

## Two-step sparse boosting for high-dimensional longitudinal data with varying coefficients

Yue, Mu; Li, Jialiang; CHENG, Ming-Yen

*Published in:*  
Computational Statistics and Data Analysis

*DOI:*  
[10.1016/j.csda.2018.10.002](https://doi.org/10.1016/j.csda.2018.10.002)

Published: 01/03/2019

*Document Version:*  
Peer reviewed version

[Link to publication](#)

*Citation for published version (APA):*  
Yue, M., Li, J., & CHENG, M-Y. (2019). Two-step sparse boosting for high-dimensional longitudinal data with varying coefficients. *Computational Statistics and Data Analysis*, 131, 222-234.  
<https://doi.org/10.1016/j.csda.2018.10.002>

### General rights

Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent publication URLs

# Two-step Sparse Boosting for High-Dimensional Longitudinal Data with Varying Coefficients

Mu Yue<sup>a,b,\*</sup>, Jialiang Li<sup>b</sup>, Ming-Yen Cheng<sup>c</sup>

<sup>a</sup>*University of Electronic Science and Technology of China, National University of Singapore*

<sup>b</sup>*National University of Singapore, Duke University-NUS Graduate Medical School, Singapore Eye Research Institute*

<sup>c</sup>*Hong Kong Baptist University*

---

## Abstract

Varying-coefficient models are widely used to model nonparametric interaction and recently adopted to analyze longitudinal data measured repeatedly over time. We focus on high-dimensional longitudinal observations in this article. A novel two-step sparse boosting approach is proposed to carry out the variable selection and the model-based prediction. As a new machine learning tool, boosting provides seamless integration of model estimation and variable selection for complicated regression functions. Specifically, in the first step the sparse boosting technique assuming independence is applied to facilitate an initial estimate of the correlation structure while in the second step the estimated correlation structure is incorporated in the loss function of the sparse boosting algorithm. Extensive numerical examples illustrate the advantage of the two-step sparse boosting method. An application of yeast cell cycle gene expression data is further provided to demonstrate the proposed methodology.

*Keywords:* Sparse boosting, Variable selection, Longitudinal data, Varying-coefficient model, Minimum description length

---

\*Corresponding author

*Email address:* [yuemu.moon@gmail.com](mailto:yuemu.moon@gmail.com) (Mu Yue)

## 1. Introduction

High-dimensional model is a focus in statistical literature and many new machine learning approaches have been developed for data with very large number of features. Examples of such features include disease genes and quantitative trait loci in the human genome, biomarkers responsible for disease pathways, and stocks generating profits in investment portfolios. The selection of causal features is a crucial aspect of these scientific and business studies. The situation, where the sample size  $n$  is relatively small but the number of features  $p$  under consideration is extremely large, poses a serious challenge to the selection of causal features. Among all the available techniques, regularization methods such as lasso [1], SCAD [2], MCP [3] and others [4, 5, 6] have been thoroughly studied and widely used to perform variable selection and estimation simultaneously in order to improve the prediction accuracy and interpretability of the statistical model. However, one major disadvantage for these methods is that they all require to specify some tuning parameters dictating the degree of regularization and the optimal selection of the tuning parameters is usually through a time-consuming numerical procedure such as the cross-validation.

We consider boosting as an effective alternative tool for high-dimensional modelling in this article. Boosting was proposed as a family of powerful machine-learning techniques that iteratively forms base learners to minimize the loss function. The original boosting algorithms were proposed by Schapire [7] and Mason et al. [8] while both of them did not take full advantage of the base learners. Then Freund and Schapire [9] developed the AdaBoost algorithm which could adapt to the base learners very easily. The major variation among many boosting algorithms is the loss function. For example, AdaBoost with the exponential loss function,  $L_2$ Boosting [10] with the squared error loss function, Sparse $L_2$ Boosting [11] with the penalized loss function and HingeBoost [12] with the weighted hinge loss function. Moreover, there are more recent various versions of boosting algorithms have been proposed. See, for example, Komori and Eguchi [13] for the pAUCBoost; Bühlmann and Hothorn [14] for the Twin Boosting; Wang [15] for the Twin HingeBoost; Yang and Zou [16] for the ER-Boost and Zhao [17] for GSBoosting. In addition to these extensions, much progress has been made in

understanding the advantages of boosting such as relatively smaller computational cost, lower over-fitting risk and simpler adjustment to incorporate extra constraints. Furthermore, properties of the boosting have been extensively studied. With fixed predictor dimension, the statistical convergence and consistency for boosting-type algorithms  
35 has been established by [18, 19, 20] under classical data setting and further proved by [21, 22] under high-dimensional data setting. In this paper we will examine a novel sparse boosting approach for varying-coefficient models with high-dimensionality.

Varying-coefficient models are an important class of nonparametric models where the regression coefficients vary as functions of random variables. They have been widely  
40 applied in many research fields such as economics [23], finance [24] and epidemiology [25]. For example, in a study to examine the Boston housing price, [26] considered varying-coefficient model by assuming that the impact of predictor variables on housing price varies with the distance to the business hubs, which is an important indicator of geographical accessibility to employment. Yue and Li [27] adopted varying-  
45 coefficient model to identify a set of gene expression signatures highly predictive in the lung cancer where the effects are functions of age. Gao et al. [28] considered varying-coefficient model in a health care study and specified the impacts of BMI determinants as functions of demographic variables to quantify the varying impacts of BMI across demographic groups. Recently varying-coefficient models have been used for high-  
50 dimensional variable screening and variable selection. The assumption of sparsity is usually adopted when dealing with the high-dimensional model [29]. We will propose a novel two-step sparse boosting method for varying-coefficient models. Specifically, we will examine longitudinal data which are frequently encountered in biomedical studies.

Longitudinal data consist of repeated measures on the same subject over time. The as-  
55 sumption that all observations are independent is violated for such data. One important issue in longitudinal analysis is how to take into account of the correlation within subjects and make efficient inference. This issue has been thoroughly investigated by many authors [30, 31] for parametric models. Nonparametric and semiparametric models for longitudinal data are also proposed in the literature, see [32, 33]. There are some recent  
60 development on high-dimensional longitudinal data using varying-coefficient models

[34, 35, 36, 37]. All previous authors considered the penalty methods.

In this paper, a novel two-step sparse boosting approach is proposed to carry out the variable selection and the model-based prediction. In particular, we consider to fit high-dimensional varying-coefficient models with longitudinal data. In the first step of  
65 two-step sparse boosting, the sparse boosting technique is used to yield an estimate of the correlation structure and in the second step, the within-subject correlation structure is taken into account and variable selection and estimation are conducted by sparse boosting again. Although (sparse) boosting algorithms has been widely studied in literature, to the best of our knowledge, there is no existing work about the two-step  
70 (sparse) boosting. Extending known boosting methods for such a complicated task is methodologically challenging and needs significant progress. The detailed development in this work makes a substantial contribution for such data analysis. Simulation results confirm the effectiveness of the two-step sparse boosting method.

The main contents of this paper are arranged as follows. In section 2, the varying-  
75 coefficient model for longitudinal data is formulated and a two-step sparse boosting algorithm is proposed. In section 3, simulation studies are conducted to demonstrate the validity of this two-stage method. In section 4, the performance of two-step sparse boosting method is evaluated by analyzing yeast cell cycle gene expression data. Concluding remarks are given in section 5.

## 80 **2. Methodology**

Varying-coefficient models are commonly used for examining the time-dependent effects of covariates on repeated measured responses. Such models can be applied for longitudinal data where the measurements within each of the subjects are correlated with each other. In this section, we propose a two-step sparse boosting method to  
85 conduct variable selection and estimation simultaneously for varying-coefficient model with longitudinal data.

## 2.1. Model and estimation

Let  $Y_{ij}$  be the continuous outcome for the  $j$ th measurement of individual  $i$  taken at time  $t_{ij} \in T$ , where  $T$  is the time interval on which the measurements are taken. Denote  $\mathbf{X}_{ij} = (X_{ij,1}, \dots, X_{ij,p-1})$  to be the corresponding  $(p-1)$ -dimensional covariate vector. We consider the varying-coefficient model which can capture the dynamical impacts of the covariates on the response variable:

$$Y_{ij} = \beta_0(t_{ij}) + \sum_{d=1}^{p-1} X_{ij,d} \beta_d(t_{ij}) + \varepsilon_{ij}, \quad i = 1, \dots, n, j = 1, \dots, n_i, \quad (1)$$

where  $\beta_0(\cdot), \beta_1(\cdot), \dots, \beta_{p-1}(\cdot)$  are the unknown smooth coefficient functions of time and  $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{in_i})^T, i = 1, \dots, n$  are multivariate error terms with mean zero. We assume errors are uncorrelated for different  $i$ , but components of  $\varepsilon_i$  are correlated with each other. Without loss of generality, we assume the longitudinal study is balanced in the following implementation, i.e.,  $t_{ij} = t_{kj}$ , and  $n_i = m$  for all  $i$ .

We next outline the estimation procedure. In the first step, we first ignore the within-subject correlation and estimate the coefficients by minimizing the following least squares loss function:

$$\sum_{i=1}^n \sum_{j=1}^m (Y_{ij} - \beta_0(t_{ij}) - \sum_{d=1}^{p-1} X_{ij,d} \beta_d(t_{ij}))^2. \quad (2)$$

Splines are well-known smoothing methods to approximate the unknown functions. We consider the B-spline basis to estimate the coefficient functions  $\beta_0(\cdot), \beta_1(\cdot), \dots, \beta_{p-1}(\cdot)$ . Denote  $B(\cdot) = (B_1(\cdot), \dots, B_L(\cdot))^T$  to be an equal-spaced B-spline basis of dimension  $L$ . Under certain smoothness conditions, function  $\beta_d(\cdot)$  can be approximated by

$$\beta_d(\cdot) \approx B^T(\cdot) \gamma_d, \quad d = 0, \dots, p-1, \quad (3)$$

where  $\gamma_d$  is a loading vector of length  $L$ . Then the least squares loss function (2) is close to

$$\sum_{i=1}^n \sum_{j=1}^m (Y_{ij} - B^T(t_{ij}) \gamma_0 - \sum_{d=1}^{p-1} X_{ij,d} B^T(t_{ij}) \gamma_d)^2. \quad (4)$$

We further denote  $Y_i = (Y_{i1}, \dots, Y_{im})^T$ ,  $Y = (Y_1^T, \dots, Y_n^T)^T$ ,  $X_{ij,0} = 1$ ,  $\tilde{\mathbf{X}}_{i,d} = (B(t_{i1})X_{i1,d}, \dots, B(t_{im})X_{im,d})^T$ ,  $\tilde{\mathbf{X}}_i = (\tilde{\mathbf{X}}_{i,0}, \dots, \tilde{\mathbf{X}}_{i,p-1})$ ,  $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}_1^T, \dots, \tilde{\mathbf{X}}_n^T)^T$  and  $\gamma =$

$(\gamma_0^T, \dots, \gamma_{p-1}^T)^T$ . Then the target function (4) can be expressed in the matrix format:

$$\sum_{i=1}^n (Y_i - \tilde{\mathbf{X}}_i \gamma)^T (Y_i - \tilde{\mathbf{X}}_i \gamma) \equiv (Y - \tilde{\mathbf{X}} \gamma)^T (Y - \tilde{\mathbf{X}} \gamma). \quad (5)$$

Denote  $\gamma^{[\widehat{K}_1]}$  to be the estimator of  $\gamma$  by sparse boosting with squared loss function (5) being loss function, where  $\widehat{K}_1$  is the estimated stopping iterations in this step. There  
 95 is no exact closed form for  $\gamma^{[\widehat{K}_1]}$  as it is based on an iterative algorithm. However in a computer implementation it can be evaluated very quickly. The detailed sparse boosting algorithm will be given in the next subsection.

