



RESEARCH ARTICLE

Penalized integrative semiparametric interaction analysis for multiple genetic datasets

Yang Li^{1,2,3}  | Rong Li^{2,3} | Cunjie Lin^{1,2,3} | Yichen Qin⁴ | Shuangge Ma^{2,5} 

¹Center for Applied Statistics, Renmin University of China, Beijing, China

²School of Statistics, Renmin University of China, Beijing, China

³Statistical Consulting Center, Renmin University of China, Beijing, China

⁴Department of Operations, Business Analytics and Information Systems, University of Cincinnati, Cincinnati, Ohio

⁵Department of Biostatistics, Yale University, New Haven, Connecticut

Correspondence

Cunjie Lin, School of Statistics, Renmin University of China, Beijing 100872, China.

Email: lincunjie@ruc.edu.cn

Shuangge Ma, Department of Biostatistics, Yale University, New Haven, CT 06520.

Email: shuangge.ma@yale.edu

Present Address

59 Zhongguancun St., Beijing, China

Funding information

National Institutes of Health, Grant/Award Number: CA204120, CA121974, and CA196530; National Natural Science Foundation of China, Grant/Award Number: 11701561, 71771211, and 81774206; the MOE Project of Key Research Institute of Humanities and Social Sciences at Universities, Grant/Award Number: 16JJD910002

In this article, we consider a semiparametric additive partially linear interaction model for the integrative analysis of multiple genetic datasets. The goals are to identify important genetic predictors and gene-gene interactions and to estimate the nonparametric functions that describe the environmental effects at the same time. To find the similarities and differences of the genetic effects across different datasets, we impose a group structure on the regression coefficients matrix under the homogeneity assumption, ie, models for different datasets share the same sparsity structure, but the coefficients may differ across datasets. We develop an iterative approach to estimate the parameters of main effects, interactions and nonparametric functions, where a reparametrization of interaction parameters is implemented to meet the strong hierarchy assumption. We demonstrate the advantages of the proposed method in identification, estimation, and prediction in a series of numerical studies. We also apply the proposed method to the Skin Cutaneous Melanoma data and the lung cancer data from the Cancer Genome Atlas.

KEYWORDS

Gene-gene interaction analysis, hierarchical constraint, integrative analysis, semiparametric model

1 | INTRODUCTION

High-dimensional genetic data has been extensively analyzed to discover complicated biological mechanisms.¹ For a single genetic dataset with a small sample size but a large number of genes, various variable selection methods, such as Lasso, adaptive Lasso, smoothly clipped absolute deviation, minimax concave penalty, and others,²⁻⁵ have been applied to study the associations between genetic measurements and clinical outcomes. Considering that multiple genes have coordinated biological functions and/or correlated expressions, selection in groups⁶ has also been applied. For many diseases like cancer, it is more difficult to identify important genetic factors since one gene can have different effects in different clinical stages. An example is the skin cutaneous melanoma (SKCM) data from the Cancer Genome Atlas

(TCGA). One goal of the study was to investigate genetic and environmental effects on the development of melanoma. For patients in different clinical stages, genes can have different effects on the Breslow thickness, which is a continuous variable and has been extensively used as a prognostic indicator for melanoma.^{7,8} Therefore, subjects with the three stages should not be analyzed as a single population, which motivates us to develop an appropriate analysis method to handle such differences across multiple datasets.

One possible solution is to conduct integrative analysis, which efficiently assembles and jointly analyzes raw data from multiple datasets, and enables us to improve the estimation by borrowing information across datasets while allowing differences of signals in different datasets.⁹⁻¹¹ As mentioned in the work of Liu et al,¹¹ datasets can be described under the homogeneity structure (or heterogeneity structure), ie, datasets have the same (or different) sets of disease-associated genetic measurements. It is desirable to advocate the homogeneity assumption for multiple datasets from different clinical stages or subtypes of the same disease. For example, with the SKCM data analyzed in this study, datasets are collected on the same cancer with different clinical stages and it is quite likely that a gene has either null effect or important but different influences in different stages, which means that the models for different stages share the same sparsity structure, but their coefficients have different magnitudes. In this study, we conduct integrative analysis under the homogeneity assumption and identify significant genes for all datasets simultaneously but leave the degrees of influence to be different in different datasets.¹²⁻¹⁴

