

Horton T, Dettling M and Bühlmann P (2005): Ensemble methods of computational inference. pp#293. In Gentleman R, Carey VJ, Huber W, Irizarry RA and Dudoit S eds *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Springer.

Hurvich C, Simonoff J and Tsai C-L (1998): Smoothing parameter selection in non-parametric regression using an improved Akaike information criterion. *Journal of Royal Statistical Society, Series B*, 60: 271-293.

Kanehisa M and Goto S (2002): KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28: 27-30.

Li H and Luan Y (2005): Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data. *Bioinformatics*, 21: 2403-2409.

Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, Pawitan Y, Hall P, Klaar S, Liu E and Bergh J (2005): An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of National Academy of Sciences*, 102: 13550-13555.

Nakopoulou L, Tsirmpa I, Alexandrou P, Louvrou A, Ampela C, Markaki S, Davaris PS (2003): MMP-2 protein in invasive breast cancer and the impact of MMP-2/TIMP-2 phenotype on overall survival. *Breast Cancer Research Treatment*,77(2):145-55.

Neale BM and Sham PC (2004): The Future of Association Studies: Gene-Based Analysis and Replication. *American Journal of Human Genetics*, 75(3): 353-362.

Park M-Y, Hastie T and Tibshirani R (2006): Averaged gene expressions for regression. *Biostatistics*, in press.

Pellikainen JM, Ropponen KM, Kataja VV, Kellokoski JK, Eskelinen MJ, Kosma VM (2004): Expression of matrix metalloproteinase (MMP)-2 and MMP-9 in breast cancer with a special reference to activator protein-2, HER2, and prognosis. *Clinical Cancer Research*, 10(22):7621-8.

Robins J and Rotnitsky A (1992): *Recovery of information and adjustment for dependent censoring using surrogate markers.* Chapter Aids epidemiology, Methodoligical issues. Bikhauser.

Scorilas A, Karameris A, Arnogiannaki N, Ardavanis A, Bassilopoulos P, Trangas T and Talieri M (2001): Overexpression of matrix-metalloproteinase-9 in human breast cancer: a potential favourable indicator in node-negativepatients *British Journal of Cancer*, 84:1488-1496.

Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-Kains B, Desmedt C, Larsimont D, Cardoso F, Peterse H, Nuyten D, Buyse M, Van de Vijver MJ, Bergh J, Piccart M, and DelorenziM (2006): Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of National Cancer Institute*, 98: 262-272.

Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane I and Park P (2005): Discovering statistically significant pathways in expression profiling studies. *Proceedings of National Academy of Sciences*, 103: 13544-13549.

Tibshirani R (1995): Regression shrinkage and selection via the Lasso. *Journal of Royal Statistical Society B*, 58: 267-288.

van der Laan MJ, Robins JM (2002): *Unified Methods for Censored Longitudinal Data and Causality*, Springer-Verlag: New York.

Vapnik V (1998): *Statistical Learning Theory.* Wiley.

Wei LJ (1992): The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine* 11:1871-1879.

Wei Z and Li H (2006): Nonparametric pathway-based regression models for analysis of genomic data. *Biostatisics*, in press.

Yuan M and Lin Y (2006): Model selection and estimation in regression with grouped variables. *Journal of Royal Statistical Society, Series B*, 68: 49-67.

Table 1: Simulation results to compare the predictive performance as measured by the prediction mean square errors (MSEs), mean, median and variance of MSEs based on 200 testing data sets are presented for both component-wise gradient descent boosting (C-DGB) and group gradient descent boosting (G-DGB).

		Model 1		Model 2		Model 3	
		C-GDB	G-GDB	C-GDB	G-GDB	C-GDB	G-GDB
$\sigma^2 = 1$	Mean	.85	.33	1.34	.82	2.55	2.40
	Median	.42	.18	.63	.35	1.19	1.12
	Variance	1.14	.19	3.34	1.35	13.24	12.02
$\sigma^2 = 4$	Mean	1.60	.57	5.91	2.23	3.97	2.84
	Median	.72	.25	2.87	.99	1.70	1.26
	Variance	4.49	.58	65.42	8.75	34.92	17.60

Table 2: Pathways considered in breast cancer data analysis, including the numbers of genes in each pathway and a description of the pathways. The last set includes 188 genes that do not belong to a particular pathway.