The first step coefficient estimates are given by

$$\tilde{\beta}_d(t) = B^T(t) \gamma_d^{[\widehat{K}_1]}, \quad d = 0, \dots, p-1. \quad (6)$$

Write  $\widehat{\varepsilon}_i = Y_i - \tilde{\mathbf{X}}_i \gamma^{[\widehat{K}_1]}$ ,  $i = 1, \dots, n$ . We can estimate the  $m \times m$  covariance matrix  $\text{Cov}(Y_i) \equiv \Sigma$  by the following empirical estimator

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \widehat{\varepsilon}_i \widehat{\varepsilon}_i^T. \quad (7)$$

In the second step, we then take into account of the estimated correlation structure within repeated measurements to form the following weighted least squares loss function

$$\sum_{i=1}^n (Y_i - \tilde{\mathbf{X}}_i \gamma^*)^T \widehat{\Sigma}^{-1} (Y_i - \tilde{\mathbf{X}}_i \gamma^*) \equiv (Y - \tilde{\mathbf{X}} \gamma^*)^T W (Y - \tilde{\mathbf{X}} \gamma^*), \quad (8)$$

where  $W = \text{diag}(\widehat{\Sigma}^{-1}, \dots, \widehat{\Sigma}^{-1})$  is the estimated  $(n \times m) \times (n \times m)$  weight matrix. Denote  $\gamma^{*[\widehat{K}_2]}$  to be the estimator of  $\gamma^*$  by sparse boosting with weighted loss function (8) being the loss function, where  $\widehat{K}_2$  is the estimated stopping iterations in the second step. The second step coefficient estimates are given by

$$\widehat{\beta}_d(t) = B^T(t) \gamma_d^{*[\widehat{K}_2]}, \quad d = 0, \dots, p-1. \quad (9)$$

We could then obtain reliable estimates for the coefficient functions. We provide in the following subsection more details about how to use sparse boosting to get  $\gamma^{[\widehat{K}_1]}$  and  
 100  $\gamma^{*[\widehat{K}_2]}$ .

## 2.2. Two-step Sparse Boosting Techniques

The generic boosting algorithm is sketched in the Appendix. Sparse boosting can be viewed as iteratively pursuing gradient descending in function space using penalized empirical risk function which integrate squared loss and complexity of boosting measure. Since the trace of boosting operator is used as the degree of freedom in the respect of boosting by Bühlmann [38] and is the measure of the complexity of model [39, 40], we will adopt g-prior minimum description length (gMDL) [41], a combination of squared loss and the trace of boosting operator, as the penalized empirical risk function to estimate the update criterion in each iteration and the stopping criterion. gMDL can be written in the following form:

$$\begin{aligned} \text{gMDL}(RSS, \text{trace}(\mathcal{B})) &= \log(F) + \frac{\text{trace}(\mathcal{B})}{n \times m} \log\left(\frac{Y^T Y - RSS}{\text{trace}(\mathcal{B}) \times F}\right), \\ F &= \frac{RSS}{n \times m - \text{trace}(\mathcal{B})}, \end{aligned} \quad (10)$$

where  $RSS$  is the residual sum of squares and  $\mathcal{B}$  is the boosting operator. The model achieve shortest description of data will be chosen. We use it because it has a data driven penalty to avoid the selection of the tuning parameter.

We present the two-step sparse boosting approach more specifically. In the first step, the start value of  $\gamma$  is set to zero vector, i.e.  $\gamma^{[0]} = \mathbf{0}$ , and in each of the  $k_1$ th iteration ( $0 < k_1 \leq K_1$ , and  $K_1$  is the maximum number of iterations considered in the first step), we use the residual  $R^{[k_1]} = Y - \tilde{\mathbf{X}}\gamma^{[k_1-1]}$  in present iteration to fit each of the  $d$ th component  $\tilde{\mathbf{X}}_{\cdot,d} = (\tilde{\mathbf{X}}_{1,d}^T, \dots, \tilde{\mathbf{X}}_{n,d}^T)^T$ ,  $d = 0, \dots, p-1$  by treating all the within-subject observations uncorrelated. Then the fit denoted by  $\hat{\lambda}_d^{[k_1]}$  can be calculated by minimizing the squared loss function  $(R^{[k_1]} - \tilde{\mathbf{X}}_{\cdot,d}\lambda)^T (R^{[k_1]} - \tilde{\mathbf{X}}_{\cdot,d}\lambda)$  with respect to  $\lambda$ . Therefore, the least squares estimate is  $\hat{\lambda}_d^{[k_1]} = [(\tilde{\mathbf{X}}_{\cdot,d})^T (\tilde{\mathbf{X}}_{\cdot,d})]^{-1} (\tilde{\mathbf{X}}_{\cdot,d})^T R^{[k_1]}$ , the corresponding hat matrix is  $\mathcal{H}_d = (\tilde{\mathbf{X}}_{\cdot,d}) [(\tilde{\mathbf{X}}_{\cdot,d})^T (\tilde{\mathbf{X}}_{\cdot,d})]^{-1} (\tilde{\mathbf{X}}_{\cdot,d})^T$  and the residual sum of squares is  $RSS_d^{[k_1]} = (R^{[k_1]} - \tilde{\mathbf{X}}_{\cdot,d} \hat{\lambda}_d^{[k_1]})^T (R^{[k_1]} - \tilde{\mathbf{X}}_{\cdot,d} \hat{\lambda}_d^{[k_1]})$ . The chosen element  $\hat{s}_{k_1}$  is attained by:

$$\hat{s}_{k_1} = \text{argmin}_{0 \leq d \leq p-1} \text{gMDL}(RSS_d^{[k_1]}, \text{trace}(\mathcal{B}_d^{[k_1]})), \quad (11)$$

where  $\mathcal{B}_d^{[1]} = \mathcal{H}_d$  and  $\mathcal{B}_d^{[k_1]} = I - (I - \mathcal{H}_d)(I - \nu \mathcal{H}_{\hat{s}_{k_1-1}}) \dots (I - \nu \mathcal{H}_{\hat{s}_1})$  for  $k_1 > 1$  is the first step boosting operator for choosing  $d$ th element in the  $k_1$ th iteration. Hence,



there is an unique element  $\tilde{\mathbf{X}}_{\widehat{s}_{k_1}}$  to be selected at each iteration, and only the corresponding coefficient vector  $\gamma_{\widehat{s}_{k_1}}^{[k_1]}$  changes, i.e.,  $\gamma_{\widehat{s}_{k_1}}^{[k_1]} = \gamma_{\widehat{s}_{k_1}}^{[k_1-1]} + v\widehat{\lambda}_{\widehat{s}_{k_1}}^{[k_1]}$ , where  $v$  is the pre-specified step-size parameter. All the other  $\gamma_d^{[k_1]}$  for  $d \neq \widehat{s}_{k_1}$  keep unchanged. We repeat this procedure for  $K_1$  times and the number of iterations  $K_1$  can be estimated by

$$\widehat{K}_1 = \operatorname{argmin}_{1 \leq k_1 \leq K_1} \mathfrak{gMDL}(RSS_{\widehat{s}_{k_1}}^{[k_1]}, \operatorname{trace}(\mathcal{B}^{[k_1]})), \quad (12)$$

105 where  $\mathcal{B}^{[k_1]} = I - (I - v\mathcal{H}_{\widehat{s}_{k_1}}) \cdots (I - v\mathcal{H}_{\widehat{s}_1})$ .

From the first step of sparse boosting, we get the estimator of  $\gamma$  by  $\gamma^{[\widehat{K}_1]} = ((\gamma_0^{[\widehat{K}_1]})^T, \dots, (\gamma_{p-1}^{[\widehat{K}_1]})^T)^T$ . Then the weight matrix  $W$  can be easily obtained.

In the second step, we use sparse boosting again by incorporating the estimated correlation structure for the repeated measurements estimated in the first step. The start value of  $\gamma^*$  is set to be the coefficient estimator from the first step of sparse boosting, i.e.  $\gamma^{*[0]} = \gamma^{[\widehat{K}_1]}$ , and in each of the  $k_2$ th iteration ( $0 < k_2 \leq K_2$ , and  $K_2$  is the maximum number of iterations considered in the second step), we use the residual  $R^{*[k_2]} = Y - \tilde{\mathbf{X}}\gamma^{*[k_2-1]}$  in current iteration to fit each of the  $d$ th working component  $\tilde{\mathbf{X}}_{\cdot,d}, d = 0, \dots, p-1$  by incorporating the estimated within-subject correlation from the first step. Then the fit denoted by  $\widehat{\lambda}_d^{*[k_2]}$  can be attained by minimizing the weighted squared loss function  $(R^{*[k_2]} - \tilde{\mathbf{X}}_{\cdot,d}\lambda)^T W (R^{*[k_2]} - \tilde{\mathbf{X}}_{\cdot,d}\lambda)$  with respect to  $\lambda$ . Thus, the weighted least squares estimate is  $\widehat{\lambda}_d^{*[k_2]} = [(\tilde{\mathbf{X}}_{\cdot,d})^T W (\tilde{\mathbf{X}}_{\cdot,d})]^{-1} (\tilde{\mathbf{X}}_{\cdot,d})^T W R^{*[k_2]}$ , the corresponding hat matrix is  $\mathcal{H}_d^* = (\tilde{\mathbf{X}}_{\cdot,d}) [(\tilde{\mathbf{X}}_{\cdot,d})^T W (\tilde{\mathbf{X}}_{\cdot,d})]^{-1} (\tilde{\mathbf{X}}_{\cdot,d})^T W$  and the weighted residual sum of squares is  $RSS_d^{*[k_2]} = (R^{*[k_2]} - \tilde{\mathbf{X}}_{\cdot,d}\widehat{\lambda}_d^{*[k_2]})^T W (R^{*[k_2]} - \tilde{\mathbf{X}}_{\cdot,d}\widehat{\lambda}_d^{*[k_2]})$ . The chosen element  $\widehat{s}_{k_2}$  can be obtained by:

$$\widehat{s}_{k_2} = \operatorname{argmin}_{0 \leq d \leq p-1} \mathfrak{gMDL}(RSS_d^{*[k_2]}, \operatorname{trace}(\mathcal{B}_d^{*[k_2]})), \quad (13)$$

where  $\mathcal{B}_d^{*[1]} = I - (I - \mathcal{B}^{[\widehat{K}_1]})(I - \mathcal{H}_d^*)$  and  $\mathcal{B}_d^{*[k_2]} = I - (I - \mathcal{B}^{[\widehat{K}_1]})(I - \mathcal{H}_d^*)(I - v\mathcal{H}_{\widehat{s}_{k_2-1}}^*) \cdots (I - v\mathcal{H}_{\widehat{s}_1}^*)$  for  $k_2 > 1$  is the second step boosting operator for choosing  $d$ th element in the  $k_2$ th iteration. Thus, there is an unique element  $\tilde{\mathbf{X}}_{\widehat{s}_{k_2}}$  to be selected at each time, and only the corresponding coefficient vector  $\gamma_{\widehat{s}_{k_2}}^{*[k_2]}$  change, i.e.,  $\gamma_{\widehat{s}_{k_2}}^{*[k_2]} = \gamma_{\widehat{s}_{k_2}}^{*[k_2-1]} + v\widehat{\lambda}_{\widehat{s}_{k_2}}^{*[k_2]}$ . While all the other  $\gamma_d^{*[k_2]}$  for  $d \neq \widehat{s}_{k_2}$  remain the same. We

repeat this procedure for  $K_2$  times and the estimated stopping iterations  $\widehat{K}_2$  is

$$\widehat{K}_2 = \operatorname{argmin}_{1 \leq k_2 \leq K_2} \text{gMDL}(RSS_{\widehat{s}_{k_2}}^{*[k_2]}, \operatorname{trace}(\mathcal{B}^{*[k_2]})), \quad (14)$$

where  $\mathcal{B}^{*[k_2]} = I - (I - \mathcal{B}^{[\widehat{K}_1]})(I - \mathbf{v}\mathcal{H}_{\widehat{s}_{k_2}}^*) \cdots (I - \mathbf{v}\mathcal{H}_{\widehat{s}_1}^*)$ .