For many complex diseases, the main genetic effects alone may not be sufficient to capture the relationship between the response and predictors. The gene-gene interactions can also contribute to improving prediction accuracy and assessing the functions of genes. For example, Ochoa et al¹⁵ confirmed that the interaction between *PPAR γ 2* and *ADRB3* increases obesity risk significantly in children and adolescents. However, fitting regression models for genetic data with gene-gene interactions brings more challenges due to its extremely high-dimensional issue and hierarchical structure. For a dataset with n samples and p predictors, the number of unknown parameters increases to $(p^2 + p)/2$ in interaction analysis, which would be much larger than n even for a moderate p . Moreover, a more complex algorithm is needed to address the “main effects, interactions” hierarchical structure,¹⁶ because directly applying the existing variable selection procedures may violate the hierarchy assumption and cause trouble in interpretation and inference. For a single dataset, some studies have offered solutions to ensure the hierarchical structure in variable selection. For example, Zhao et al¹⁷ proposed the composite absolute penalties family for grouped and hierarchical variable selection by defining groups and combining L_γ -norm penalties; Choi et al¹⁸ extended the Lasso method for simultaneously fitting a regression model and identifying interactions under the strong hierarchy constraint, ie, an interaction term can be selected only if the corresponding main effects are also selected. In integrative analysis, a feasible approach should be studied to accommodate multiple datasets and meet the hierarchy assumption simultaneously.

In addition to genetic measurements, environmental measurements can also influence the development of diseases. Different from genetic measurements, though, we consider a nonparametric approach to model the effects of environmental factors. This is because a preliminary analysis of the SKCM data shows that the linear relationship between age and Breslow thickness is denied. As illustrated in Figure 1, the curve in Stage I is close to zero while there are obviously nonlinear trends in other stages, which means that the effects of age on the response variable are nonlinear and different for different stages. As a result, a semiparametric regression model is developed in this study, where the genetic and environmental measurements are included in the model as the parametric and nonparametric components, respectively. In order to estimate the nonparametric component functions while carrying out variable selection on the parametric part, we modify the algorithm of Du et al,¹⁹ which achieves simultaneous model selection for both nonparametric and parametric parts. They introduced a penalty combining the adaptive empirical L_2 -norms for the nonparametric component estimation and the smoothly clipped absolute deviation penalty for the parametric component selection. In this study, we resort to this iterative algorithm to identify genetic factors with interactions, but leave the nonparametric functions approximated using B-spline techniques without selection.²⁰⁻²³

Consequently, the goals of this study are to identify important genetic factors and gene-gene interactions, and to improve the estimates of parameters by jointly analyzing multiple datasets, while allowing differences across datasets. This can be achieved by conducting a penalized integrative interaction analysis under semiparametric regression modeling. Different from the integrative analysis of gene-environment interactions in the work of Wu et al,²⁴ we focus on the identification of gene-gene interactions and the nonlinear relationship between the environmental factors and outcome. Here, we build on the work of Liu et al,¹¹ which is on penalized integrative analysis under the homogeneity structure for multiple datasets, and extend the work of Choi et al,¹⁸ which is on penalized interaction analysis under the strong hierarchical constraint for a single dataset, meanwhile, we introduce nonparametric functions to describe the effects of environmental factors. A modified iterative algorithm of Du et al¹⁹ is proposed to account for the penalized integrative interaction analyses and estimate nonparametric functions with multiple datasets. Finally, the SKCM data and lung cancer data from TCGA are analyzed.