Pathway ID	# of Genes	Description
1	18	Anti-apoptosis
2	4	VHLCaspase activation
3	3	DNA damage response
4	24	Factors involved in other aspects of apoptosis
5	8	Induction of apoptosis
6	10	Induction of apoptosis by signals
7	6	Regulation of apoptosis
8	3	Apoptosis others
9	13	Cell cycle arrest
10	4	Cell cycle checkpoint
11	29	Factors involved in other aspect of cell cycle
12	81	Regulation of cell cycle
13	6	Cell differentiation/ cell fate determination
14	63	Cell growth and/or maintenance
15	41	Cell proliferation
16	11	Growth factors
17	46	Regulation of cell proliferation, differentiation, growth and volume
18	10	Cell migration and motility
19	2	Cell-cell adhesion
20	6	Cell-matrix adhesion
21	10	Metalloendopeptidases (MMPs) and MMP inhibitors
22	13	Cell surface receptor-linked signal transduction
23	9	Frizzled and frizzled-2 Signaling Pathways
24	17	G-protein coupled receptor protein signaling pathway
25	2	Insulin receptor signaling pathway
26	4	integrin-mediated signaling pathway
27	29	Intracellular signaling cascade
28	6	JAK-STAT cascade
29	2	Notch signaling pathway
30	3	RAS protein signal transduction
31	4	Rho protein signal transduction
32	13	Small GTPase mediated signal transduction
33	16	Wnt receptor signaling pathway
34	188	Other cancer-related genes