From the second step of sparse boosting, we arrive the estimator of  $\gamma^*$  by  $\gamma^{*[\widehat{K}_2]} = ((\gamma_0^{*[\widehat{K}_2]})^T, \dots, (\gamma_{p-1}^{*[\widehat{K}_2]})^T)^T$ . We can summarize the two-step sparse boosting algorithm for varying-coefficient model with longitudinal data in the following form:

---

### Two-step Sparse Boosting Algorithm

---

*Step I: Use sparse boosting to estimate covariance matrix.*

- a. Initialization. Let  $k_1 = 0$  and  $\gamma_0^{[k_1]} = \mathbf{0}, \dots, \gamma_{p-1}^{[k_1]} = \mathbf{0}$ .
- b. Increase  $k_1$  by 1. Calculate  $\widehat{s}_{k_1} = \operatorname{argmin}_{0 \leq d \leq p-1} \text{gMDL}(RSS_d^{[k_1]}, \operatorname{trace}(\mathcal{B}_d^{[k_1]}))$ , where  $\mathcal{B}_d^{[1]} = \mathcal{H}_d$  and  $\mathcal{B}_d^{[k_1]} = I - (I - \mathcal{H}_d)(I - \mathbf{v}\mathcal{H}_{\widehat{s}_{k_1-1}}) \cdots (I - \mathbf{v}\mathcal{H}_{\widehat{s}_1})$  for  $k_1 > 1$ .
- c. Update.  $\gamma_{\widehat{s}_{k_1}}^{[k_1]} = \gamma_{\widehat{s}_{k_1}}^{[k_1-1]}$  for  $d \neq \widehat{s}_{k_1}$  and  $\gamma_{\widehat{s}_{k_1}}^{[k_1]} = \gamma_{\widehat{s}_{k_1}}^{[k_1-1]} + \mathbf{v}\widehat{\lambda}_{\widehat{s}_{k_1}}^{[k_1]}$ , where  $\mathbf{v}$  is the step-size parameter.
- d. Iteration. Repeat step (b)-(c) for some large iteration number  $K_1$ .
- e. Stopping. The optimal iteration number can be taken as  $\widehat{K}_1 = \operatorname{argmin}_{1 \leq k_1 \leq K_1} \text{gMDL}(RSS_{\widehat{s}_{k_1}}^{[k_1]}, \operatorname{trace}(\mathcal{B}^{[k_1]}))$ , where  $\mathcal{B}^{[k_1]} = I - (I - \mathbf{v}\mathcal{H}_{\widehat{s}_{k_1}}) \cdots (I - \mathbf{v}\mathcal{H}_{\widehat{s}_1})$ .

Thus,  $\gamma^{[\widehat{K}_1]} = ((\gamma_0^{[\widehat{K}_1]})^T, \dots, (\gamma_{p-1}^{[\widehat{K}_1]})^T)^T$  is the first step estimator for  $\gamma$  from sparse boosting and  $\widetilde{\beta}_d(t) = B^T(t)\gamma_d^{[\widehat{K}_1]}$ ,  $d = 0, \dots, p-1$  are the varying coefficient estimates ignoring the within-subject correlation.  $\operatorname{Cov}(Y_i)$  can be estimated by

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (Y_i - \widetilde{\mathbf{X}}_i \gamma^{[\widehat{K}_1]})(Y_i - \widetilde{\mathbf{X}}_i \gamma^{[\widehat{K}_1]})^T.$$

115

*Step II: Use sparse boosting again by incorporating covariance matrix estimator.*

- a. Initialization. Let  $k_2 = 0$  and  $\gamma^{*[k_2]} = \gamma^{[\widehat{K}_1]}$ .
- b. Increase  $k_2$  by 1. Calculate  $\widehat{s}_{k_2} = \operatorname{argmin}_{0 \leq d \leq p-1} \text{gMDL}(RSS_d^{*[k_2]}, \operatorname{trace}(\mathcal{B}_d^{*[k_2]}))$ , where  $\mathcal{B}_d^{*[1]} = I - (I - \mathcal{B}^{[\widehat{K}_1]})(I - \mathcal{H}_d^*)$  and  $\mathcal{B}_d^{*[k_2]} = I - (I - \mathcal{B}^{[\widehat{K}_1]})(I - \mathcal{H}_d^*)(I -$

120  $\mathbf{v} \mathcal{H}_{\widehat{s}_{k_2-1}}^* \dots (I - \mathbf{v} \mathcal{H}_{\widehat{s}_1}^*)$  for  $k_2 > 1$ .

c. Update.  $\gamma_{\widehat{s}_{k_2}}^{*[k_2]} = \gamma_{\widehat{s}_{k_2}}^{*[k_2-1]}$  for  $d \neq \widehat{s}_{k_2}$  and  $\gamma_{\widehat{s}_{k_2}}^{*[k_2]} = \gamma_{\widehat{s}_{k_2}}^{*[k_2-1]} + \mathbf{v} \widehat{\lambda}_{\widehat{s}_{k_2}}^{*[k_2]}$ .

d. Iteration. Repeat step (b)-(c) for some large iteration number  $K_2$ .

e. Stopping. The optimal iteration number can be taken as  $\widehat{K}_2 = \operatorname{argmin}_{1 \leq k_2 \leq K_2} \text{gMDL}(RSS_{\widehat{s}_{k_2}}^{*[k_2]}, \text{trace}(\mathcal{B}^{*[k_2]}))$ , where  $\mathcal{B}^{*[k_2]} = I - (I - \mathcal{B}^{[\widehat{K}_1]})(I - \mathbf{v} \mathcal{H}_{\widehat{s}_{k_2}}^*) \dots (I - \mathbf{v} \mathcal{H}_{\widehat{s}_1}^*)$ .

125 Therefore,  $\gamma^{*[\widehat{K}_2]} = ((\gamma_0^{*[\widehat{K}_2]})^T, \dots, (\gamma_{p-1}^{*[\widehat{K}_2]})^T)^T$  and  $\widehat{\beta}_d(t) = B^T(t) \gamma_d^{*[\widehat{K}_2]}$ ,  $d = 0, \dots, p-1$  are the final estimator for  $\gamma^*$  and varying coefficient estimates by the two-step sparse boosting. The final estimate for  $Y$  is  $\widehat{Y} = \widetilde{\mathbf{X}} \gamma^{*[\widehat{K}_2]}$ .

The selection of step-size parameter  $\mathbf{v}$  is of minor importance [11]. Although smaller  $\mathbf{v}$  may produce more accurate estimation, it is also more computational intensive. We will use  $\mathbf{v} = 0.1$  since it is commonly used in literature [42].

### 3. Simulation

We conduct simulation studies to assess the performance of the proposed two-step sparse boosting algorithm. We compare the following four methods in their performance of variable selection and function estimation. (a) two-step boosting (introduced in Appendix, use squared loss for update criterion and gMDL for stopping criterion); (b) two-step sparse boosting; (c) two-step lasso (performs lasso regression in the first step to obtain estimated within-subject correlation structure using (7), and performs lasso regression in the second step by incorporating the estimated error correlation structure) and (d) two-step elastic net regression (similar as (c) with the elastic net mixing parameter 0.5). The tuning parameters in (c) and (d) are chosen by 5-fold cross-validation.

We consider the sparse model:

$$Y_{ij} = \beta_0(t_{ij}) + \sum_{d=1}^5 X_{ij,d} \beta_d(t_{ij}) + \varepsilon_{ij}, \quad i = 1, \dots, n, j = 1, \dots, m, \quad (15)$$

where  $\mathbf{X}_{ij} = (X_{ij,1}, \dots, X_{ij,p-1})$ ,  $i = 1, \dots, n, j = 1, \dots, m$  are i.i.d. from multivariate normal distribution with mean  $\mathbf{0}_{p-1}$ , and  $\text{Cov}(X_{ij,d}, X_{ij,l}) = 0.2^{|d-l|}$ . In each  $i$ th

145 group, the within subject error term  $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{im})^T$  are generated from normal distribution with mean  $\mathbf{0}_m$  and  $Cov(\varepsilon_{ij}, \varepsilon_{ik}) = \rho^{|j-k|}$ , where  $\rho$  is set to be 0.2 and 0.8. There are no correlation between groups. The time  $t'_{ij}$ s are generated independently from uniform distribution on  $[0, 1]$ . We consider  $n = 150, m = 4$  with  $p = 150, 300$ . The maximum number of boosting iterations in both steps are limited to  $K_1 = 500$  and  
150  $K_2 = 500$  respectively. The dimension of B-spline basis is set to 6 and 100 Monte Carlo simulations are conducted to produce reliable results. We examine the following four cases for nonparametric components:

**Case I** Let  $\beta_0(t) = -1.5$ ,  $\beta_1(t) = 1.2$ ,  $\beta_2(t) = 3t$ ,  $\beta_3(t) = t - t^2$ ,  $\beta_4(t) = t^3$  and  $\beta_5(t) = 2 \exp(-2t)$ .

155 **Case II** Let  $\beta_0(t) = -1.5t$ ,  $\beta_1(t) = t^2$ ,  $\beta_2(t) = 3t$ ,  $\beta_3(t) = t - t^2$ ,  $\beta_4(t) = t^3$  and  $\beta_5(t) = 2 \exp(-2t)$ .

**Case III** Let  $\beta_0(t) = \sin(2\pi t)$ ,  $\beta_1(t) = (1-t)^2$ ,  $\beta_2(t) = \exp(-(3t-1)^2) + \exp(-(4t-3)^2)$ ,  $\beta_3(t) = \sqrt{t}$ ,  $\beta_4(t) = 1-t$  and  $\beta_5(t) = 1 - \cos(2\pi t)$ .

**Case IV** Let  $\beta_0(t) = 3.5 \sin(2\pi t)$ ,  $\beta_1(t) = 5(1-t)^2$ ,  $\beta_2(t) = 3.5 \exp(-(3t-1)^2) + 3.5 \exp(-(4t-3)^2)$ ,  $\beta_3(t) = 3.5\sqrt{t}$ ,  $\beta_4(t) = 6-2t$  and  $\beta_5(t) = 2-3 \cos(4\pi t)$ .  
160

After 100 simulations, we summarized the performance of variable selection in Table 1 for the four considered cases. We reported the median of true positives (TF), the median of false positives (FP) as well as the median of model size (Size). We observe  
165 the following from Table 1: (1). In most cases, all methods produce true positives equal to the true number of important variables. However, in terms of sparsity, our proposed two-step sparse boosting method achieve best performance with zero false positives in most cases. Both penalization methods produce much more irrelevant covariates than boosting methods, with elastic net produce the most. (2). For two-step sparse  
170 boosting, variable selection results do not change much from step I to step II while for the other three methods, the false positives and thus model sizes from step I and step II are increasing. (3). When increasing the contribution of all components to response (from Case III to Case IV), the false positives and total model size are decreasing for

two-step boosting, but not for penalized approaches.