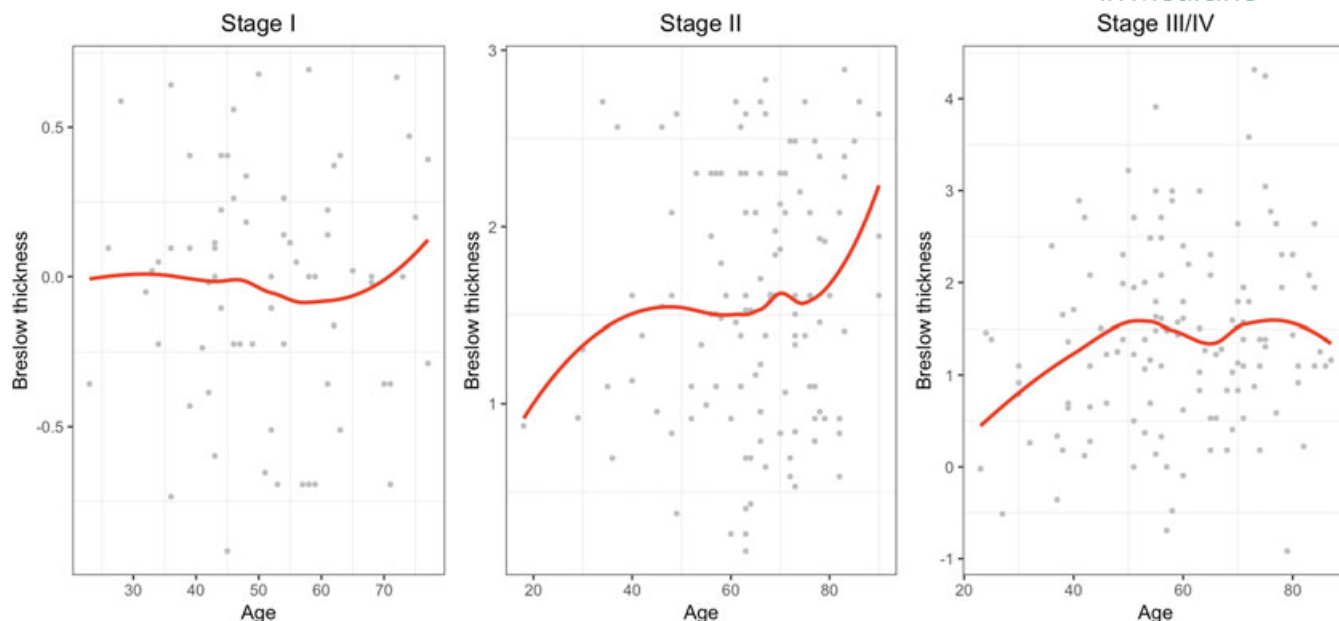


FIGURE 1 Breslow thickness against age; black points are the observations and the red line is the locally weighted scatterplot smoothing [Colour figure can be viewed at wileyonlinelibrary.com]

The rest of this paper is organized as follows. In Section 2, the semiparametric model and corresponding estimation method are introduced in detail and the algorithms from a single dataset to multiple datasets, are also presented. In Section 3, extensive numerical studies are carried out to evaluate the performance of the proposed method. Real datasets from TCGA are analyzed in Section 4 to illustrate the application of the proposed method. Concluding remarks are given in Section 5.

2 | METHODOLOGY

2.1 | Semiparametric model under strong hierarchical constraint

First, consider the case of a single dataset. Let $\mathbf{y} = (y_1, \dots, y_n)'$ be the response vector, $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_p) \in \mathbb{R}^{n \times p}$ be the covariate matrix for the genetic measurements and $\mathbf{u} = (\mathbf{u}_1, \dots, \mathbf{u}_q) \in \mathbb{R}^{n \times q}$ be the covariate matrix for the environmental measurements. Consider the gene-gene interactions $\mathbf{x}_j \mathbf{x}_{j'} = (x_{1j}x_{1j'}, \dots, x_{nj}x_{nj'})'$, $j < j'$, $j, j' = 1, \dots, p$. The semiparametric additive partially linear interaction model is given by

$$\mathbf{y} = \sum_{j=1}^p \beta_j \mathbf{x}_j + \sum_{j=1}^{p-1} \sum_{j'=j+1}^p \alpha_{jj'} \mathbf{x}_j \mathbf{x}_{j'} + \sum_{k=1}^q f_k(\mathbf{u}_k) + \boldsymbol{\epsilon}, \quad (1)$$

where β_j is the coefficient of the main effect of \mathbf{x}_j , and $\alpha_{jj'}$ is the coefficient of the interaction term $\mathbf{x}_j \mathbf{x}_{j'}$. $f_k(\cdot)$ is an unknown smooth function. $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)'$ is the vector of mean zero error terms. All the variables are centered before modeling so there is no intercept in (1).