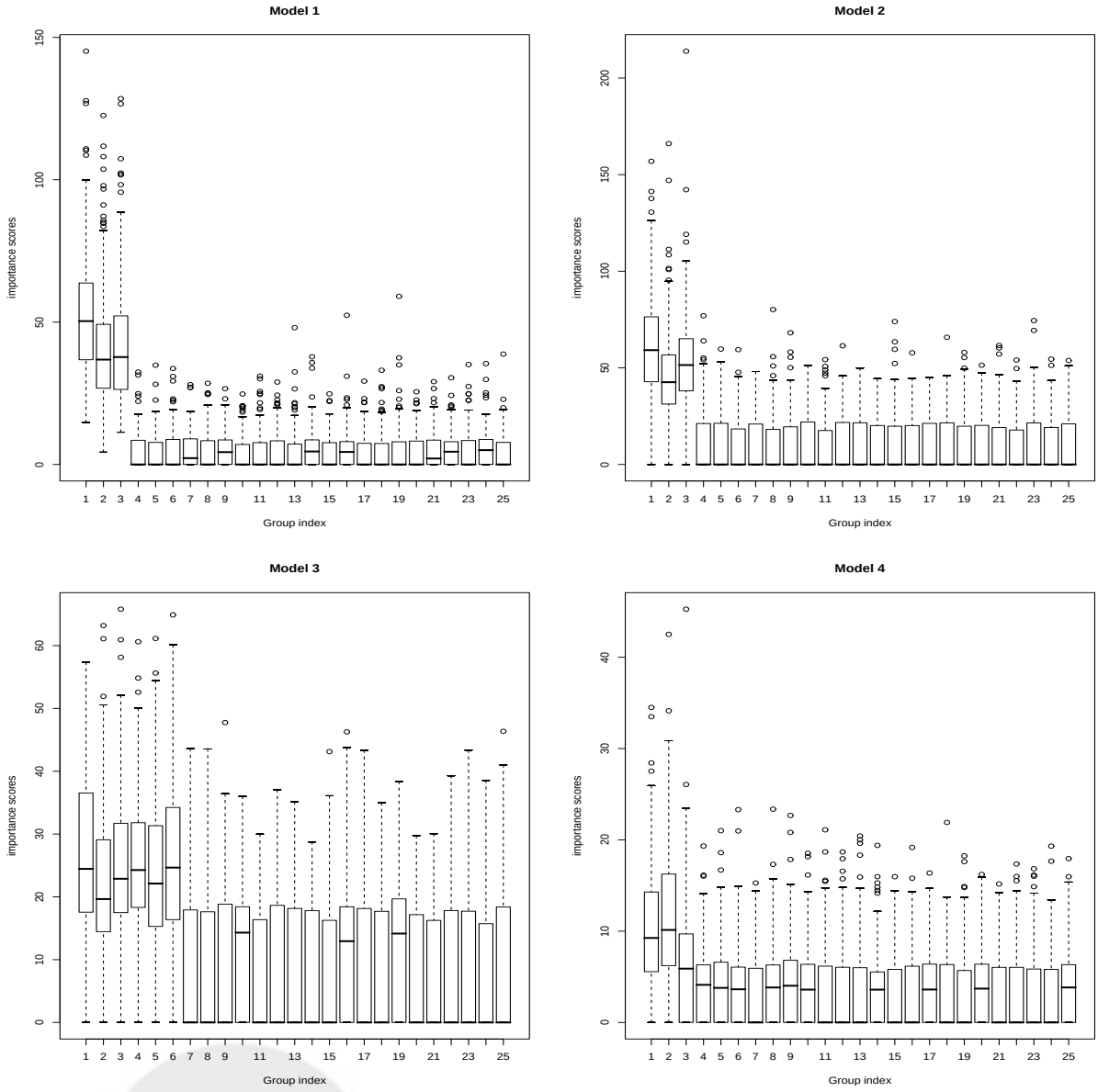

Figure 1: *Simulation results for models 1-4 with $\sigma^2 = 4$: the boxplots of the variable importance scores over 200 replications for each pathway based on the G-GDBoosting procedure. Top panel: Models 1 and 2; bottom panel: Models 3 and 4.*