175 In Table 2, we reported the bias for each estimated coefficients, the average reduction  
of bias among all estimated coefficients from step I to step II and the bias of the esti-  
mated covariance matrix with  $p = 150$  gene expressions. The results for  $p = 300$  are  
similar, and are thus skip here. We observe the following from Table 2: (1). Two-  
step sparse boosting produces smallest bias for the estimated coefficients among the  
180 competing methods. (2). The refined estimates after taking into account of the within-  
subject correlation in general perform better than the initial estimates without consid-  
ering within-subject correlation since the four methods achieve reduction of bias under  
most settings, especially when the within-subject correlation is high. (3). When the  
within-subject correlation is 0.8, our proposed two-step sparse boosting performs best  
185 as it reduced the average bias by 12.54%-36.09% under different settings. (4). When  
the within-subject correlation is high ( $\rho = 0.8$ ), the reduction of bias from step I to step  
II are much bigger than those when within-subject correlation is low ( $\rho = 0.2$ ). This is  
intuitive because in the second step, we have taken into account of the within-subject  
correlation structure estimated from the first step. (5). The bias of estimated covariance  
190 matrix under smaller within-subject correlation are smaller than the those under larger  
within-subject correlation. (6). When the within-subject correlation is 0.8, the bias of  
the estimated covariance matrix for our proposed two-step sparse boosting is smallest  
among the four compared methods.

We also reported the mean integrated absolute error (MIAE) for each estimated coeffi-  
195 cients, the average reduction of MIAE among all estimated coefficients from step I to  
step II with 150 gene expressions in Appendix. They have the similar pattern as the  
bias reported in Table 2. In conclusion, the variable selection and functional coeffi-  
cients estimation results for two-step sparse boosting are quite satisfactory.

Table 1: Variable selection results in simulation.  $\rho$ : within-subject correlation; p: No. of gene expressions; Method: (a) two-step boosting; (b) two-step sparse boosting; (c) two-step lasso; and (d) two-step elastic net; TP: the median of true positives; FP: the median of false positives; Size: the median of model size; The values in the parentheses are the robust standard deviations.

Case	$\rho$	Methods	p = 150						p = 300					
			TP		FP		Size		TP		FP		Size	
			Step I	Step II	Step I	Step II	Step I	Step II	Step I	Step II	Step I	Step II	Step I	Step II
I	0.2	(a)	6 (0.20)	6 (0.14)	8 (2.80)	10 (3.18)	14 (2.82)	16 (3.17)	6 (0.20)	6 (0.00)	10 (4.78)	12 (5.65)	16 (4.77)	18 (5.65)
		(b)	5 (0.41)	5 (0.42)	0 (0.00)	0 (0.00)	5 (0.41)	5 (0.42)	5 (0.41)	5 (0.41)	0 (0.00)	0 (0.00)	5 (0.41)	5 (0.41)
		(c)	6 (0.00)	6 (0.00)	38 (13.74)	50 (13.62)	44 (13.74)	56 (13.62)	6 (0.00)	6 (0.10)	54 (15.75)	66 (16.63)	60 (15.75)	72 (16.63)
		(d)	6 (0.00)	6 (0.00)	48 (12.62)	58 (12.10)	54 (12.62)	64 (12.10)	6 (0.00)	6 (0.10)	61 (18.02)	77 (19.59)	67 (18.02)	83 (19.59)
	0.8	(a)	6 (0.23)	6 (0.00)	8 (3.21)	11 (3.82)	14 (3.25)	17 (3.82)	6 (0.20)	6 (0.00)	10 (4.27)	14 (5.49)	16 (4.26)	20 (5.49)
		(b)	5 (0.42)	6 (0.47)	0 (0.00)	0 (0.00)	5 (0.42)	6 (0.47)	5 (0.38)	6 (0.50)	0 (0.00)	0 (0.00)	5 (0.38)	6 (0.50)
		(c)	6 (0.00)	6 (0.00)	41 (11.77)	60 (11.21)	47 (11.77)	66 (11.21)	6 (0.00)	6 (0.00)	47 (16.34)	74 (14.90)	53 (16.34)	80 (14.90)
		(d)	6 (0.00)	6 (0.00)	48 (15.30)	67 (13.21)	54 (13.30)	73 (13.21)	6 (0.00)	6 (0.00)	57 (19.01)	85 (18.35)	63 (19.01)	91 (18.35)
II	0.2	(a)	6 (0.17)	6 (0.17)	10 (3.17)	13 (3.56)	16 (3.24)	19 (3.63)	6 (0.20)	6 (0.20)	14 (5.47)	17 (5.75)	20 (5.45)	23 (5.75)
		(b)	5 (0.45)	6 (0.40)	0 (0.00)	0 (0.00)	5 (0.45)	6 (0.40)	5 (0.47)	6 (0.28)	0 (0.00)	0 (0.00)	5 (0.48)	6 (0.28)
		(c)	6 (0.00)	6 (0.00)	30 (12.63)	38 (12.32)	36 (12.63)	44 (12.32)	6 (0.00)	6 (0.00)	34 (19.42)	45 (20.90)	40 (19.42)	51 (20.90)
		(d)	6 (0.00)	6 (0.00)	37 (12.60)	44 (12.41)	43 (12.60)	50 (12.41)	6 (0.00)	6 (0.00)	44 (20.20)	58 (21.01)	50 (20.20)	64 (21.07)
	0.8	(a)	6 (0.00)	6 (0.00)	10 (3.51)	14 (4.24)	16 (3.51)	20 (4.24)	6 (0.00)	6 (0.00)	16 (4.34)	22 (5.88)	22 (4.31)	28 (5.88)
		(b)	5 (0.50)	6 (0.39)	0 (0.00)	0 (0.00)	5 (0.50)	6 (0.39)	5 (0.49)	6 (0.33)	0 (0.00)	0 (0.00)	5 (0.49)	6 (0.33)
		(c)	6 (0.00)	6 (0.00)	28 (14.74)	45 (12.95)	34 (14.74)	51 (12.95)	6 (0.00)	6 (0.00)	36 (19.80)	59 (18.64)	42 (19.80)	65 (18.64)
		(d)	6 (0.00)	6 (0.00)	34 (12.93)	52 (11.13)	40 (12.93)	58 (11.13)	6 (0.00)	6 (0.00)	44 (22.26)	71 (20.66)	50 (22.26)	77 (20.66)
III	0.2	(a)	6 (0.00)	6 (0.00)	11 (3.59)	13 (3.87)	17 (3.59)	19 (3.58)	6 (0.00)	6 (0.00)	20 (6.51)	24 (7.47)	26 (6.51)	30 (7.47)
		(b)	6 (0.00)	6 (0.00)	0 (0.00)	0 (0.00)	6 (0.00)	6 (0.00)	6 (0.00)	6 (0.00)	0 (0.00)	0 (0.21)	6 (0.00)	6 (0.21)
		(c)	6 (0.00)	6 (0.00)	28 (13.57)	36 (13.82)	34 (13.57)	42 (13.82)	6 (0.00)	6 (0.00)	36 (22.03)	45 (21.79)	42 (22.03)	51 (21.79)
		(d)	6 (0.00)	6 (0.00)	34 (14.54)	39 (14.35)	40 (14.54)	45 (14.35)	6 (0.00)	6 (0.00)	42 (20.46)	53 (22.13)	48 (20.46)	59 (22.13)
	0.8	(a)	6 (0.00)	6 (0.00)	12 (4.67)	17 (4.63)	18 (4.67)	23 (4.63)	6 (0.00)	6 (0.00)	20 (5.70)	26 (6.56)	26 (5.70)	32 (6.56)
		(b)	6 (0.00)	6 (0.17)	0 (0.17)	0 (0.17)	6 (0.17)	6 (0.21)	6 (0.00)	6 (0.00)	0 (0.45)	0 (0.45)	6 (0.45)	6 (0.45)
		(c)	6 (0.00)	6 (0.00)	27 (12.43)	42 (11.72)	33 (12.43)	48 (11.72)	6 (0.00)	6 (0.00)	40 (23.19)	56 (21.71)	46 (23.19)	62 (21.71)
		(d)	6 (0.00)	6 (0.00)	34 (12.42)	49 (10.79)	40 (12.42)	55 (10.79)	6 (0.00)	6 (0.00)	46 (22.02)	64 (20.91)	52 (22.02)	70 (20.91)
IV	0.2	(a)	6 (0.00)	6 (0.00)	2 (1.35)	4 (1.54)	8 (1.35)	10 (1.54)	6 (0.00)	6 (0.00)	3 (1.63)	5 (2.03)	9 (1.63)	11 (2.03)
		(b)	6 (0.00)	6 (0.00)	0 (0.00)	0 (0.00)	6 (0.00)	6 (0.00)	6 (0.00)	6 (0.00)	0 (0.00)	0 (0.00)	6 (0.00)	6 (0.00)
		(c)	6 (0.00)	6 (0.00)	32 (14.15)	46 (15.63)	38 (14.15)	52 (15.63)	6 (0.00)	6 (0.00)	37 (17.88)	54 (20.92)	43 (17.88)	60 (20.92)
		(d)	6 (0.00)	6 (0.00)	37 (12.65)	50 (14.35)	43 (12.65)	56 (14.35)	6 (0.00)	6 (0.00)	41 (19.25)	59 (20.87)	47 (19.25)	65 (20.87)
	0.8	(a)	6 (0.00)	6 (0.00)	2 (1.38)	4 (1.92)	8 (1.38)	10 (1.92)	6 (0.00)	6 (0.00)	4 (1.88)	6 (2.92)	10 (1.88)	12 (2.92)
		(b)	6 (0.00)	6 (0.00)	0 (0.00)	0 (0.00)	6 (0.00)	6 (0.00)	6 (0.00)	6 (0.00)	0 (0.00)	0 (0.00)	6 (0.00)	6 (0.00)
		(c)	6 (0.00)	6 (0.00)	34 (13.09)	44 (14.57)	40 (13.09)	50 (14.57)	6 (0.00)	6 (0.00)	36 (18.28)	55 (22.28)	42 (18.28)	61 (22.28)
		(d)	6 (0.00)	6 (0.00)	38 (13.05)	49 (12.83)	44 (13.05)	55 (12.83)	6 (0.00)	6 (0.00)	42 (20.00)	59 (22.72)	48 (20.00)	65 (22.72)

Table 2: The bias (BIAS) for each estimated coefficients, the average reduction of bias in percentage among all estimated coefficients from step I to step II (RED) and the bias of the estimated covariance matrix in simulation with 150 gene expressions;  $\rho$ : within-subject correlation; Method: (a) two-step boosting; (b) two-step sparse boosting; (c) two-step lasso; and (d) two-step elastic net.

