Multivariate Boosting for Integrative Analysis of High-Dimensional Cancer Genomic Data

Lie Xiong[1], Pei-Fen Kuan[2], Jianan Tian[1], Sunduz Keles[1,3] and Sijian Wang[1,3]

[1]Department of Statistics, University of Wisconsin, Madison, WI, USA. [2]Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY, USA. [3]Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI, USA.

Supplementary Issue: Array Platform Modeling and Analysis (B)

ABSTRACT: In this paper, we propose a novel multivariate component-wise boosting method for fitting multivariate response regression models under the high-dimension, low sample size setting. Our method is motivated by modeling the association among different biological molecules based on multiple types of high-dimensional genomic data. Particularly, we are interested in two applications: studying the influence of DNA copy number alterations on RNA transcript levels and investigating the association between DNA methylation and gene expression. For this purpose, we model the dependence of the RNA expression levels on DNA copy number alterations and the dependence of gene expression on DNA methylation through multivariate regression models and utilize boosting-type method to handle the high dimensionality as well as model the possible nonlinear associations. The performance of the proposed method is demonstrated through simulation studies. Finally, our multivariate boosting method is applied to two breast cancer studies.

KEYWORDS: boosting, breast cancer, integrative genomic analysis, multivariate regression

SUPPLEMENT: Array Platform Modeling and Analysis (B)

CITATION: Xiong et al. Multivariate Boosting for Integrative Analysis of High-Dimensional Cancer Genomic Data. *Cancer Informatics* 2014:13(S7) 123–131 doi: 10.4137/CIN.S16353.

RECEIVED: September 18, 2014. **RESUBMITTED:** March 16, 2015. **ACCEPTED FOR PUBLICATION:** March 20, 2015.

ACADEMIC EDITOR: J.T. Efird, Editor in Chief

TYPE: Methodology

FUNDING: The work of Keles was supported by NIH grants HG003747 and HG007019. The work of Wang was supported by NIH Grant HG007377-02. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: swang@biostat.wisc.edu

Introduction

High-resolution microarrays and second-generation sequencing platforms are powerful tools to investigate genome-wide alterations in DNA copy number, methylation, and gene expression associated with a disease. An integrated genomic profiling approach measures multiple omics data types simultaneously in the same set of biological samples. As integrated genomic studies emerge, it has become increasingly clear that true oncogenic mechanisms are more visible when combining evidence across patterns of alterations in DNA copy number, methylation, gene expression, and mutational profiles.[1,2] Integrative analysis of multiple omic data types can help the search of potential drivers by uncovering genomic features that tend to be dysregulated by multiple mechanisms.[3]

This paper is motivated by two specific applications of integrative analysis of cancer genomic data. The first application is to study the influence of DNA copy number alterations (from array comparative genomic hybridization (aCGH)) on RNA transcript (from microarray gene expression experiments) levels. While useful information has been revealed by analyzing expression arrays alone or CGH arrays alone, careful integrative analysis of DNA copy numbers and expression data is necessary as these two types of data provide complimentary information in gene characterization. Specifically, RNA data contain information on genes that are over/underexpressed, whereas DNA copy numbers contain information on gains and losses that are potential drivers of cancer. Therefore, integrating DNA and RNA data may help discern more subtle (yet biologically important) genetic regulatory relationships in cancer cells.[4]

The second application is to investigate the association between DNA methylation and gene expression. The advancement of global DNA methylation arrays and next-generation RNA sequencing transcriptome studies now allow scientists to investigate the functional consequence of DNA methylation in various genomic regions, including CpG islands (CGIs), which have been extensively investigated in the literature.[5–7] Generally, methylation of regulatory CGIs is thought to downregulate transcription by promoting the formation of heterochromatin and preventing the binding of transcription factors (TFs).[8] In normal cells, CGIs are protected from methylation. However, hypermethylation of promoter CGIs of important genes, ie, tumor suppressor genes (TSGs), is frequently observed in cancer cells.[9] Thus, integrating DNA methylation and RNA data could bring a step closer in unraveling the complex genetic regulatory relationships in cancer cells.

The most straightforward way to model the dependence of RNA levels on DNA copy numbers or the dependence

of gene expression on methylation is through a multivariate (multiple-response) regression model with gene expression as responses and the DNA copy numbers or methylation as covariates. Many other important biological problems can be modeled using the multivariate regression model as well. For example, in expression quantitative trait loci (eQTL) mapping, we can treat gene expressions as responses and genomic markers as covariates to identify genomic locations to which expression traits are linked.[10,11] In the integrative analysis of gene expression and chromatin immunoprecipitation (chip-chip) data, we can treat gene expression as responses and TFs as covariates to identify TFs that are related to cell cycle regulation.[12]

While the multivariate linear regression is well studied in statistical literatures, the current problems pose new challenges because of (i) the complicated relationships among response variables, (ii) high dimensionality in terms of both covariates and responses, and (iii) possible nonlinear association between responses and covariates. Because of challenge (i), the naive approach of regressing each response onto the covariates separately is unlikely to produce satisfactory results, as such methods often lead to high variability and overfitting.[13,14] For challenge (ii), sparse regularization schemes have been utilized in high-dimensional multivariate regression.[15–18] Factor analysis has been considered to reduce the dimensionality as well.[19–21] For challenge (iii), most of existing methods are based on linear models and may have inadequate performance when the linear assumption is violated.

Boosting, originally proposed as an ensemble scheme for classification, ie, AdaBoost,[22] has attracted a lot of attention both in the machine learning and statistics literature, mainly because of its flexibility in modeling possible nonlinear association and excellent prediction performance. Friedman et al.[23] demonstrated that the AdaBoost ensemble method can be represented as a stagewise forward additive modeling procedure, which leads to many new versions of boosting. One particular method is component-wise L_2-boosting, which has been demonstrated to be a powerful method for univariate regression with single response.[24,25]

In this paper, we introduce a novel multivariate component-wise boosting method, which extends the univariate component-wise boosting from the single-response setting to the multiple-response setting. Inherited from the univariate component-wise boosting, the proposed method is not only able to model the nonlinear association between response and covariates when nonlinear base learners are used but also able to avoid overfitting in high-dimensional setting because of the implicit shrinkage in the estimation of the method. By jointly fitting regression models on all responses, the proposed method is able to borrow strength across different responses and is shown to have better performance in both prediction and variable selection than the separate univariate boosting method.