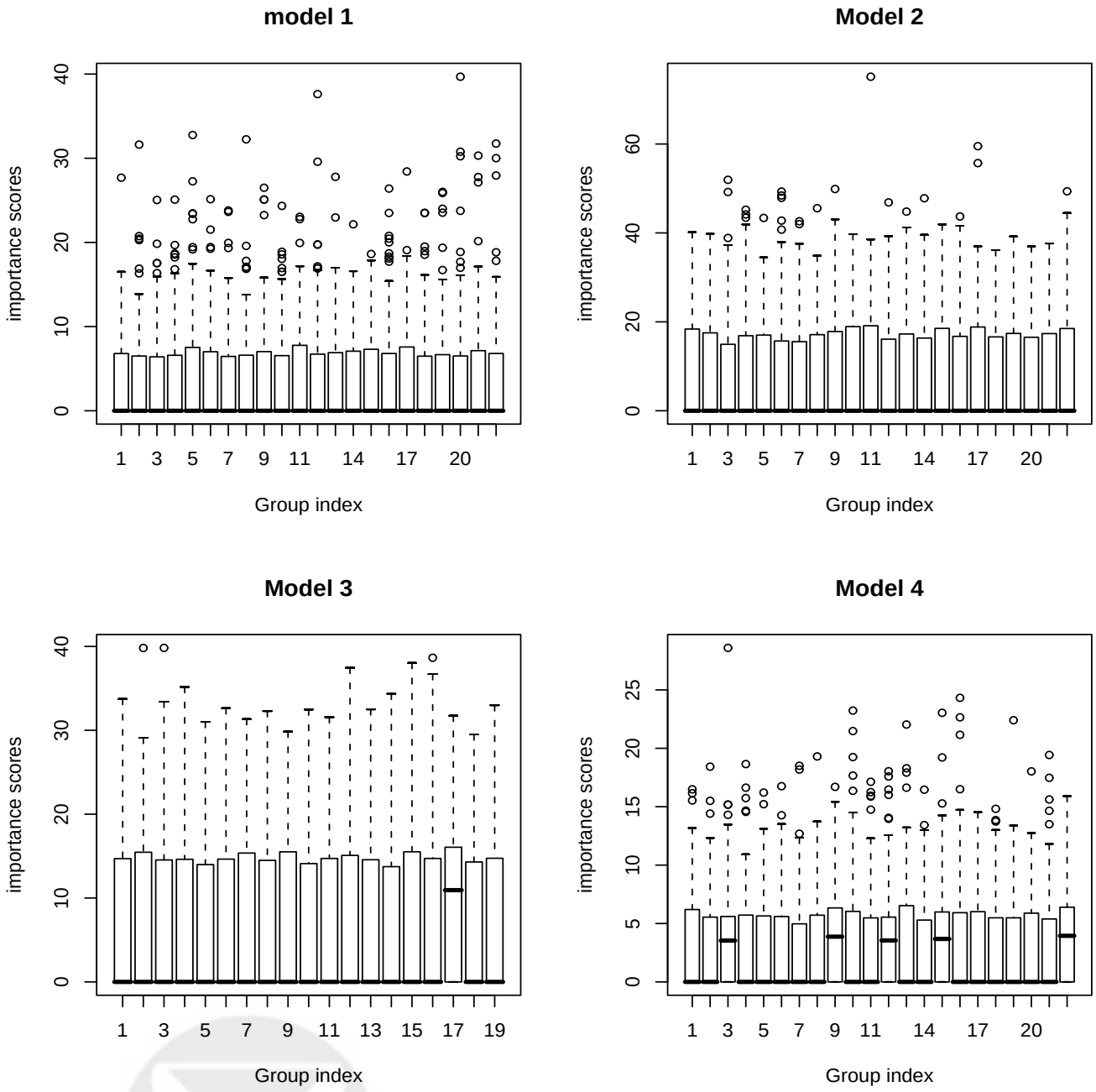

Figure 2: *Simulation results for Models 1-4 with* $\sigma^2 = 4$*: the boxplots of the variable importance scores over 200 replications for each pathway based on the G-GDBoosting procedure when the relevant pathways are not included in the analysis. Top panel: Models 1 and 2; bottom panel: Models 3 and 4.*

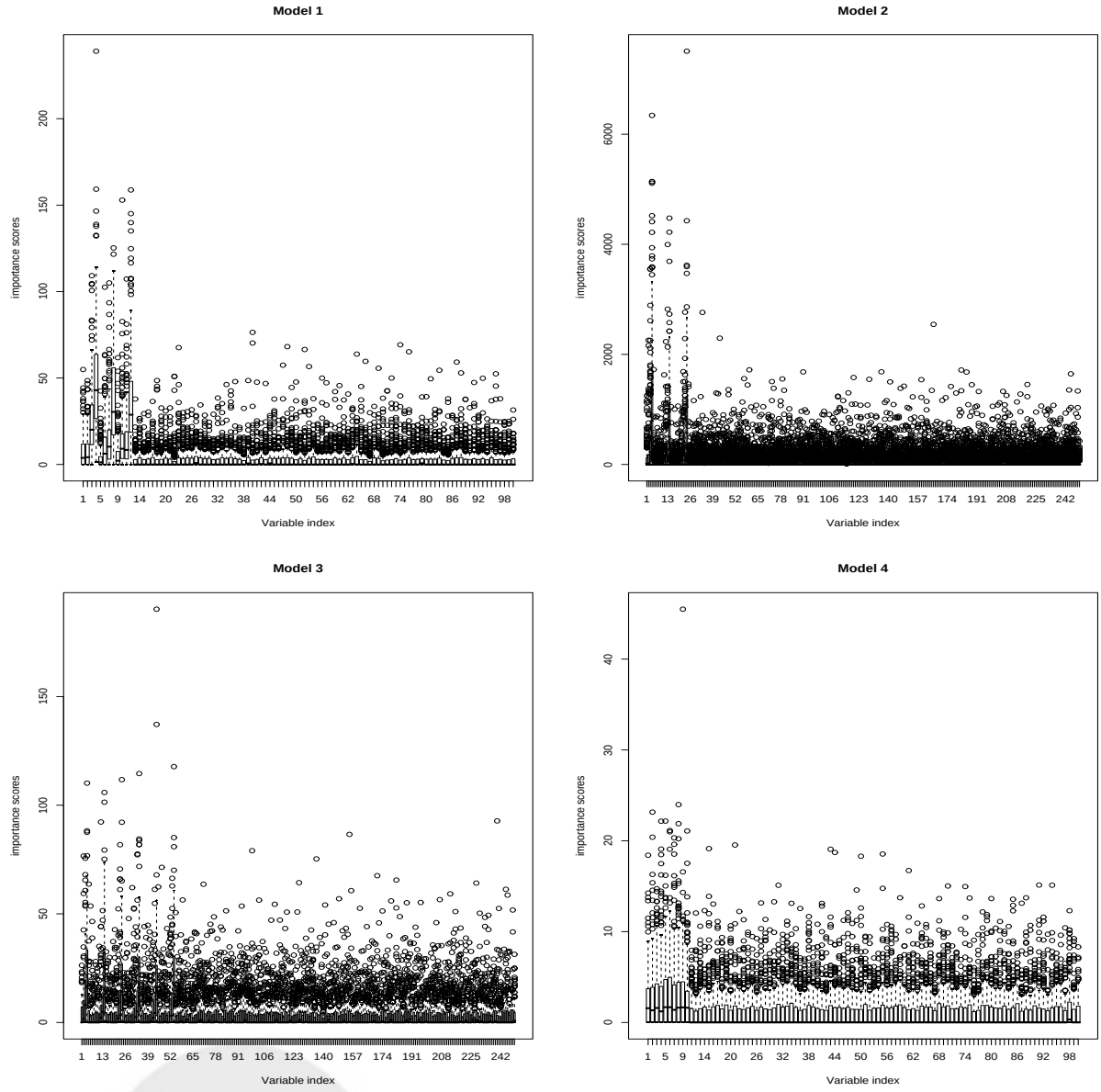

Figure 3: *Simulation results for Models 1-4 with* $\sigma^2 = 4$*: the boxplots of the variable importance scores over 200 replications for each variable based on the component-wise gradient descent boosting procedure. Top panels: Models 1 and 2; bottom panel: Models 3 and 4.*