Case	$\rho$	Methods	BIAS( $\hat{\beta}_0(t)$ )		BIAS( $\hat{\beta}_1(t)$ )		BIAS( $\hat{\beta}_2(t)$ )		BIAS( $\hat{\beta}_3(t)$ )		BIAS( $\hat{\beta}_4(t)$ )		BIAS( $\hat{\beta}_5(t)$ )		RED	BIAS( $\hat{\Sigma}$ )
			Step I	Step II	Step I	Step II	Step I	Step II	Step I	Step II	Step I	Step II	Step I	Step II		
I	0.2	(a)	-0.172	-0.171	0.141	0.137	0.066	0.062	0.108	0.102	0.067	0.063	0.141	0.139	3.74	-0.039
		(b)	-0.058	-0.056	0.078	0.070	0.013	0.006	-0.112	-0.097	0.079	0.072	0.061	0.054	16.88	-0.024
		(c)	-0.151	-0.160	0.148	0.145	0.100	0.097	0.060	0.059	0.060	0.060	0.170	0.166	0.51	-0.013
		(d)	-0.159	-0.170	0.155	0.153	0.113	0.110	0.056	0.055	0.059	0.059	0.176	0.173	0.09	-0.012
	0.8	(a)	-0.204	-0.195	0.138	0.093	0.069	0.043	0.081	0.065	0.059	0.039	0.149	0.116	25.08	0.140
		(b)	-0.079	-0.077	0.066	0.034	-0.020	-0.022	0.151	0.139	0.066	0.030	0.069	0.051	21.60	0.137
		(c)	-0.152	-0.270	0.150	0.113	0.097	0.067	0.062	0.048	0.063	0.044	0.169	0.134	8.57	0.213
		(d)	-0.160	-0.282	0.157	0.119	0.111	0.077	0.058	0.045	0.062	0.044	0.177	0.140	8.49	0.227
II	0.2	(a)	-0.134	-0.133	0.098	0.092	0.068	0.067	0.073	0.070	0.083	0.079	0.163	0.158	3.39	-0.014
		(b)	-0.061	-0.055	0.100	0.092	-0.033	-0.037	0.137	0.135	0.094	0.088	0.082	0.076	3.48	-0.019
		(c)	-0.108	-0.111	0.080	0.078	0.106	0.104	0.061	0.061	0.061	0.060	0.162	0.159	0.85	0.004
		(d)	-0.112	-0.123	-0.079	-0.075	0.125	0.122	0.058	0.056	0.061	0.059	0.173	0.168	1.21	0.008
	0.8	(a)	-0.120	-0.082	0.081	0.051	0.054	0.023	0.091	0.060	0.070	0.041	0.136	0.105	37.40	0.180
		(b)	-0.040	-0.009	0.086	0.053	-0.057	-0.049	0.154	0.115	0.076	0.038	0.053	0.047	36.09	0.149
		(c)	-0.098	-0.154	0.086	0.061	0.109	0.070	0.065	0.049	0.062	0.044	0.162	0.125	14.03	0.211
		(d)	-0.100	-0.155	0.083	0.059	0.127	0.084	0.060	0.046	0.061	0.044	0.170	0.132	13.55	0.222
III	0.2	(a)	0.009	0.009	0.107	0.102	0.085	0.080	0.090	0.087	0.080	0.076	0.093	0.091	3.51	-0.015
		(b)	0.010	0.011	0.103	0.096	0.023	0.013	0.061	0.059	0.074	0.069	0.015	0.009	15.05	-0.022
		(c)	-0.005	-0.006	0.094	0.090	0.110	0.108	0.094	0.090	0.106	0.104	0.082	0.079	-0.69	0.031
		(d)	-0.005	-0.006	0.094	0.091	0.116	0.114	0.097	0.093	0.106	0.104	0.089	0.085	-0.76	0.032
	0.8	(a)	-0.002	-0.001	0.106	0.081	0.089	0.056	0.070	0.041	0.109	0.083	0.079	0.047	36.08	0.226
		(b)	-0.002	-0.002	0.095	0.068	0.031	0.016	0.034	0.010	0.097	0.070	0.001	-0.002	12.54	0.193
		(c)	-0.003	-0.006	0.097	0.073	0.105	0.071	0.100	0.066	0.099	0.078	0.077	0.053	7.25	0.246
		(d)	-0.003	-0.005	0.097	0.074	0.111	0.077	0.103	0.068	0.100	0.079	0.085	0.060	12.01	0.252
IV	0.2	(a)	-0.014	-0.014	0.220	0.213	0.164	0.149	0.144	0.136	0.273	0.278	0.047	0.035	6.93	-0.130
		(b)	-0.016	-0.014	0.222	0.184	0.132	0.063	0.154	0.099	0.238	0.181	0.033	-0.016	32.18	-0.120
		(c)	0.004	0.004	0.205	0.206	0.162	0.162	0.110	0.106	0.282	0.283	-0.048	-0.047	0.81	-0.135
		(d)	0.005	0.005	0.211	0.214	0.170	0.170	0.109	0.105	0.301	0.302	-0.045	-0.046	-0.05	-0.132
	0.8	(a)	0.001	0.002	0.219	0.199	0.143	0.114	0.122	0.087	0.286	0.256	0.044	0.021	3.48	-0.152
		(b)	0.003	0.004	0.210	0.161	0.123	0.059	0.132	0.059	0.250	0.185	0.033	-0.006	34.19	-0.150
		(c)	0.004	0.004	0.196	0.196	0.162	0.147	0.118	0.110	0.278	0.273	-0.044	-0.045	2.59	-0.166
		(d)	0.004	0.004	0.201	0.202	0.168	0.155	0.117	0.110	0.296	0.292	-0.042	-0.041	2.83	-0.161

#### 4. Real Data Analysis

200 The cell cycle is one of the life's most important processes by which cells grow, replicate their DNA, segregate their chromosomes, and divide into daughter cells. Therefore, identification of cell cycle regulated genes has become greatly important. Using a model-based approach, Luan and Li [43] identified  $n = 297$  cell cycle regulated genes based on the  $\alpha$ -factor synchronization experiments. All gene expression levels were  
205 measured at  $m = 18$  time points covering two cell-cycle periods. Using the same subset of the original data as in [44], we include total  $p = 96$  transcriptional factors (TFs) as covariates in the following analysis. Wei, Huang and Li [45] confirmed that the effects of the TFs on gene expression levels are time-dependent. Our goal is to determine the key TFs involved in gene regulation based on these 297 cell cycle regulated genes.

210 Firstly, we use the independence screening approach by  $l^2$ -norm [27] to screen out the unimportant covariates at first step. As suggested by [34, 27, 46], we keep the top  $\lfloor \frac{n}{\log n} \rfloor$  variables in the ranked list and screen out the rest of the covariates. 51 variables are kept in the screening step.

Note that through variable screening, there still remains a significant number of false  
215 positives. Therefore, we further adopt variable selection mechanisms to select the important variables. Besides two-step boosting and two-step sparse boosting which incorporate the within-subject correlation in the second step, we also consider one-step boosting and one-step sparse boosting which do not incorporate the within-subject correlation for better comparison. In other words, the one-step (sparse) boosting method  
220 is just the first step of the two-step (sparse) boosting method. In addition, we also consider some two step penalized approaches: two-step lasso, two-step adaptive lasso and two-step elastic net (the elastic net mixing parameter 0.5). Specifically, each of them perform penalized regression twice, and the second step should incorporate the correlation structure estimated from the first step and minimize the weighted loss function  
225 plus penalty. The three penalized methods are implemented by **glmnet** function in R.

The maximum number of boosting iterations in one-step (sparse) boosting and in both steps of two-step (sparse) boosting are all set to 500. The tuning parameters in penal-



ized methods are chosen by 5-fold cross-validation. The dimension of B-spline basis is set to 6. The performance of variable selection, estimation and prediction are presented in Table 3. We report the number of genes selected (No.), the number of overlapping TFs selected by the corresponding method and the two-step sparse boosting (Overlap), In-sample prediction error (ISPE) and 5-fold cross-validation out-of-sample prediction error (OSPE). We observe the following from Table 3: (1). Boosting methods select sparser model than the penalized approaches. (2). Sparse boosting produces sparser model and smaller estimation and prediction error than boosting. (3). Two-step (sparse) boosting methods perform better than one-step (sparse) boosting methods with smaller estimation and prediction errors. (4). Two-step sparse boosting methods produces the most sparse model, achieves the smallest ISPE and OSPE among all methods. (5). The number of overlapping TFs between two-step sparse boosting and each of the other methods are all quite close to 20 (the number of TFs selected by two-step sparse boosting), indicating a high correlation between two-step sparse boosting and each of the other methods. In summary, our two-step sparse boosting procedure performs very well in terms of variable selection, estimation and prediction and can provide meaningful information in identification of the important TFs that play roles in the network of regulations.

Applying two-step sparse boosting algorithm for variable selection, the following 20 TFs are selected: HIR2, HIR1, FKH2, NDD1, SWI4, SWI5, SKN7, MCM1, SMP1, PHD1, MBP1, ABF1, INO4, CBF1, RLM1, SWI6, MET4, CIN5, MTH1 and FKH1. Among the selected 20 TFs, 11 of them (FKH2, NDD1, SWI4, SWI5, SKN7, MCM1, MBP1, ABF1, CBF1, SWI6 and FKH1) are known and experimentally verified cell cycle related TFs reported in [44], 3 of them (HIR2, HIR1 and MET4) were identified by [47, 34], 2 of them (RLM1 and CIN5) were identified by [48] and 1 of them (SMP1) was identified by [34]. The identified TFs also include many pairs of cooperative or synergistic pairs of TFs involved in the yeast cycle process reported in the literature. For example, HIR1-HIR2 [49], SWI4-SWI6 [50], FKH2-NDD1 [51], SWI5-SMP1 [52], MCM1-FKH2 [53], MBP1-SWI6 [50], FKH1-FKH2 [53, 54] and RLM1-SWI5 [47]. Next, we apply the structure identification procedure proposed in [55] to identify the constant coefficients. Among the 20

TFs, INO4, CBF1, RLM1, SWI6, MET4, CIN5, MTH1 and FKH1 have constant coefficients and the other 12 TFs have varying coefficients. Among the 12 TFs with varying coefficients, FKH2, NDD1 and SWI5 were also identified as important TFs in [34, 36, 44]; HIR2, HIR1, SMP1 and ABF1 were also identified as important TFs in [34, 44]; MCM1 was also identified as important TFs in [44, 36]; SWI4, SKN7 and ABF1 were also identified as important TFs in [44]. Of the 8 TFs with constant coefficients, RLM1 was identified before with constant coefficient in [34]; SWI6 and MET4 were also identified by [34] but their effects were estimated as varying coefficients instead of constant ones. Finally, by fitting the 20 TFs in the semi-varying coefficient model, we obtain the refined estimators with their standard errors for the 8 constant coefficients. One-step sparse boosting selects the same set of 20 TFs as two-step sparse boosting. However, by applying structure identification procedure, 5 TFs have constant coefficients and the other 15 TFs have varying coefficients. The difference comes from the fact that for the two-step sparse boosting, the log-likelihood function takes into account of the within-subject correlation structure estimated in the first step and thus weighted loss function is used in the log-likelihood function while for the one-step sparse boosting, independent observations within subjects is assumed. We report the constant coefficient estimators with their standard errors from one-step and two-step sparse boosting in Table 4. The refined estimated transcriptional effects of the selected TFs with varying coefficients along with their 95% confidence band for one-step and two-step sparse boosting are plotted in Figure 1 and 2 respectively, almost all showing time-dependent effects of these TFs on gene expression levels.

Table 3: Yeast Cell Cycle example: variable selection, estimation and prediction performance. No.: Number of genes selected; Overlap: Number of overlapping TFs selected by the corresponding method and two-step sparse boosting; ISPE: In-sample prediction error  $\frac{1}{mn} \sum_{i=1}^n \sum_{j=1}^m (Y_{ij} - \hat{Y}_{ij})^2$ ; OSPE: 5-fold cross-validation out-of-sample prediction error.