Methods

We start our exposition by assuming that we have n observations. For the ith subject, we observe q responses, $y_i^1,\ldots,y_i^q$, and p covariates, $x_{i1},\ldots, x_{ip}$. We consider the following multivariate regression model:

$$y_i^g = F^g(x_{i1},\ldots,x_{ip}) + \epsilon_i^g, \quad i = 1,\ldots,n, \quad g = 1,\ldots,q, \qquad (1)$$

where $F^g(\cdot)$ is the regression function for the gth response and ϵ_i^g's are error terms. Without loss of generality, we assume that all responses and covariates are standardized, ie, have zero mean and unit variance, before the analysis.

In the remaining of this section, we first review the univariate component-wise boosting method with single response, followed by presenting our multivariate component-wise boosting method with multiple responses.

Univariate component-wise boosting with single response. The component-wise boosting[25,26] is a special version of the general boosting method.[22–24,27,28] It starts from the null model, ie, the model with no covariate, and in each iteration, the model is updated by adding the covariate that yields the greatest improvement in the model fit at the current stage. We summarize the algorithm as follows.

Algorithm: univariate component-wise boosting

Step 1 (Initialization) Set iteration index $m = 0$. Initialize a starting model $\hat{F}_0$, for example, $\hat{F}_0 = 0$.

Step 2 (Component-wise regression) Compute residuals $u_i = y_i - \hat{F}_m(x_i)$, $i = 1,\ldots,n$. For each covariate x_j, fit a univariate regression model such that

$$\hat{h}_j = \arg\min_{h(\cdot)} \sum_{i=1}^{n} (u_i - h(x_{ij}))^2.$$

Step 3 (Model update) Select the covariate x_j such that

$$j = \arg\min_{k} \sum_{i=1}^{n} (u_i - \hat{h}_k(x_{ik}))^2.$$

Update $\hat{F}_{m+1} = \hat{F}_m + v\hat{h}_j(x_j)$, where $0 < v < 1$ is a shrinkage factor, for example, $v = 0.01$.

Step 4 (Iteration) Increase iteration index m by one and repeat Step 2 and Step 3 until reaching a stopping time M.

In Step 2, $h(\cdot)$ is known as a base learner, which is a function to model the association between the residual and a single covariate. Three popular choices for $h(\cdot)$ are linear function $h(x) = x\beta$, stump (one-level regression tree), and smoothing spline.[25] When the stump or smoothing spline is used, the final model may capture the nonlinear association between the response and the covariates. The number of iterations M is a tuning parameter that can be selected using a validation set or cross-validation. The importance of each covariate can be evaluated by the cumulative deduction of mean square error (MSE) in the fitting procedure.[29]

The parameter v in Step 3 can be regarded as controlling the learning rate of the boosting procedure. Smaller values of v (more shrinkage) result in larger training risk for the same number of iterations M. Thus, both v and M control prediction risk on the training data. However, these parameters do not operate independently. Smaller values of v lead to larger values of M for the same training risk, so that there is a tradeoff between them. In terms of selection of v, empirically it has been found[24] that smaller values of v favor better test error and require correspondingly larger values of M. In fact, Friedman[24] suggested that the best strategy appears to be to set v to be very small ($v < 0.1$) and then choose M by early stopping. In this paper, we fix $v = 0.01$ in all numerical studies.

An advantage of the component-wise boosting method is the computation efficiency, since in each iteration, only univariate models are fitted. This makes the boosting method very suitable for analyzing high-dimensional data.

Multivariate component-wise boosting. We now consider the multiple-response setting. The key of our extension from single-response boosting to multi-response boosting is to modify the way to select the best covariate to update the model to borrow strength across multiple responses. We use the integrative analysis of DNA copy numbers and expression data as an example. Suppose a DNA copy number alteration is associated with several genes, but these individual associations are all relatively weak. If we look at the association between the DNA copy number alteration and gene expressions one by one (single-response regression), we may fail to discover this alteration. However, if we can combine the signal across different genes and consider an overall association between this alteration and all genes, we may have a better chance of identifying the alteration. This enlightens us to consider selecting x_j, which yields the smallest overall MSE (which is equivalent to the best overall-fit). Specifically, in each iteration, suppose $u_i^g = y_i^g - \hat{F}_m^g(x_i)$ is the current residual associated with the gth response. We select the covariate x_j such that

$$j = \arg\min_k \sum_{g=1}^{q} \sum_{i=1}^{n} \frac{(u_i^g - \hat{h}_k^g(x_{ik}))^2}{\sum_{i=1}^{n}(u_i^g)^2}, \qquad (2)$$

where we adjust the MSE by the square of Euclidean norm of the corresponding residual. This adjustment is shown to work better than the unadjusted method in our numerical analysis. When there is only one response, we can see that the criterion given by (2) is equivalent to the criterion used in the univariate component-wise boosting.

Once the covariate x_j is selected, the next step is to update the models for q responses. One option is to update all q models by including x_j, but this leads to an all-in-all-out updating strategy, with which the covariate is expected to be either associated with all responses or not associated with any response. In many practical problems, however, it is more likely that one covariate is associated with only a subset of responses in our multivariate boosting method. For example, in the integrative analysis of copy number variation and gene expression, we may expect a particular copy number affects only a small set of genes. This enlightens us to update a subset of responses. Specifically, we first rank all q responses in an increasing order according to their adjusted MSEs $\sum_{i=1}^{n}(u_i^g - \hat{h}_k^g(x_j))^2 / \sum_{i=1}^{n}(u_i^g)^2$. Then we only update models for the first s responses, which have the strongest association with the covariate x_j. The number s can be a tuning parameter or fixed at $s = 1$ for simplicity. We summarize our multivariate component-wise boosting algorithm as follows.