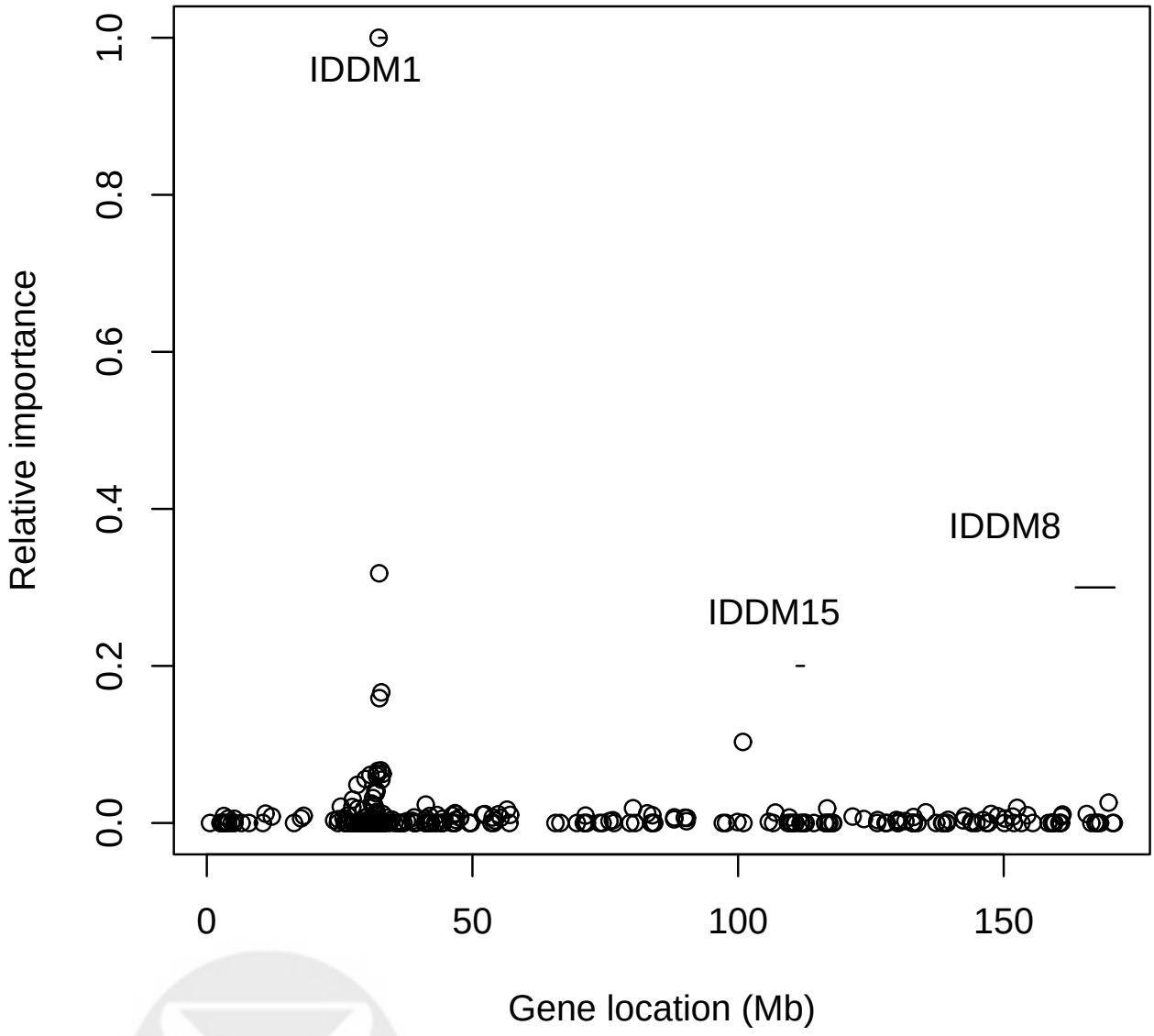

Figure 4: *Relative importance scores for genes on chromosome 6 for type 1 diabetes. IDDM1, IDDM8 and DDM15 are the regions that were shown to be linked to type 1 diabetes by linkage analysis.*