Method	No.	Overlap	ISPE	OSPE
One-step boosting	23	20	0.205	0.285
One-step sparse boosting	20	20	0.204	0.273
Two-step boosting	27	20	0.204	0.275
Two-step sparse boosting	20	-	0.201	0.253
Two-step lasso	31	15	0.268	0.277
Two-step adaptive lasso	30	18	0.266	0.470
Two-step elastic net	50	20	0.254	0.272

Table 4: Yeast Cell Cycle example: constant coefficients estimators with standard errors by one-step and two-step sparse boosting. TF: transcriptional factors; SE: 500 bootstraps standard error.

TF	One-step sparse boosting		Two-step sparse boosting	
	Estimation	SE	Estimation	SE
INO4	-0.0022	0.0066	-0.0023	0.0072
CBF1	0.0030	0.0064	0.0033	0.0063
RLM1	-0.0053	0.0101	-0.0052	0.0099
SWI6	-0.0210	0.0121	-0.0211	0.0114
MET4	-0.0186	0.0117	-0.0187	0.0113
CIN5	-	-	0.0015	0.0083
MTH1	-	-	-0.0001	0.0083
FKH1	-	-	0.0004	0.0067

Figure 1: Estimated transcriptional effects with their 95% confidence bands based on 500 bootstraps for the selected TFs with varying coefficients in the final semivarying-coefficient model by one-step sparse boosting.

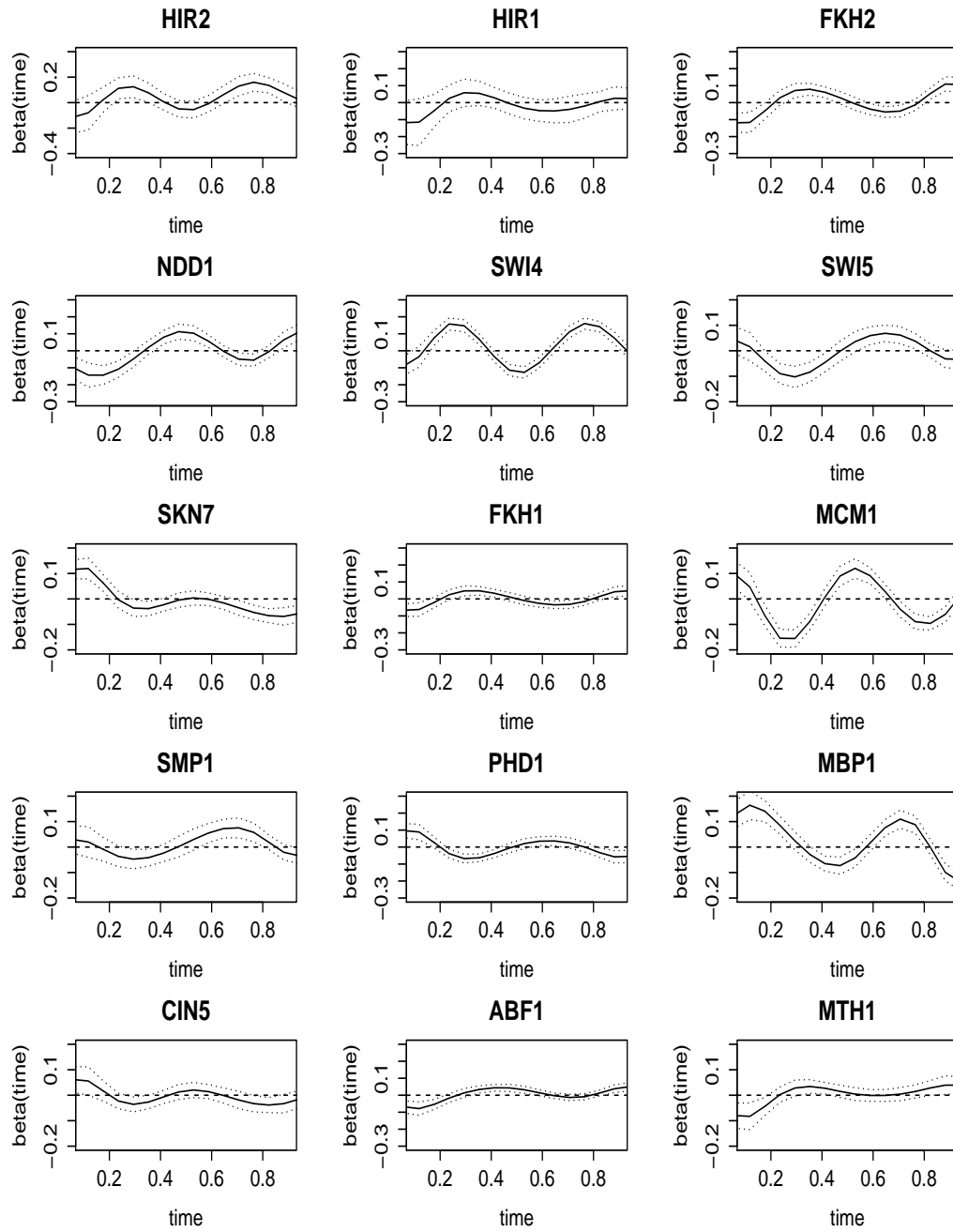
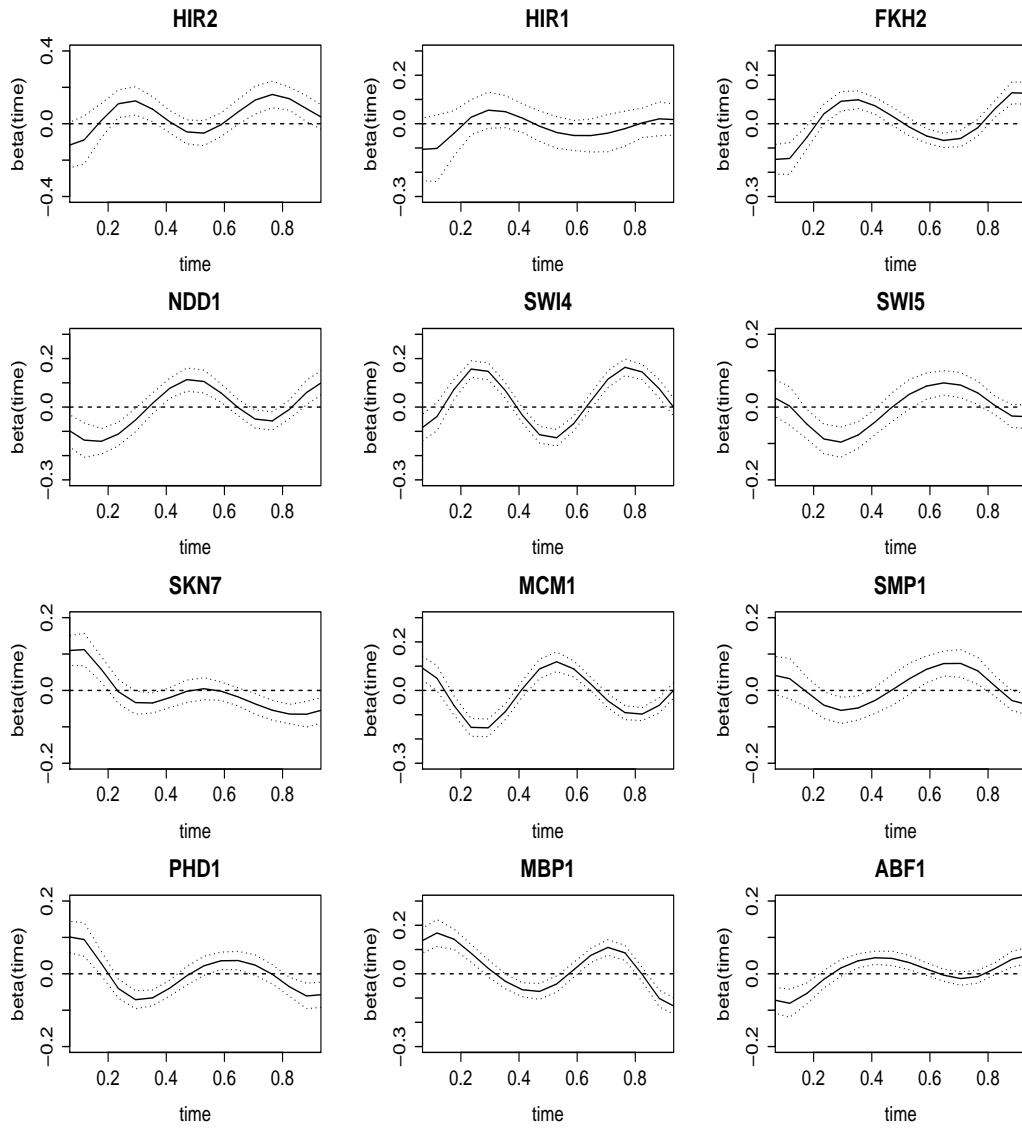


Figure 2: Estimated transcriptional effects with their 95% confidence bands based on 500 bootstraps for the selected TFs with varying coefficients in the final semivarying-coefficient model by two-step sparse boosting.



## 280 5. Concluding Remarks

Motivated by a biomedical study for identifying the TFs that can explain the variation of gene expression over time, we propose a two-step sparse boosting procedure for nonparametric varying-coefficient models. It can simultaneously perform variable selection and estimation of functional coefficients. Our methods are illustrated by analyzing repeatedly measured longitudinal data. The numerical studies indicate that our proposed method can effectively select the important covariates and accurately estimate the smooth coefficient functions, and the real data analysis also demonstrate the usefulness of our proposed method in identifying the key TFs in gene regulation based on cell cycle regulated genes.

290 The simulation study and real data analysis shows that the two-step (sparse) boosting is superior to one-step (sparse) boosting. By incorporating the estimated error structure from first step, the performance of variable selection and model estimation is improved. Moreover, the two-step sparse boosting is better than two-step boosting. They only differ regard to the update criterion in each iteration of the algorithm (RSS for two-step boosting vs. gMDL for two-step sparse boosting). The gMDL model selection criterion, similar to AIC and BIC, uses a data driven penalty to the squared loss and consequently achieves sparser model by not recruiting irrelevant covariates in the model. The numerical results also indicate that our proposed method works better than the penalized methods such as lasso and elastic net and our two-step sparse boosting produces smallest bias among all compared methods. Although regularization methods are not unbiased, the bias and variance trade-off should be balanced by choosing appropriate regularization parameters [56]. Incurring a slight bias in coefficients would yields more stable estimates to outliers and over-fitting. Thus, choosing the optimal tuning parameters become an important question of interest. We use cross-validation as it is the default method for **glmnet** function. In addition, since it makes greater use of the available data [57], cross-validation is used in a lot of studies [58, 59] and achieved good results. However, since we use gMDL as the model selection criterion, these might cause large parts of the differences between boosting and penalized approaches observed in the results.

310 Yue, Li and Ma [55] proposed the sparse boosting approach to conduct model selection and estimation for the right censored survival data with varying coefficients. This paper extends their sparse boosting methods to a more complicated version using a two-step algorithm. Because the data complexity, we require an initial step to obtain covariance estimates. The advantage of incorporating within-subject correlation in our proposed  
315 two-step sparse boosting procedure is demonstrated by Wang [60], who proposed a kernel smoothing method and showed that by accounting for within-subject correlation, the variance of the proposed method is uniformly smaller than the most efficient working independence approach. The calculation of the loss functions and the stopping criteria thus differs from [55]. Our methods are most suitable for correlated data such  
320 as the longitudinal measurements, clustered observations or multi-level structures [61], [62] [63].