Algorithm: multivariate component-wise boosting

Step 1 (Initialization) Set iteration index $m = 0$. Initialize starting models $\hat{F}_0^g$, for example, $\hat{F}_0^g = 0, \quad g = 1,\ldots,p$.

Step 2 (Component-wise regression) Compute residuals $u_i^g = y_i^g - \hat{F}_m^g(x_i), \quad i = 1,\ldots,n, \quad g = 1,\ldots,q$. For each covariate x_j, fit a univariate regression model such that

$$\hat{h}_j^g = \arg\min_{h(\cdot)} \sum_{i=1}^{n} (u_i^g - h^g(x_{ij}))^2.$$

Step 3 (Model update) For each covariate x_j and each response, calculate the adjusted MSE (adj-MSE) e_j^q:

$$e_j^q = \sum_{i=1}^{n} \frac{(u_i^g - \hat{h}_k^g(x_{ij}))^2}{\sum_{i=1}^{n}(u_i^g)^2}, \quad j = 1,\ldots,p, \quad g = 1,\ldots,q.$$

Select the covariate x_j such that $j = \text{argmin}_k \sum_{g=1}^{q} e_k^q$. For this covariate x_j, sort e_j^q in the increasing order. For the responses with s smallest e_j^q, update $\hat{F}_{m+1}^g = \hat{F}_m^g + v\hat{h}_j^g(x_j)$, where $0 < v < 1$. One can take $v = 0.01$ as in Section 2.1.

Step 4 (Iteration) Increase iteration index m by one and repeat Step 2 and Step 3 until reaching a stopping time M.

The covariate importance can be evaluated using the similar strategy as component-wise boosting with single response described in Section 2.1. For each covariate, the response-specific importance score is calculated by the cumulative adjusted MSE for each response. The overall importance score for a covariate is the summation of its response-specific importance scores. The R code is available upon request.

We would like to point out that Lutz and Buhlmann[30] also proposed a multivariate component-wise L_2-boosting, but their approach is still within the linear model framework. In addition, it requires the estimation of inverse covariance matrix of covariates, which can be challenging when fitting high-dimensional genomic data.

Simulation Study

In this section, we conduct simulation studies to demonstrate the operating characteristics of our multivariate component-wise boosting method. We consider five methods: separate

analysis with the lasso method (slasso), multivariate adaptive regression splines (MARS),[31] regularized multivariate regression for identifying master predictors (RemMap),[18] the separate univariate boosting method (sboost), and our multivariate boosting method (mboost). For boosting methods, we consider three base learners: ordinary least square (ols), one-level regression tree (tree), and smoothing spline (spline). The lasso method is implemented using R package glmnet, the MARS method is implemented using R package earth, and the RemMap method is implemented using R package remMap.

Simulation setup. We consider two scenarios: (1) the true model is linear and (2) the true model is nonlinear. In scenario (1), we simulate data for $n = 50$ observations, $p = 100$ covariates, and $q = 100$ responses. The true model is

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}, \tag{3}$$

where $\mathbf{Y}$ is a 50×100 response matrix, $\mathbf{B}$ is a 100×100 coefficient matrix, $\mathbf{X}$ is a 50×100 covariate matrix, and $\mathbf{E}$ is a 50×100 error matrix. Each row of the covariate matrix $\mathbf{X}$ is drawn independently from MVN(0, Σ_X), where $\Sigma_X = [\sigma_{X_{ij}}]_{100\times100}$ with $\sigma_{X_{ij}} = 0.7^{|i-j|}$. The rows of error matrix $\mathbf{E}$ are drawn independently from MVN(0, Σ_E), where $\Sigma_E = [\sigma_{E_{ij}}]_{100\times100}$ with $\sigma_{E_{ij}} = \rho_E^{|i-j|}$. We consider two values of ρ_E: 0 and 0.9. The coefficient matrix $\mathbf{B}$ is generated by matrix element-wise product $\mathbf{B} = \mathbf{W}^*\mathbf{K}^*\mathbf{Q}$. The element of $\mathbf{W}$ is drawn independently from $N(0,1)$, the element of $\mathbf{K}$ is drawn independently from Bernoulli distribution with success probability 0.5, and $\mathbf{Q}$ has rows that are either all ones or all zeros, where p-independent Bernoulli variables with success probability 0.1 are drawn to determine whether a row is one or zero.

In scenario (2), we simulate data for $n = 100$ observations, $p = 50$ covariates, and $q = 20$ responses. The true model is

$$y_i^g = \begin{cases} 5f_1(x_{i1}) + 3f_2(x_{i2}) + 4f_3(x_{i3}) + \varepsilon_i^g & \text{for } g = 1,\ldots,10, \\ 5f_1(x_{i1}) + 3f_2(x_{i2}) + 6f_4(x_{i4}) + \varepsilon_i^g & \text{for } g = 11,\ldots,15, \\ 5f_1(x_{i1}) + 3f_2(x_{i2}) + \varepsilon_i^g & \text{for } g = 15,\ldots,20, \end{cases} \tag{4}$$

where $f_1(x) = x$, $f_2(x) = (2x-1)^2$, $f_3(x) = \sin(2\pi x)/(2-\sin(2\pi x))$, and $f_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin^2(2\pi x) + 0.4\cos^3(2\pi x) + 0.5\sin^3(2\pi x)$. The domain of all four functions is [0,1]. The covariates are generated as $x_j = (w_j + u)/2$, $j = 1,\ldots, 50$, where elements of w_j, $j = 1, \ldots, 50$ and u are drawn independently from uniform distribution $U(0,1)$. The elements of ε_i^g are drawn independently from $N(0, \sigma^2)$, where σ^2 is chosen such that the average signal-to-noise ratio of the model is 1 or 3.