In analysis, we reinforce the strong hierarchical constraint, ie,

$$\alpha_{jj'} \neq 0 \rightarrow \beta_j \neq 0 \quad \text{and} \quad \beta_{j'} \neq 0, \forall j, j' = 1, \dots, p. \quad (2)$$

Without loss of generality, assume that the environmental predictors \mathbf{u}_k are continuous and the smooth function $f_k(\cdot)$ satisfies $\int_0^1 f_k(u) du = 0$ ($k = 1, \dots, q$) to ensure identifiability. To estimate $f_k(\cdot)$ ($k = 1, \dots, q$), we use the cubic B-spline approximation. Following smoothness assumptions as in the works of Liu et al²⁰ and Wang et al,²¹ $f_k(\cdot)$ ($k = 1, \dots, q$) can be approximated by spline functions, ie,

$$f_k(u) \approx \sum_{d=1}^{D_k} \phi_k^d B_k^d(u) = \mathbf{B}_k'(u) \boldsymbol{\phi}_k, \quad (3)$$

where $\mathbf{B}_k(u) = (B_k^1(u), \dots, B_k^{D_k}(u))'$ and $\{B_k^i : i = 1, \dots, D_k\}$ are the B-spline basis functions and $\boldsymbol{\phi}_k = (\phi_k^1, \dots, \phi_k^{D_k})'$ is the unknown parameter vector. Thus, the estimation of nonparametric functions is transformed to estimating $\boldsymbol{\phi}_k$ ($k = 1, \dots, q$).

Denote $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$, $\boldsymbol{\alpha} = (\alpha_{12}, \dots, \alpha_{1p}, \dots, \alpha_{p-1,p})'$, $\boldsymbol{\phi} = (\boldsymbol{\phi}_1', \dots, \boldsymbol{\phi}_q')'$; we propose the penalized objective function

$$Q(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\phi}) = \frac{1}{2n} \left\| \mathbf{y} - \sum_j \beta_j \mathbf{x}_j - \sum_{j < j'} \alpha_{jj'} \mathbf{x}_j \mathbf{x}_{j'} - \sum_k \mathbf{Z}_k \boldsymbol{\phi}_k \right\|_2^2 + \rho(\boldsymbol{\beta}, \boldsymbol{\alpha}), \quad (4)$$

where $\mathbf{Z}_k = (\mathbf{B}_k(u_{1k}), \dots, \mathbf{B}_k(u_{nk}))'$ is the $n \times D_k$ dimensional B-spline matrix and $\rho(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is the penalty function. Here, we use the Lasso penalty with $\rho(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \lambda_\beta \sum_j |\beta_j| + \lambda_\alpha \sum_{j < j'} |\alpha_{jj'}|$, where λ_β and λ_α are the tuning parameters. However, the direct use of Lasso does not guarantee the hierarchy.

To account for the strong hierarchical constraint, we rewrite the coefficients of the interaction terms $\boldsymbol{\alpha}$ as the product of the corresponding main effects $\boldsymbol{\beta}$ and additional parameters $\boldsymbol{\gamma}$, ie, $\alpha_{jj'} = \gamma_{jj'} \beta_j \beta_{j'}$; the model itself guarantees the hierarchical constraint via reparametrization. That is, whenever any one of $\gamma_{jj'}$, β_j , $\beta_{j'}$ is zero, $\alpha_{jj'}$ is automatically zero. Thus, the minimization problem in (4) is equivalent to finding $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ to minimize

$$Q(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\phi}) = \frac{1}{2n} \left\| \mathbf{y} - \sum_j \beta_j \mathbf{x}_j - \sum_{j < j'} \beta_j \beta_{j'} \gamma_{jj'} \mathbf{x}_j \mathbf{x}_{j'} - \sum_k \mathbf{Z}_k \boldsymbol{\phi}_k \right\|_2^2 + \lambda_\beta \sum_j |\beta_j| + \lambda_\gamma \sum_{j < j'} |\gamma_{jj'}|, \quad (5)$$

where $\boldsymbol{\gamma} = (\gamma_{12}, \dots, \gamma_{1p}, \dots, \gamma_{p-1,p})'$.