When the number of covariates  $p$  and the number of repeated measurement  $m$  are both smaller than sample size  $n$ , the traditional covariance matrix estimator, the sample covariance matrix, is known to be unbiased and invertible [64]. Under our setting, we  
325 consider the number of covariates  $p$  is larger than the sample size  $n$  and the number of repeated measurement  $m$  is moderate in size. This is also different from the situation that the number of unknown parameters inside the covariance matrix is larger than the sample size while the number of covariates is moderate in size. In the latter case, least squares can be used to estimate the residuals [65], while the first step of  
330 our proposed two-step boosting procedure uses biased regression estimates affected by shrinkage. When the number of repeated measurement  $m$  is larger than sample size  $n$ , the empirical covariance matrix estimator becomes less useful or even degenerate [66]. For example, to study the multivariate two-sample testing problem for high dimensional data, alternative tests have been proposed by Chen and Qin [67], Biswas  
335 and Ghosh [68] and Mondal et al. [69] instead of using the Hotelling  $T^2$  test since it is not defined as the sample covariance matrix may not be invertible. In order to overcome the difficulty, certain covariance matrix structure is usually imposed. Some popular covariance matrix structures adopted in literatures including sparse covariance matrix through thresholding [70], bandable covariance matrices through banding [71],



340 tapering [72] or block banding [73], sparse inverse covariance matrix via graphical  
lasso [74], graphical Dantzig [75] and CLIME [76]. The overlook of the additional  
structural assumptions could be a weakness of our proposed procedure and the influ-  
ence of dimensionality on the estimation of covariance matrices should be taken into  
consideration in future studies.

345 We also acknowledge the following limitations. Firstly, although we did find in our  
simulation and real data example that two-step sparse boosting outperforms two-step  
boosting under various criteria, we have no theoretical guarantee. In fact Bühlmann  
and Yu [11] commented that there is no general superiority of sparse boosting over  
boosting. Secondly, the number of repeated measurements considered in the numerical  
350 studies are relatively small and may not be applicable to frequent observations such as  
financial time series data. Finally, even though we found success using our model to  
study the impact of different TFs on cell cycle, further application of the same model  
to other high-dimensional data sets requires validation and evaluation.

### **Acknowledgments**

355 We thank David Nott for helpful comments that improve the clarity of the presentation.  
The work was partly supported by Academic Research Funds R-155-000-195-114 and  
a Tier 2 MOE grant in Singapore.

## Appendix

Given  $(X_i, Y_i), i = 1, \dots, n$ , where  $X_i \in \mathbb{R}^p$  is a  $p$ -dimensional covariate vector and  $Y_i$   
360 a one-dimensional response variable.  $(X_i, Y_i)$  are usually assumed to be independent  
sample. Let  $L(Y, f)$  denote the loss function, our purpose is to estimate the unknown  
function  $f^*(\cdot) = \operatorname{argmin}_{f(\cdot)} E(L(Y, f(X)))$ . Estimate  $f^*(\cdot)$  can be done by boosting  
which descends the empirical risk  $n^{-1} \sum_{i=1}^n L(Y_i, f(X_i))$  via steepest gradient descent in  
function space, where the function space is provided by base learners. We summarize  
365 a generic boosting algorithm [77, 42] as follows:

---

### Generic Boosting Algorithm

---

- a.** Initialization. Let  $k = 0$  and let  $f^{[k]} = \bar{Y}$ .
- 370 **b.** Iteration. Increase  $k$  by 1. Compute negative gradient of the loss function  $-\frac{\partial L(Y_i, f)}{\partial f}$  and evaluate at  $f^{[k-1]}(X_i)$ , i.e.  $R_i^{[k]} = -\frac{\partial L(Y_i, f)}{\partial f}|_{f=f^{[k-1]}(X_i)}$  for  $i = 1, \dots, n$ . Fit negative gradient vector  $(R_1^{[k]}, \dots, R_n^{[k]})^T$  by the base learner, i.e.  $(X_i, R_i^{[k]})_{i=1}^n \xrightarrow{\text{base procedure}} \hat{g}^{[k]}(\cdot)$ .
- c.** Update.  $f^{[k]}(X) = f^{[k-1]}(X) + \nu \hat{g}^{[k]}(\cdot)$ , where  $0 < \nu \leq 1$  is the pre-defined step-size  
375 parameter.
- d.** Iteration. Repeat step (b)-(c) until  $k = K$  for some stopping criterion  $K$ .

For  $L_2$ Boosting, the squared loss function is a special case of generic boosting algorithm with  $L = (Y_i - f(X_i))^2/2$  and  $R_i^{[k]} = Y_i - f^{[k-1]}(X_i)$  in iteration  $k$ .

Table 5: The mean integrated absolute error (MIAE) for each estimated coefficients, the average reduction of MIAE in percentage among all estimated coefficients from step I to step II (RED) in simulation with 150 gene expressions;  $\rho$ : within-subject correlation; Method: (a) two-step boosting; (b) two-step sparse boosting; (c) two-step lasso; and (d) two-step elastic net.

Case	$\rho$	Methods	MIAE( $\hat{\beta}_0(t)$ )		MIAE( $\hat{\beta}_1(t)$ )		MIAE( $\hat{\beta}_2(t)$ )		MIAE( $\hat{\beta}_3(t)$ )		MIAE( $\hat{\beta}_4(t)$ )		MIAE( $\hat{\beta}_5(t)$ )		RED
			Step I	Step II	Step I	Step II	Step I	Step II	Step I	Step II	Step I	Step II	Step I	Step II	
I	0.2	(a)	0.202	0.200	0.168	0.166	0.126	0.123	0.106	0.104	0.120	0.117	0.209	0.206	1.73
		(b)	0.155	0.154	0.143	0.141	0.085	0.085	0.131	0.119	0.122	0.117	0.194	0.193	2.64
		(c)	0.193	0.201	0.177	0.176	0.162	0.160	0.185	0.184	0.098	0.098	0.210	0.208	-0.14
		(d)	0.194	0.200	0.178	0.176	0.171	0.167	0.189	0.186	0.098	0.097	0.209	0.207	-0.20
	0.8	(a)	0.212	0.205	0.178	0.149	0.134	0.100	0.104	0.084	0.116	0.097	0.203	0.182	15.15
		(b)	0.166	0.151	0.148	0.127	0.095	0.066	0.159	0.141	0.117	0.087	0.185	0.169	16.56
		(c)	0.192	0.270	0.179	0.141	0.154	0.110	0.184	0.135	0.103	0.078	0.208	0.180	12.26
		(d)	0.192	0.278	0.178	0.141	0.163	0.118	0.185	0.138	0.102	0.078	0.207	0.177	11.17
II	0.2	(a)	0.127	0.127	0.110	0.109	0.124	0.121	0.096	0.094	0.119	0.118	0.206	0.203	1.28
		(b)	0.088	0.087	0.113	0.109	0.09	0.094	0.156	0.156	0.119	0.115	0.187	0.187	0.60
		(c)	0.127	0.131	0.100	0.099	0.131	0.128	0.077	0.078	0.078	0.078	0.210	0.208	-0.03
		(d)	0.127	0.132	0.096	0.094	0.140	0.136	0.075	0.075	0.077	0.077	0.209	0.207	0.33
	0.8	(a)	0.150	0.115	0.113	0.086	0.123	0.086	0.094	0.076	0.128	0.101	0.199	0.178	21.35
		(b)	0.101	0.066	0.119	0.082	0.088	0.060	0.145	0.120	0.127	0.083	0.181	0.167	26.20
		(c)	0.130	0.165	0.101	0.067	0.129	0.082	0.077	0.058	0.078	0.053	0.215	0.180	19.36
		(d)	0.129	0.169	0.097	0.065	0.137	0.090	0.074	0.057	0.076	0.053	0.213	0.178	17.66
III	0.2	(a)	0.140	0.135	0.136	0.134	0.133	0.129	0.125	0.123	0.144	0.141	0.138	0.134	2.44
		(b)	0.100	0.094	0.132	0.129	0.105	0.103	0.106	0.103	0.136	0.133	0.086	0.082	3.31
		(c)	0.115	0.115	0.118	0.119	0.127	0.127	0.119	0.118	0.124	0.124	0.095	0.095	0.00
		(d)	0.118	0.117	0.117	0.117	0.127	0.127	0.117	0.115	0.122	0.122	0.101	0.101	0.43
	0.8	(a)	0.146	0.100	0.135	0.111	0.125	0.099	0.121	0.093	0.135	0.114	0.135	0.096	22.94
		(b)	0.103	0.067	0.132	0.102	0.102	0.077	0.104	0.072	0.127	0.103	0.083	0.055	27.60
		(c)	0.121	0.086	0.118	0.091	0.129	0.089	0.119	0.081	0.123	0.101	0.097	0.063	27.95
		(d)	0.124	0.088	0.117	0.091	0.131	0.091	0.117	0.080	0.121	0.098	0.099	0.067	27.46
IV	0.2	(a)	0.230	0.217	0.407	0.404	0.265	0.254	0.195	0.189	0.497	0.494	0.263	0.251	3.13
		(b)	0.205	0.130	0.404	0.389	0.248	0.191	0.196	0.148	0.489	0.469	0.242	0.172	20.13
		(c)	0.164	0.165	0.372	0.370	0.203	0.202	0.152	0.154	0.485	0.475	0.326	0.327	0.14
		(d)	0.166	0.167	0.367	0.365	0.206	0.203	0.148	0.152	0.483	0.472	0.316	0.318	0.06
	0.8	(a)	0.218	0.193	0.403	0.397	0.254	0.232	0.207	0.185	0.491	0.481	0.256	0.229	7.47
		(b)	0.192	0.120	0.401	0.384	0.237	0.175	0.213	0.154	0.479	0.453	0.232	0.153	22.51
		(c)	0.162	0.153	0.367	0.362	0.211	0.196	0.153	0.142	0.482	0.467	0.322	0.307	4.83
		(d)	0.165	0.155	0.361	0.356	0.215	0.200	0.151	0.140	0.479	0.464	0.313	0.297	4.99

380 **References**

- [1] R. Tibshirani, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* (1996) 267–288.
- [2] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American statistical Association* 96 (456) (2001) 1348–1360.
- 385 [3] C.-H. Zhang, Nearly unbiased variable selection under minimax concave penalty, *The Annals of statistics* (2010) 894–942.
- [4] H. Zou, The adaptive lasso and its oracle properties, *Journal of the American statistical association* 101 (476) (2006) 1418–1429.
- [5] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2) (2005) 301–320.
- 390 [6] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68 (1) (2006) 49–67.
- [7] R. E. Schapire, The strength of weak learnability, *Machine learning* 5 (2) (1990) 197–227.
- [8] L. Mason, J. Baxter, P. L. Bartlett, M. R. Frean, Boosting algorithms as gradient descent, in: *Advances in neural information processing systems*, 2000, pp. 512–518.
- 395 [9] Y. Freund, R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, in: *European conference on computational learning theory*, Springer, 1995, pp. 23–37.
- [10] P. Bühlmann, B. Yu, Boosting with the  $l_2$  loss: regression and classification, *Journal of the American Statistical Association* 98 (462) (2003) 324–339.
- 400 [11] P. Bühlmann, B. Yu, Sparse boosting, *Journal of Machine Learning Research* 7 (Jun) (2006) 1001–1024.
- [12] Z. Wang, Hingeboost: Roc-based boost for classification and variable selection, *The International Journal of Biostatistics* 7 (1) (2011) 1–30.
- [13] O. Komori, S. Eguchi, A boosting method for maximizing the partial area under the roc curve, *BMC bioinformatics* 11 (1) (2010) 314.
- 405 [14] P. Bühlmann, T. Hothorn, Twin boosting: improved feature selection and prediction, *Statistics and Computing* 20 (2) (2010) 119–138.
- [15] Z. Wang, Multi-class hingeboost, *Methods of information in medicine* 51 (02) (2012) 162–167.
- [16] Y. Yang, H. Zou, Nonparametric multiple expectile regression via er-boost, *Journal of Statistical Computation and Simulation* 85 (7) (2015) 1442–1458.
- 410 [17] J. Zhao, General sparse boosting: Improving feature selection of  $l_2$  boosting by correlation-based penalty family, *Communications in Statistics-Simulation and Computation* 44 (6) (2015) 1612–1640.
- [18] R. E. Schapire, Y. Freund, *Boosting: Foundations and algorithms*, MIT press, 2012.