Simulation analysis. For both scenarios, we repeated the simulation for 100 times. All tuning parameters were selected using five-fold cross-validation. To evaluate the variable selection performance of a method, we consider the following two measurements, namely sensitivity and specificity:

$$Sensitivity = \frac{\#\left\{(g,j): \hat{d}_j^g \neq 0 \text{ and } d_j^g \neq 0\right\}}{\#\left\{(g,j): d_j^g \neq 0\right\}}, \tag{5}$$

$$Specificity = \frac{\#\left\{(g,j): \hat{d}_j^g = 0 \text{ and } d_j^g = 0\right\}}{\#\left\{(g,j): d_j^g = 0\right\}}, \tag{6}$$

where $d_j^g = 1$ if the gth response is associated with the jth covariate in the true model, $\hat{d}_j^g = 1$ if the gth response is associated with the jth covariate in the fitted model, and $\hat{d}_j^g = 0$ otherwise.

To evaluate the prediction performance of the model, in each simulation, we generated a test set with $n = 10{,}000$ observations using the same distribution as the training data. Following suggestion from a reviewer, we calculated the following scaled average mean square error (SAMSE) on the test set:

$$\text{SAMSE} = \frac{1}{nq}\sum_{i=1}^{n}\sum_{k=1}^{q}(y_{ik} - \hat{y}_{ik})^2/\sigma_k^2, \tag{7}$$

where n is the sample size, q is the number of responses, and σ_k^2 is the variance of error term for the kth response.

Simulation results. The simulation results are summarized in Table 1. In both linear and nonlinear scenarios, the multivariate boosting method has smaller average MSEs, higher sensitivity, and higher specificity than the separate univariate boosting, no matter which base learner is used. When the true model is linear, the multivariate boosting with OLS as base learner has smaller average MSEs than the other two multivariate regression methods: MARS and RemMap. When the true model is nonlinear, the multivariate boosting with a nonlinear base learner (tree or spline) has a clear advantage in prediction over MARS and RemMap. The multivariate boosting method also has either significantly higher or comparable sensitivity and specificity than MARS and RemMap.

Real Application

In this section, we analyze two publicly available breast cancer datasets from integrative cancer genomic studies.

Integrative analysis of gene expression data and copy number variation data. In this subsection, we analyze a breast cancer dataset described in Sørlie et al.[32] The dataset contains both the microarray gene expression and aCGH profiled for 172 breast cancer specimens. For each specimen, we have expression profiles of 578 genes and 384 copy number alteration intervals (CNAIs). Since variations in DNA copy number play an important role in cancer development through altering the expression levels of cancer-related genes, we would like to identify important genes in aCGH experiment, model the association of RNA transcript level with gene copy

Table 1. Results of simulation studies. In each cell, the number outside the parenthesis is the average value over 100 replications and the number within the parenthesis is the corresponding standard error.

METHOD	AMSE	SENSITIVITY	SPECIFICITY	AMSE	SENSITIVITY	SPECIFICITY
	Scenario (1), $\rho_E = 0$			Scenario (1), $\rho_E = 0.9$		
mboost.ols	0.366 (0.007)	0.81 (0.003)	0.97 (0.001)	0.369 (0.007)	0.80 (0.003)	0.96 (0.001)
sboost.ols	0.400 (0.007)	0.73 (0.003)	0.88 (0.003)	0.399 (0.007)	0.73 (0.003)	0.87 (0.004)
mboost.spline	0.406 (0.006)	0.77 (0.008)	0.96 (0.002)	0.404 (0.008)	0.77 (0.83)	0.95 (0.20)
sboost.spline	0.444 (0.008)	0.69 (0.005)	0.82 (0.004)	0.443 (0.008)	0.69 (0.49)	0.85 (0.46)
mboost.tree	0.559 (0.005)	0.75 (0.004)	0.87 (0.003)	0.562 (0.006)	0.75 (0.40)	0.87 (0.26)
sboost.tree	0.931 (0.006)	0.67 (0.003)	0.73 (0.006)	0.926 (0.006)	0.68 (0.38)	0.72 (0.56)
MARS	0.604 (0.010)	0.89 (0.02)	1 (0)	0.598 (0.010)	0.89 (1.48)	0.96 (0.08)
RemMap	0.370 (0.008)	0.89 (0.002)	0.59 (0.005)	0.374 (0.008)	0.89 (0.20)	0.66 (0.64)
slasso	0.421 (0.008)	0.65 (0.004)	0.04 (0.006)	0.423 (0.008)	0.65 (0.38)	0.94 (0.06)
	Scenario (2), Average SNR = 1			Scenario (2), Average SNR = 3		
mboost.spline	0.72 (0.01)	0.91 (0.01)	0.87 (0.005)	0.44 (0.01)	0.99 (0.002)	0.83 (0.005)
sboost.spline	0.75 (0.01)	0.87 (0.01)	0.81 (0.005)	0.44 (0.02)	0.99 (0.003)	0.76 (0.006)
mboost.tree	0.76 (0.01)	0.89 (0.01)	0.81 (0.007)	0.51 (0.01)	0.99 (0.002)	0.72 (0.007)
sboost.tree	0.75 (0.01)	0.84 (0.01)	0.74 (0.006)	0.52 (0.01)	0.98 (0.003)	0.63 (0.008)
MARS	0.78 (0.01)	0.76 (0.02)	0.94 (0.002)	0.45 (0.01)	0.98 (0.009)	0.95 (0.003)
RemMap	0.89 (0.01)	0.72 (0.02)	0.72 (0.014)	0.73 (0.01)	0.72 (0.014)	0.72 (0.013)
slasso	0.90 (0.01)	0.58 (0.01)	0.90 (0.003)	0.74 (0.01)	0.63 (0.007)	0.89 (0.003)

number, and predict the RNA transcript levels using DNA copy number variations in aCGH experiments. The details of data preprocessing can be found in Peng et al.[18]

In our analysis, the 172 samples were randomly split into a training set (120 samples) and a test set (52 samples). All methods were applied to training set to build the predictive models. All tuning parameters were selected using five-fold cross-validation. Once an optimal model was obtained, it was used to predict the gene expression profiles for samples in the test set. To evaluate the prediction performance of the model, we calculate the following SAMSE on the test set:

$$\text{SAMSE} = \frac{1}{nq}\sum_{i=1}^{n}\sum_{k=1}^{q}\left(y_{ik} - \hat{y}_{ik} \right)^2 / \hat{\sigma}_k^2 \qquad (8)$$

where n is the sample size, q is the number of responses, and $\hat{\sigma}_k^2 = \sum_{i=1}^{n}\left(y_{ik} - \bar{y}_{ik}\right)^2 / (n-1)$ is the sample standard deviation of the kth responses on the test set.