2.2 | Extension to multiple datasets

Consider the integrative analysis of L datasets. We further add superscript “(l)” to indicate the l th dataset. Suppose that there are n_l subjects in the l th ($= 1, \dots, L$) dataset, and $\mathbf{y}^{(l)} = (y_1^{(l)}, \dots, y_{n_l}^{(l)})'$ is the observed response vector. Denote $\mathbf{x}^{(l)} = (\mathbf{x}_1^{(l)}, \dots, \mathbf{x}_p^{(l)}) \in \mathbb{R}^{n_l \times p}$ and $\mathbf{u}^{(l)} = (u_1^{(l)}, \dots, u_q^{(l)}) \in \mathbb{R}^{n_l \times q}$ as the covariate matrices of the genetic and environmental measurements, respectively. The l th semiparametric additive partially linear interaction model is

$$\mathbf{y}^{(l)} = \sum_{j=1}^p \beta_j^{(l)} \mathbf{x}_j^{(l)} + \sum_{j=1}^{p-1} \sum_{j'=j+1}^p \alpha_{jj'}^{(l)} \mathbf{x}_j^{(l)} \mathbf{x}_{j'}^{(l)} + \sum_{k=1}^q f_k^{(l)}(\mathbf{u}_k^{(l)}) + \boldsymbol{\epsilon}^{(l)},$$

where $\beta_j^{(l)}$ and $\alpha_{jj'}^{(l)}$ denote the coefficients of the main effects and interactions in the l th dataset, respectively. $f_k^{(l)}(\cdot)$ is the unknown function to capture the environmental effect of the l th dataset. $\boldsymbol{\epsilon}^{(l)} = (\epsilon_1^{(l)}, \dots, \epsilon_{n_l}^{(l)})'$ is the error term with mean zero and variance $\sigma^2 I_{n_l}$.

We reparametrize the coefficients for the interaction terms $\alpha_{jj'}^{(l)}$ as $\alpha_{jj'}^{(l)} = \gamma_{jj'}^{(l)} \beta_j^{(l)} \beta_{j'}^{(l)}$, $j < j'$, $j, j' = 1, \dots, p$ to guarantee the hierarchical constraint. The difference is that we conduct integrative analysis under the homogeneity assumption, which implies that $I(\beta_j^{(l)} = 0) = I(\beta_j^{(k)} = 0)$ for all (l, k, j) 's, and $I(\gamma_{jj'}^{(l)} = 0) = I(\gamma_{jj'}^{(k)} = 0)$ for all (l, k, j, j') 's. Here, we use $\boldsymbol{\beta}_j = (\beta_j^{(1)}, \dots, \beta_j^{(L)})'$ to denote the coefficients of the j th main effect and $\boldsymbol{\gamma}_{jj'} = (\gamma_{jj'}^{(1)}, \dots, \gamma_{jj'}^{(L)})'$ to denote the additional parameters of the corresponding interactions in all L datasets. Besides, we approximate the nonparametric functions $f_k^{(l)}(\cdot)$, $k = 1, \dots, q$, $l = 1, \dots, L$ with B-splines and use $\boldsymbol{\phi}_k^{(l)}$ to denote the parameter vector of the k th nonparametric covariate in the l th dataset, and $\boldsymbol{\phi}_k = (\boldsymbol{\phi}_k^{(1)}, \dots, \boldsymbol{\phi}_k^{(L)})'$ denotes the corresponding spline parameter in all L datasets. We construct the loss function by integrating all individual loss functions, and the penalized objective function is given by

$$Q(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\phi}) = \sum_{l=1}^L \frac{1}{2n_l} \left\| \mathbf{y}^{(l)} - \sum_j \beta_j^{(l)} \mathbf{x}_j^{(l)} - \sum_{j < j'} \beta_j^{(l)} \beta_{j'}^{(l)} \gamma_{jj'}^{(l)} \mathbf{x}_j^{(l)} \mathbf{x}_{j'}^{(l)} - \sum_k \mathbf{Z}_k^{(l)} \boldsymbol{\phi}_k^{(l)} \right\|_2^2 + \rho(\boldsymbol{\beta}, \boldsymbol{\gamma}),$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p)$ is the $L \times p$ matrix of main effect coefficients, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_{12}, \dots, \boldsymbol{\gamma}_{1p}, \dots, \boldsymbol{\gamma}_{p-1,p})$ is the $L \times \frac{p(p-1)}{2}$ matrix of the additional parameters of interactions, and $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_q)$ as the $L \times d$ ($d = \sum_{k=1}^q D_k$) matrix of the spline parameters.