- 415 [19] P. Bühlmann, J. Gertheiss, S. Hieke, T. Kneib, S. Ma, M. Schumacher, G. Tutz, C.-Y. Wang, Z. Wang, A. Ziegler, et al., Discussion of the evolution of boosting algorithms and extending statistical boosting, *Methods of information in medicine* 53 (6) (2014) 436–445.
- [20] A. C. Lozano, S. R. Kulkarni, R. E. Schapire, Convergence and consistency of regularized boosting with weakly dependent observations, *IEEE Transactions on Information Theory* 60 (1) (2014) 651–660.
- [21] C.-K. Ing, T. L. Lai, A stepwise regression method and consistent model selection for high-dimensional sparse linear models, *Statistica Sinica* (2011) 1473–1513.
- 420 [22] Y. Luo, M. Spindler, High-dimensional  $L_2$  boosting: Rate of convergence, arXiv preprint arXiv:1602.08927.
- [23] S. Centorrino, J. S. Racine, Semiparametric varying coefficient models with endogenous covariates, *Annals of Economics and Statistics/Annales d'Économie et de Statistique* (128) (2017) 261–295.
- [24] C. Zhang, X. Yin, Trending time-varying coefficient market models, *Quantitative Finance* 12 (10) (2012) 1533–1546.
- 425 [25] C. O. Wu, X. Tian, H. Bang, A varying-coefficient model for the evaluation of time-varying concomitant intervention effects in longitudinal studies, *Statistics in medicine* 27 (16) (2008) 3042–3056.
- [26] J. Fan, Y. Ma, W. Dai, Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models, *Journal of the American Statistical Association* 109 (507) (2014) 1270–1284.
- [27] M. Yue, J. Li, Improvement screening for ultra-high dimensional data with censored survival outcomes and varying coefficients, *International Journal of Biostatistics*.
- 430 [28] J. Gao, B. Peng, Z. Ren, X. Zhang, et al., Variable selection for a categorical varying-coefficient model with identifications for determinants of body mass index, *The Annals of Applied Statistics* 11 (2) (2017) 1117–1145.
- [29] C. Giraud, *Introduction to High-dimensional Statistics*, CRC Press, 2015.
- [30] P. Diggle, *Analysis of longitudinal data*, Oxford University Press, 2002.
- [31] G. M. Fitzmaurice, N. M. Laird, J. H. Ware, *Applied longitudinal analysis*, Vol. 998, John Wiley & Sons, 2012.
- 435 [32] X. Lin, R. J. Carroll, Semiparametric regression for clustered data using generalized estimating equations, *Journal of the American Statistical Association* 96 (455) (2001) 1045–1056.
- [33] J. Fan, T. Huang, R. Li, Analysis of longitudinal data with semiparametric estimation of covariance function, *Journal of the American Statistical Association* 102 (478) (2007) 632–641.
- 440 [34] M.-Y. Cheng, T. Honda, J. Li, H. Peng, et al., Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data, *The Annals of Statistics* 42 (5) (2014) 1819–1849.
- [35] H. Wang, Y. Xia, Shrinkage estimation of the varying coefficient model, *Journal of the American Statistical Association* 104 (486) (2009) 747–757.
- [36] L. Wang, H. Li, J. Z. Huang, Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements, *Journal of the American Statistical Association* 103 (484) (2008) 1556–1569.
- 445 [37] M.-Y. Cheng, T. Honda, J. Li, et al., Efficient estimation in semivarying coefficient models for longitudinal/clustered data, *The Annals of Statistics* 44 (5) (2016) 1988–2017.

- [38] P. Bühlmann, Boosting for high-dimensional linear models, *The Annals of Statistics* (2006) 559–583.
- [39] P. Green, B. Silverman, *Nonparametric regression and generalized linear models*. number 58 in *monographs on statistics and applied probability* (1994).
- 450 [40] P. Bühlmann, T. Hothorn, *Boosting algorithms: Regularization, prediction and model fitting*, *Statistical Science* (2007) 477–505.
- [41] M. H. Hansen, B. Yu, Model selection and the principle of minimum description length, *Journal of the American Statistical Association* 96 (454) (2001) 746–774.
- [42] J. H. Friedman, Greedy function approximation: a gradient boosting machine, *Annals of statistics* (2001) 1189–1232.
- 455 [43] Y. Luan, H. Li, Clustering of time-course gene expression data using a mixed-effects model with b-splines, *Bioinformatics* 19 (4) (2003) 474–482.
- [44] L. Wang, G. Chen, H. Li, Group scad regression analysis for microarray time course gene expression data, *Bioinformatics* 23 (12) (2007) 1486–1494.
- [45] F. Wei, J. Huang, H. Li, Variable selection and estimation in high-dimensional varying-coefficient models, *Statistica Sinica* 21 (4) (2011) 1515.
- 460 [46] X. Xia, B. Jiang, J. Li, W. Zhang, Low-dimensional confounder adjustment and high-dimensional penalized estimation for survival analysis, *Lifetime data analysis* 22 (4) (2016) 547–569.
- [47] H.-K. Tsai, H. H.-S. Lu, W.-H. Li, Statistical methods for identifying yeast cell cycle transcription factors, *Proceedings of the National Academy of Sciences of the United States of America* 102 (38) (2005) 13532–13537.
- 465 [48] M. Dehmer, F. Emmert-Streib, *Analysis of microarray data: a network-based approach*, John Wiley & Sons, 2008.
- [49] C. J. Loy, D. Lydall, U. Surana, Ndd1, a high-dosage suppressor of *cdc28-1n*, is essential for expression of a subset of late-s-phase-specific genes in *saccharomyces cerevisiae*, *Molecular and cellular biology* 19 (5) (1999) 3312–3327.
- [50] C. Koch, T. Moll, M. Neuberg, H. Ahorn, K. Nasmyth, A role for the transcription factors *mbp1* and *swi4* in progression from *g1* to *s* phase, *Science* 261 (5128) (1993) 1551–1557.
- 470 [51] R. Kumar, D. M. Reynolds, A. Shevchenko, A. Shevchenko, S. D. Goldstone, S. Dalton, Forkhead transcription factors, *fkh1p* and *fkh2p*, collaborate with *mcm1p* to control transcription required for *m*-phase, *Current Biology* 10 (15) (2000) 896–906.
- [52] N. Banerjee, M. Q. Zhang, Identifying cooperativity among transcription factors controlling the cell cycle in yeast, *Nucleic acids research* 31 (23) (2003) 7024–7031.
- 475 [53] M. S. Spector, A. Raff, H. DeSilva, K. Lee, M. A. Osley, *Hir1p* and *hir2p* function as transcriptional corepressors to regulate histone gene transcription in the *saccharomyces cerevisiae* cell cycle., *Molecular and cellular biology* 17 (2) (1997) 545–552.
- [54] M. Koranda, A. Schleiffer, L. Endler, G. Ammerer, Forkhead-like transcription factors recruit *ndd1* to the chromatin of *g2/m*-specific promoters, *Nature* 406 (6791) (2000) 94.
- 480 [55] M. Yue, J. Li, S. Ma, Sparse boosting for high-dimensional survival data with varying coefficients, *Statistics in medicine* 37 (5) (2018) 789–800.

- [56] G. Lugosi, N. Vayatis, On the bayes-risk consistency of regularized boosting methods, *Annals of Statistics* (2004) 30–55.
- [57] M. Schmidt, Graphical model structure learning with  $l_1$ -regularization, University of British Columbia.
- 485 [58] J. Fan, Y. Feng, Y. Wu, Network exploration via the adaptive lasso and scad penalties, *The annals of applied statistics* 3 (2) (2009) 521.
- [59] A. M. Fitch, Computationally tractable fitting of graphical models: the cost and benefits of decomposable bayesian and penalized likelihood approaches: a thesis presented in partial fulfillment of the requirements for the degree of doctor of philosophy in statistics at massey university, albany, new zealand, Ph.D. thesis, Massey University (2012).
- 490 [60] N. Wang, Marginal nonparametric kernel regression accounting for within-subject correlation, *Biometrika* 90 (1) (2003) 43–52.
- [61] C. M. Fortuin, P. W. Kasteleyn, On the random-cluster model: I. introduction and relation to other models, *Physica* 57 (4) (1972) 536–564.
- [62] A. Gelman, J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, 495 2012.
- [63] D. Rizopoulos, *Joint Models for Longitudinal and Time-to-event Data*, CRC Press, 2012.
- [64] D. TvER, The asymptotic distribution of singular values with applications to canonical correlations and correspondence analysis, *Journal of Multivariate Analysis* 50 (1994) 238–264.
- [65] J. Fan, Y. Fan, J. Lv, High dimensional covariance matrix estimation using a factor model, *Journal of Econometrics* 500 147 (1) (2008) 186–197.
- [66] Y. Cui, C. Leng, D. Sun, Sparse estimation of high-dimensional correlation matrices, *Computational Statistics & Data Analysis* 93 (2016) 390–403.
- [67] S. X. Chen, Y.-L. Qin, et al., A two-sample test for high-dimensional data with applications to gene-set testing, *The Annals of Statistics* 38 (2) (2010) 808–835.
- 505 [68] M. Biswas, A. K. Ghosh, A nonparametric two-sample test applicable to high dimensional data, *Journal of Multivariate Analysis* 123 (2014) 160–171.
- [69] P. K. Mondal, M. Biswas, A. K. Ghosh, On high dimensional two-sample tests based on nearest neighbors, *Journal of Multivariate Analysis* 141 (2015) 168–178.
- [70] P. J. Bickel, E. Levina, et al., Covariance regularization by thresholding, *The Annals of Statistics* 36 (6) (2008) 2577–510 2604.
- [71] P. J. Bickel, E. Levina, et al., Regularized estimation of large covariance matrices, *The Annals of Statistics* 36 (1) (2008) 199–227.
- [72] T. T. Cai, C.-H. Zhang, H. H. Zhou, et al., Optimal rates of convergence for covariance matrix estimation, *The Annals of Statistics* 38 (4) (2010) 2118–2144.
- 515 [73] T. T. Cai, M. Yuan, et al., Adaptive covariance matrix estimation through block thresholding, *The Annals of Statistics* 40 (4) (2012) 2014–2042.

- [74] J. Friedman, T. Hastie, R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso, *Biostatistics* 9 (3) (2008) 432–441.
- [75] M. Yuan, High dimensional inverse covariance matrix estimation via linear programming, *Journal of Machine Learning Research* 11 (Aug) (2010) 2261–2286.
- 520
- [76] T. Cai, W. Liu, X. Luo, A constrained  $\ell_1$  minimization approach to sparse precision matrix estimation, *Journal of the American Statistical Association* 106 (494) (2011) 594–607.
- [77] J. Friedman, T. Hastie, R. Tibshirani, et al., Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors), *The annals of statistics* 28 (2) (2000) 337–407.