The procedure was repeated 100 times, and the results are shown in Table 2. The multivariate boosting method with tree as the base learner has the smallest average MSE among all methods, which suggests a possible nonlinear association between gene expression and copy number alteration. In addition, the multivariate boosting method tends to select a smaller number of CNAIs than the separate univariate boosting method, no matter which base learner is used. When the ordinary least square is used as the base learner, our multivariate boosting methods select a smaller number of CNAIs than the other two methods based on linear model, RemMap and slasso. The performance of MARS method is inadequate. It has the largest AMSE and selects only about four CNAIs in average. This may be because the MARS method selects CNAIs in the all-in-all-out fashion, ie, a CNAI is considered to be associated with either all genes or no gene, but the true relationship pattern is possibly that one CNAI only affects a subset of genes.

We also calculate the average covariate importance scores over 100 replicates for both univariate boosting method and multivariate boosting method. Figure 1 shows the concentration plots of the importance scores. For each concentration plot, the covariate importance scores are sorted from the largest to the smallest. The x-axis represents the sorted index of covariates, and the y-axis represents the cumulative importance scores. It is evident from these plots that the covariate importance scores given by the multivariate boosting method is more concentrated than the importance scores given by the separate univariate boosting method. For example, when the regression tree is used as the base learner, the top 100 covariates selected by the multivariate boosting method have a cumulative importance score of 90, while the separate univariate boosting method needs more than 200 covariates to reach the cumulative importance score of 90.

Table 3 lists the 10 CNAIs with top average importance scores selected by the multivariate boosting methods. We can see that there is a strong concordance among multivariate boosting methods with three different base learners on selected CNAIs with high important scores. Among the identified CNAIs, loss of heterozygosity on chromosome 19pl3[33] has been shown to be one of the most frequent alteration observed

Table 2. Results of integrative analysis of expression and copy number alteration on Sørlie's breast cancer dataset. In each cell, the number outside the parenthesis is the average value over 100 replications and the number within the parenthesis is the corresponding standard error.

METHOD	SAMSE	# OF SELECTED COVARIATES
mboost.ols	0.92 (0.01)	154.9 (3.4)
sboost.ols	0.92 (0.004)	381.8 (0.2)
mboost.tree	0.89 (0.004)	254.2 (2.7)
sboost.tree	0.90 (0.004)	384.0 (0.1)
mboost.smsp	0.92 (0.004)	81.7 (1.2)
sboost.smsp	0.94 (0.004)	379.8 (0.9)
MARS	1.00 (0.01)	3.7 (0.1)
RemMap	0.92 (0.004)	201.6 (0.9)
slasso	0.91 (0.003)	311.5 (1.0)

in breast cancer, whereas breast cancer-related genes such as PYCARD, IL4R, and PLK1 are located within chromosome 16p13.3–16p11.2.

Integrative analysis of gene expression data and methylation data. In this subsection, we analyze a breast cancer dataset described in The Cancer Genome Atlas (TCGA) Network.[34] The purpose of the analysis is to model the association between gene expression and DNA methylation across CpG sites. In our analysis, we consider a set of 808 cancer-related genes derived from published cancer gene lists. Furthermore, we focus on genes in chromosomes 13 and 17 as these two chromosomes have been shown to exhibit strong implication in breast cancer development,[35,36] thus enabling us to better relate our results to known biological findings. Finally, we remove genes whose number of missing values is greater than 10 and remove subjects with any missing values. The final dataset has 333 samples with gene expression profiles of 41 genes and methylation values of 88 CpG sites.

In our analysis, the 333 samples were randomly split into a training set (267 samples) and a test set (66 samples). The data analysis procedures are the same as described in Section 3.1. The results are shown in Table 4. Similar to the results of integrative analysis of gene expression and copy number alterations, the MARS method has an inadequate performance. It has the largest AMSE among five methods and selects only about nine CpG sites in average. This may be because the all-in-all-out selection fashion of MARS does not fit the data well. All methods except MARS have comparable prediction performances; however, our multivariate boosting method has much smaller model size compared to other methods. Figure 2 shows the concentration plots of the average importance scores over 100 replicates. From the plots, we can see that the covariate importance scores given by multivariate boosting method are more concentrated than the importance scores given by