Under the homogeneity assumption, the L datasets share the same set of disease-associated covariates, which means that the coefficients $\boldsymbol{\beta}_j$ for the j th gene in the L datasets form a natural group and should be retained or discarded as a whole. This can be achieved by imposing a group structure on the regression coefficients and applying the group Lasso penalty. The group Lasso penalty function is defined by $p_\lambda(\boldsymbol{\theta}) = \lambda \sum_{j=1}^J \|\boldsymbol{\theta}_j\|_2$, where J is the number of groups, $\|\boldsymbol{\theta}_j\|_2$ is the L_2 -norm of the corresponding coefficient vector in the j th group. This penalty can be viewed as an intermediate

between Lasso and ridge and encourages variable selection at the group level. Following the notations defined above, β_j is a group containing the coefficients of the j th gene in all L datasets. Thus, the penalty function for the main effects β can be defined by $\rho(\beta) = \lambda_\beta \sum_j \|\beta_j\|_2$. As for interactions, the group selection penalty function is defined as $\rho(\gamma) = \lambda_\gamma \sum_{j < j'} \|\gamma_{jj'}\|_2$. Overall, the proposed penalized objective function is

$$Q(\beta, \gamma, \phi) = \sum_{l=1}^L \frac{1}{2n_l} \left\| \mathbf{y}^{(l)} - \sum_j \beta_j^{(l)} \mathbf{x}_j^{(l)} - \sum_{j < j'} \beta_j^{(l)} \beta_{j'}^{(l)} \gamma_{jj'}^{(l)} \mathbf{x}_j^{(l)} \mathbf{x}_{j'}^{(l)} - \sum_k \mathbf{z}_k^{(l)} \phi_k^{(l)} \right\|_2^2 + \sum_j \lambda_\beta \|\beta_j\|_2 + \sum_{j < j'} \lambda_\gamma \|\gamma_{jj'}\|_2.$$

Note that the coefficient for the gene-gene interactions $\mathbf{x}_j^{(l)} \mathbf{x}_{j'}^{(l)}$ is expressed as the product of $\gamma_{jj'}^{(l)}$, $\beta_j^{(l)}$ and $\beta_{j'}^{(l)}$, ie, $\alpha_{jj'}^{(l)} = \gamma_{jj'}^{(l)} \beta_j^{(l)} \beta_{j'}^{(l)}$, $j < j', j, j' = 1, \dots, p$. Consequently, when either $\beta_j^{(l)}$ or $\beta_{j'}^{(l)}$ is equal to zero, $\alpha_{jj'}^{(l)}$ is automatically set to zero.

2.3 | Algorithm

In this study, we keep the dimension of nonparametric components low, and use an iterative approach to estimate the nonparametric spline parameters and main effects/interactions coefficients.¹⁹ Specifically, we estimate ϕ using ordinary least square when β and γ are fixed, and given the spline parameters, we use the idea of Choi et al¹⁸ for the main effects and the interactions iteratively.

Notice that, under the strong hierarchical constraint, the selection of main effects can affect the selection of corresponding interactions. Therefore, we estimate β_j sequentially; for $j = 1, \dots, p$, we fix $\hat{\phi}$, $\hat{\gamma}$, and $\hat{\beta}_1, \dots, \hat{\beta}_{j-1}, \hat{\beta}_{j+1}, \dots, \hat{\beta}_p$, and then the estimate of β_j becomes a group Lasso problem with only one group. In this paper, we adopt the group descent approach²⁵ to solve the group Lasso problems. Overall, we estimate the three parts of parameters, ie, coefficients of main effects (β), additional parameters of interaction effects (γ), and parameters of spline regression (ϕ), by deriving one part while fixing the other two. Denote $\hat{\beta}_j^{(l)}(t)$ as the estimate of $\beta_j^{(l)}$ at the t th iteration. The algorithm proceeds as follows.

(1) Initialization. We use ridge regression estimates $\tilde{\gamma}$, $\tilde{\beta}$, and $\tilde{\phi}$ as initial values.

(2) Update $\hat{\phi}$. Let

$$\tilde{\mathbf{y}}^{(l)} = \mathbf{y}^{(l)} - \sum_j \hat{\beta}_j^{(l)}(t-1) \mathbf{x}_j^{(l)} - \sum_{j < j'} \hat{\beta}_j^{(l)}(t-1) \hat{\beta}_{j'}^{(l)}(t-1) \hat{\gamma}_{jj'}^{(l)}(t-1) \mathbf{x}_j^{(l)} \mathbf{x}_{j'}^{(l)}.$$

Get $\hat{\phi}(t)$ by minimizing $Q_{\beta, \alpha}(\phi) = \sum_{l=1}^L \frac{1}{2n_l} \|\tilde{\mathbf{y}}^{(l)} - \sum_k \mathbf{z}_k^{(l)} \phi_k^{(l)}\|_2^2$.

(3) Update $\hat{\gamma}$. Let

$$\begin{aligned} \tilde{\mathbf{y}}^{(l)} &= \mathbf{y}^{(l)} - \sum_j \hat{\beta}_j^{(l)}(t-1) \mathbf{x}_j^{(l)} - \sum_k \mathbf{z}_k^{(l)} \hat{\phi}_k^{(l)}(t). \\ \tilde{\mathbf{x}}_{jj'}^{(l)} &= \hat{\beta}_j^{(l)}(t-1) \hat{\beta}_{j'}^{(l)}(t-1) \mathbf{x}_j^{(l)} \mathbf{x}_{j'}^{(l)}. \end{aligned}$$

Compute $\hat{\gamma}(t)$ by minimizing $Q_{\beta, \phi}(\gamma) = \sum_{l=1}^L \frac{1}{2n_l} \|\tilde{\mathbf{y}}^{(l)} - \sum_{j < j'} \gamma_{jj'}^{(l)} \tilde{\mathbf{x}}_{jj'}^{(l)}\|_2^2 + \sum_{j < j'} \lambda_\gamma \|\gamma_{jj'}\|_2$.

This can be achieved by the following iterations.

- (i) $\mathbf{r}^{(l)} = \tilde{\mathbf{y}}^{(l)} - \sum_{j < j'} \hat{\gamma}_{jj'}^{(l)} \tilde{\mathbf{x}}_{jj'}^{(l)}$. Let $\mathbf{r} = (\mathbf{r}^{(1)T}, \dots, \mathbf{r}^{(L)T})^T$.
- (ii) For $j < j', j, j' = 1, \dots, p$:
Let $\tilde{\mathbf{x}}_{jj'} = \text{diag}(\tilde{\mathbf{x}}_{jj'}^{(1)}, \dots, \tilde{\mathbf{x}}_{jj'}^{(L)})$, and $\hat{\gamma}_{jj'} = (\hat{\gamma}_{jj'}^{(1)}, \dots, \hat{\gamma}_{jj'}^{(L)})^T$. Compute $\mathbf{Z}_{jj'} \leftarrow \tilde{\mathbf{x}}_{jj'}^T \mathbf{r} + \hat{\gamma}_{jj'}$.
Define $S(\mathbf{Z}, \lambda) = S(\|\mathbf{Z}\|, \lambda) \frac{\mathbf{Z}}{\|\mathbf{Z}\|}$, where

$$S(\mathbf{Z}, \lambda) = \begin{cases} \mathbf{Z} - \lambda & \|\mathbf{Z}\| > \lambda \\ 0 & \|\mathbf{Z}\| \leq \lambda \\ \mathbf{Z} + \lambda & \|\mathbf{Z}\| < -\lambda \end{cases}$$

is the soft-thresholding operator.

Thus, $\hat{\gamma}'_{jj'} \leftarrow S(\mathbf{Z}_{jj'}, \lambda)$

Update $\mathbf{r} \leftarrow \mathbf{r} - \tilde{\mathbf{x}}_{jj'} (\hat{\gamma}'_{jj'} - \hat{\gamma}_{jj'})$.

(iii) Repeat Step (ii) until convergence.

(4) Update $\hat{\beta}$. For each j in $1, \dots, p$, let

$$\begin{aligned}\tilde{\mathbf{y}}^{(l)} &= \mathbf{y}^{(l)} - \sum_{j' \neq j} \hat{\beta}_{j'}^{(l)}(t-1)\mathbf{x}_{j'}^{(l)} - \sum_{j' < j'', j', j'' \neq j} \hat{\beta}_{j'}^{(l)}(t-1)\hat{\beta}_{j''}^{(l)}(t-1)\hat{\gamma}_{j'j''}^{(l)}(t)\mathbf{x}_{j'}^{(l)}\mathbf{x}_{j''}^{(l)} - \sum_k \mathbf{z}_k^{(l)}\hat{\phi}_k^{(l)}(t). \\ \tilde{\mathbf{x}}^{(l)} &= \mathbf{x}_j^{(l)} + \sum_{j' < j} \hat{\beta}_{j'}^{(l)}(t-1)\hat{\gamma}_{j'j}^{(l)}(t)\mathbf{x}_{j'}^{(l)}\mathbf{x}_j^{(l)} + \sum_{j' > j} \hat{\beta}_{j'}^{(l)}(t-1)\hat{\gamma}_{jj'}^{(l)}(t)\mathbf{x}_j^{(l)}\mathbf{x}_{j'}^{(l)}.\end{aligned}$$