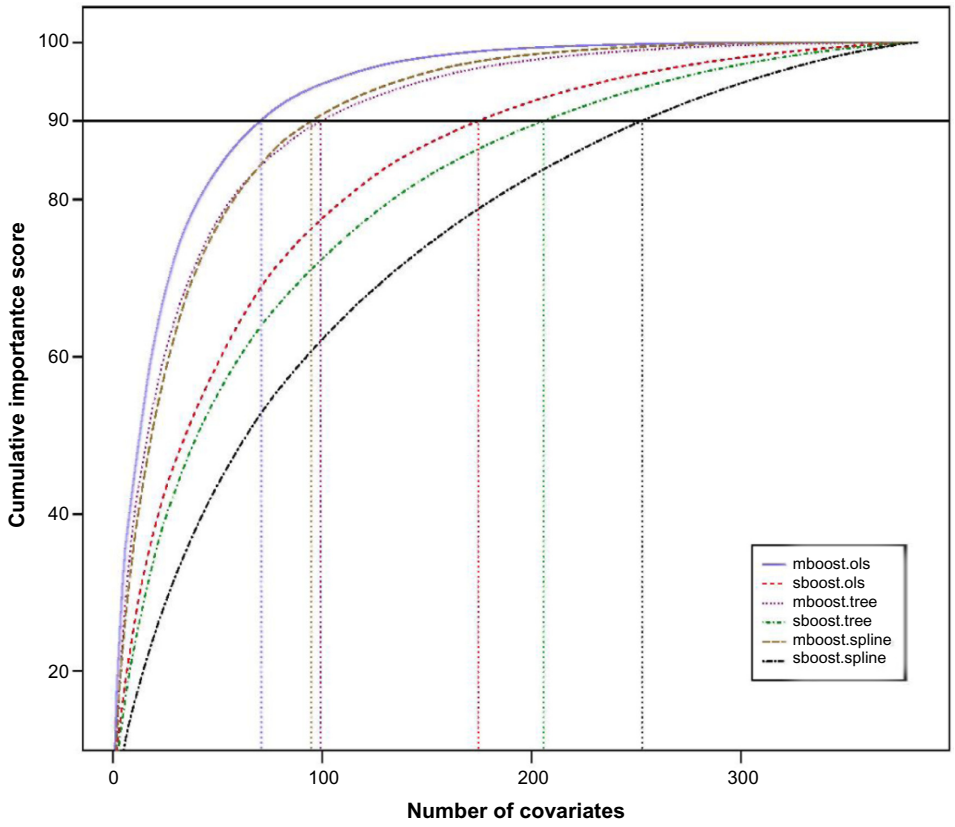

Figure 1. Concentration plots of covariate importance scores for boosting-based methods.

Table 3. Lists of top 10 CNAIs selected by multivariate boosting methods.

MBOOST.OLS		MBOOST.TREE		MBOOST.SPLINE	
SCORE (s.d.)	CYTOBAND	SCORE (s.d.)	CYTOBAND	SCORE (s.d.)	CYTOBAND
9.36 (0.55)	16p13.3–16p11.2	7.85 (0.43)	16p13.3–16p11.2	6.59 (0.39)	16p13.3–16p11.2
8.12 (0.41)	17q12–17q12	4.70 (0.36)	19p13.2–19p12	5.01 (0.34)	1p36.11–1p35.2
5.14 (0.47)	17q21.2–17q21.31	4.44 (0.36)	19p13.3–19p13.2	4.78 (0.46)	19p13.3–19p13.2
4.81 (0.48)	1p36.11–1p35.2	3.94 (0.33)	1p34.3–1p34.2	3.73 (0.32)	17q21.2–17q21.31
4.31 (0.39)	19p13.3–19p13.3	3.87 (0.32)	17q12–17q12	3.46 (0.28)	17q21.31–17q21.32
3.65 (0.35)	17q21.31–17q21.32	3.68 (0.32)	1p36.11–1p35.2	2.88 (0.22)	4p16.3–4p16.1
2.27 (0.22)	17q12–17q12	3.15 (0.27)	2q31.1–2q31.1	2.73 (0.16)	10q21.3–10q22.2
2.19 (0.23)	5q23.3–5q31.3	2.64 (0.26)	17q21.2–17q21.2	2.47 (0.17)	17q12–17q12
2.18 (0.23)	4p16.3–4p16.1	2.56 (0.28)	17q21.2–17q21.31	2.34 (0.26)	1p34.3–1p34.2
2.05 (0.28)	5q13.2–5q13.2	2.35 (0.16)	15q11.2–15q11.2	2.26 (0.15)	15q11.2–15q11.2

separate univariate boosting method. For example, when the smoothing spline is used as the base learner, the top 20 covariates selected by the multivariate boosting method have a cumulative importance score of 90, while the separate univariate boosting method needs approximately 40 covariates to reach the cumulative importance score of 90.

Table 5 lists the 10 CpG sites with top average importance scores selected by the multivariate boosting methods. We can see that there is a strong concordance among multivariate boosting methods with three different base learners on selected CpG sites with high important scores. Among these selected genes, GFAP, GRB7, ALOX12, MFAP4, and HOXB2 are well unknown to be related to breast cancer.

Discussion

In this paper, we propose a novel multivariate componentwise boosting method to fit multivariate regression models. The proposed method is able to model the possible nonlinear associations between responses and covariates. By jointly fitting models across multiple responses, our method identifies important covariates based on their overall effects across responses. The performance of the proposed methods is demonstrated using both simulation studies and real data analysis.

Table 4. Results of integrative analysis of expression data and methylation data on TCGA breast cancer dataset. In each cell, the number outside the parenthesis is the average value over 100 replications and the number within the parenthesis is the corresponding standard error.

METHOD	SAMSE	# OF SELECTED COVARIATES
mboost.ols	0.91 (0.02)	77.04 (0.30)
sboost.ols	0.93 (0.01)	87.96 (0.02)
mboost.tree	0.80 (0.01)	64.19 (0.77)
sboost.tree	0.80 (0.01)	87.74 (0.06)
mboost.spline	0.83 (0.01)	45.34 (0.71)
sboost.spline	0.89 (0.03)	85.68 (0.13)
MARS	0.84 (0.01)	8.45 (0.10)
RemMap	0.91 (0.02)	84.53 (0.17)
SLASSO	0.89 (0.01)	87.90 (0.03)

In many important problems, there could be natural grouping structures in responses and/or covariates. For example, in eQTL study, on one hand, genes belonging to the same pathway can be grouped together, and we may expect a single nucleotide polymorphism (SNP) to be associated with many genes within a pathway. On the other hand, SNPs belonging to the same gene can be grouped together, and we may expect several SNPs within a same coding region are associated with the expression of one gene. Integrating these grouping structures in the analysis may help boost the signal-to-noise ratio.[37] Our proposed method can be modified to take into account these grouping information in the analysis as follows.

Suppose that there are S groups in responses and T groups in covariates. Let $A_k \subset \{1, \ldots, G\}$, $k = 1, \ldots, S$, and $B_k \subset \{1, \ldots, p\}$, $k = 1, \ldots, T$, are the corresponding indices associated with these groups. We allow for overlaps among A_k's or B_k's. These groups create $S \times T$ blocks. In each iteration of boosting, we first select a block with the best overall association and then select a subset of covariates and a subset of response to update. Specifically, in each iteration, we first select the block (s, t) such that

$$(s,t) = \arg \min_{(q,r):1\le q\le S, 1\le r\le T} \frac{1}{|A_q||B_r|} \sum_{g\in A_q}\sum_{j\in B_r}\sum_{i=1}^{n} \frac{(u_i^g - \hat{h}_j^g(x_{ij}))^2}{\sum_{i=1}^n (u_i^g)^2}, \quad (9)$$

where $|A_q|$ and $|B_r|$ are the cardinalities of A_q and B_r, respectively. Then we rank all response–covariate combinations within the (s, t)th block in increasing order according to their corresponding

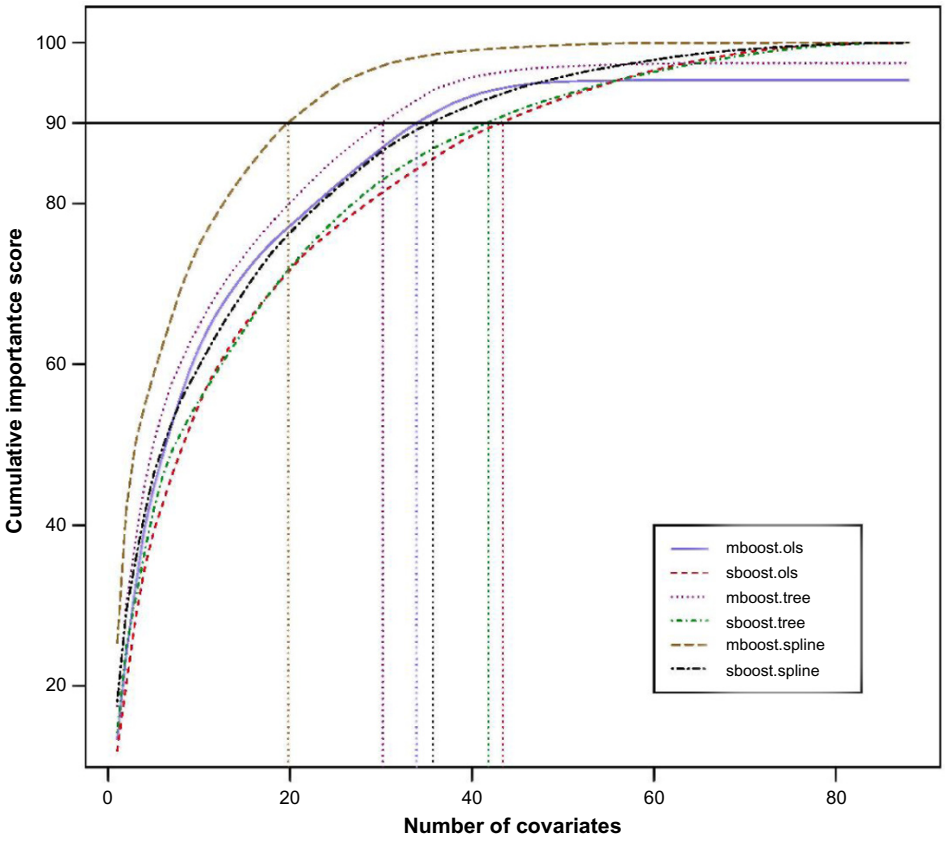

Figure 2. Concentration plots of covariate importance scores for boosting-based methods.

Table 5. Lists of top 10 CpG sties selected by multivariate boosting methods.

MBOOST.OLS			MBOOST.TREE			MBOOST.SPLINE		
COORDINATE	GENE	SCORE (s.d.)	COORDINATE	GENE	SCORE (s.d.)	COORDINATE	GENE	SCORE (s.d.)
cg21944455	GFAP	13.39 (0.87)	cg21944455	GFAP	17.48 (1.44)	cg21944455	GFAP	25.35 (1.83)
cg03684977	GRB7	10.72 (0.84)	cg03760483	ALOX12	12.78 (1.38)	cg03760483	ALOX12	16.76 (1.36)
cg03760483	ALOX12	8.09 (1.67)	cg03684977	GRB7	7.82 (0.93)	cg03684977	GRB7	8.00 (1.29)
cg13030582	MFAP4	7.36 (2.48)	cg13030582	MFAP4	6.91 (1.85)	cg25882366	HOXB2	4.73 (0.89)
cg25882366	HOXB2	5.01 (0.36)	cg25882366	HOXB2	5.26 (0.84)	cg03001305	STAT5A	3.92 (0.58)
cg05292376	AATK	4.18 (0.37)	cg03001305	STAT5A	3.98 (0.55)	cg05292376	AATK	3.63 (0.65)
cg03001305	STAT5A	3.79 (0.50)	cg13263114	ERBB2	3.47 (0.54)	cg09038914	GFAP	3.54 (0.48)
cg25465406	GUCY2D	3.56 (1.05)	cg11679069	DNAJC15	2.51 (0.54)	cg25465406	GUCY2D	3.47 (1.13)
cg09038914	GFAP	3.25 (0.44)	cg25465406	GUCY2D	2.50 (0.78)	cg13030582	MFAP4	2.89 (2.15)
cg17129388	NGFR.2	2.71 (0.57)	cg09038914	GFAP	2.16 (0.47)	cg13263114	ERBB2	2.57 (0.32)

MSEs and update the first u combinations. The number u can be a tuning parameter or fix $u = 1$ for the simplicity. A separate manuscript for this extension is in preparation.

In our current manuscript, we investigate the association between two genomic data sets, one as covariate matrix and the other as response matrix. An interesting future extension of our proposed multivariate boosting method is to consider multiple genomic datasets as covariates and/or responses, each with potentially a different distribution (eg, discrete, categorical, or continuous). Such integrative analysis might bring us a step closer in understanding the complex process of cancer cell development, in collective efforts to realizing the promise of personalized medicine.

Author Contributions

Conceived and designed the experiments: PK, SK, SW. Analyzed the data: LX, JT, SK, SW. Wrote the first draft of the manuscript: LX, SW. Contributed to the writing of

the manuscript: PK, JT, SK. Agree with manuscript results and conclusions: LX, PK, JT, SK, SW. Jointly developed the structure and arguments for the paper: LX, PK, JT, SK, SW. Made critical revisions and approved final version: PK, SK, SW. All authors reviewed and approved of the final manuscript.

REFERENCES

1. TCGA. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*. 2008;455(7216):1061–8.
2. TCGA. Integrated genomic analyses of ovarian carcinoma. *Nature*. 2011;474(7353):609–15.
3. Chin L, Gray JW. Translating insights from the cancer genome into clinical practice. *Nature*. 2008;452(7187):553–63.
4. Pollack JR, Sørlie T, Perou CM, et al. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A*. 2002;99:12963–8.
5. Esteller M. Epigenetic gene silencing in cancer: the DNA hypermethylome. *Hum Mol Genet*. 2007;16(R1):R50–9.
6. Bell JT, Pai AA, Pickrell JK, et al. DNA methylation patterns associate with genetic and gene expression variation in hapmap cell lines. *Genome Biol*. 2011;12(1):R10.
7. Pai AA, Bell JT, Marioni JC, Pritchard JK, Gilad Y. A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues. *PLoS Genet*. 2011;7(2):e1001316.
8. Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet*. 2008;9(6):465–76.
9. Sproul D, Nestor C, Culley J, et al. Transcriptionally repressed genes become aberrantly methylated and distinguish tumors of different lineages in breast cancer. *Proc Natl Acad Sci U S A*. 2011;108(11):4364–9.
10. Kendziorski C, Wang P. A review of statistical methods for expression quantitative trait loci mapping. *Mamm Genome*. 2006;17(6):509–17.
11. Kendziorski C, Chen M, Yuan M, Lan H, Attie A. Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics*. 2006;62(1):19–27.
12. Chun H, Keleş S. Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. *Genetics*. 2009;182(1):79.
13. Breiman L, Friedman J. Predicting multivariate responses in multiple linear regression. *J R Stat Soc Series B Stat Methodol*. 1997;59(1):3–54.
14. Kim, Seyoung, Kyung-Ah Sohn, Eric P. Xing. "A multivariate regression approach to association analysis of a quantitative trait network." *Bioinformatics*. 2009;25(12):i204–12.
15. Bedrick EJ, Tsai C-L. Model selection for multivariate regression in small samples. *Biometrics*. 1994;50:226–31.
16. Obozinski G, Wainwright MJ, Jordan MI. Union support recovery in high-dimensional multivariate regression. In: 2008 46th Annual Allerton Conference on Communication, Control, and Computing. Urbana-Champaign, IL: IEEE; 2008:21–6.
17. Rothman A, Levina E, Zhu J. Sparse multivariate regression with covariance estimation. *J Comput Graph Stat*. 2010;19(4):947–62.
18. Peng J, Zhu J, Bergamaschi A, et al. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann Appl Stat*. 2010;4(1):53–77.
19. Izenman AJ. Reduced-rank regression for the multivariate linear model. *J Multivar Anal*. 1975;5(2):248–64.
20. Reinsel GC, Velu RP. *Multivariate Reduced-Rank Regression*. New York: Springer; 1998.
21. Yuan M, Ekici A, Lu Z, Monteiro R. Dimension reduction and coefficient estimation in multivariate linear regression. *J R Stat Soc Series B Stat Methodol*. 2007;69(3):329–46.
22. Freund, Yoav, Robert E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting." *Journal of computer and system sciences*. 1997;55(1):119–39.
23. Friedman J, Hastie T, Tibshirani R. Special invited paper. additive logistic regression: a statistical view of boosting. *Ann Stat*. 2000;28:337–74.
24. Friedman J. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29:1189–232.
25. Buhlmann P, Yu B. Boosting with the 1 2 loss. *J Am Stat Assoc*. 2003;98(462):324–39.
26. Schmid M, Hothorn T. Boosting additive models using component-wise p-splines. *Comput Stat Data Anal*. 2008;53(2):298–311.
27. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Ann Stat*. 2000;28(2):337–407.
28. Buhlmann P. Boosting for high-dimensional linear models. *Ann Stat*. 2006;34(2):559–83.
29. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning: With Applications in R. Springer Texts in Statistics. New York: Springer; 2014.
30. Lutz R, Buhlmann P. Boosting for high-multivariate responses in high-dimensional linear regression. *Stat Sin*. 2006;16(2):471.
31. Friedman J. Multivariate adaptive regression splines. *Ann Stat*. 1991;19:1–67.
32. Sørlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA*. 2001;98(19):10869–74.
33. Oesterreich S, Allredl D, Mohsin S, et al. High rates of loss of heterozygosity on chromosome 19p13 in human breast cancer. *Br J Cancer*. 2001;84(4):493.
34. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012;490(7418):61–70.
35. Wooster R, Neuhausen SL, Mangion J, et al. Localization of a breast cancer susceptibility gene, brca2, to chromosome 13q12–13. *Science*. 1994;265(5181):2088–90.
36. Orsetti B, Nugoli M, Cervera N, et al. Genomic and expression profiling of chromosome 17 in breast cancer reveals complex patterns of alterations and novel candidate genes. *Cancer Res*. 2004;64(18):6453–60.
37. Newton MA, He Q, Kendziorski C. A model-based analysis to infer the functional content of a gene list. *Stat Appl Genet Mol Biol*. 2012;11(2):1–27.