Obtain $\hat{\beta}_j(t)$ by minimizing $Q_{\alpha, \phi}(\beta_j) = \sum_{l=1}^L \frac{1}{2n_l} \|\tilde{\mathbf{y}}^{(l)} - \tilde{\mathbf{x}}^{(l)}\beta_j\|_2^2 + \lambda_\beta \|\beta_j\|_2$. This can be achieved using the same group descent approach in Step (3).

(5) Iteration. Repeat (2) to (4) until convergence. In numerical study, $\max\{\|\hat{\beta}(t) - \hat{\beta}(t-1)\|_2, \|\hat{\gamma}(t) - \hat{\gamma}(t-1)\|_2\} \leq 10^{-3}$ is used as the convergence criterion.

To select the regularization parameters λ_β and λ_γ , we adopt the Bayesian information criterion-based approach. Denote the total sample size $n = n_1 + \dots + n_L$, total residual sum of squares as RSS, and total degree of freedom as df. The Bayesian information criterion is $\log(\text{RSS}) + \text{df} \times \log(n)/n$.

2.4 | Extension to accelerated failure time model

The proposed method can be naturally extended to the accelerated failure time (AFT) model with survival data. Let $\mathbf{T}^{(l)}$ be the logarithm of survival time in the l th dataset, and consider the AFT model

$$\mathbf{T}^{(l)} = \sum_{j=1}^p \beta_j^{(l)} \mathbf{x}_j^{(l)} + \sum_{j=1}^{p-1} \sum_{j'=j+1}^p \alpha_{jj'}^{(l)} \mathbf{x}_j^{(l)} \mathbf{x}_{j'}^{(l)} + \sum_{k=1}^q f_k(\mathbf{u}_k^{(l)}) + \epsilon^{(l)}, \quad (6)$$

where $\epsilon^{(l)}$ is the random error with an unknown distribution; $\mathbf{x}_j^{(l)}$ and $\mathbf{u}_k^{(l)}$ are genetic and environmental predictors, respectively. Suppose we have a random sample $(y_i^{(l)}, \delta_i^{(l)}, \mathbf{x}_i^{(l)}, \mathbf{u}_i^{(l)})$, $i = 1, \dots, n_l$ in the l th dataset for $l = 1, \dots, L$, where $y_i^{(l)} = \min(T_i^{(l)}, C_i^{(l)})$, $C_i^{(l)}$ is the logarithm of the censoring time, and $\delta_i^{(l)} = I\{T_i^{(l)} \leq C_i^{(l)}\}$ is the censoring indicator.

Assume $y_{(1)}^{(l)} \leq \dots \leq y_{(n_l)}^{(l)}$ are the order statistics of $y_i^{(l)}$'s and $\delta_{(1)}^{(l)}, \dots, \delta_{(n_l)}^{(l)}$ are the associated censoring indicators. Similarly, let $\mathbf{x}_{(1)}^{(l)}, \dots, \mathbf{x}_{(n_l)}^{(l)}$ and $\mathbf{u}_{(1)}^{(l)}, \dots, \mathbf{u}_{(n_l)}^{(l)}$ be the associated covariates of the ordered $y_i^{(l)}$'s. The Kaplan-Meier weights can be computed as

$$\omega_1^{(l)} = \frac{\delta_{(1)}^{(l)}}{n_l}, \quad \omega_i^{(l)} = \frac{\delta_{(i)}^{(l)}}{n_l - i + 1} \prod_{j=1}^{i-1} \left(\frac{n_l - j}{n_l - j + 1} \right)^{\delta_{(j)}^{(l)}}, \quad i = 2, \dots, n_l.$$

We center $\mathbf{x}_{(i)}^{(l)}$, $\mathbf{u}_{(i)}^{(l)}$, and $y_{(i)}^{(l)}$ with their $\omega_i^{(l)}$ -weighted means, respectively, ie,

$$\bar{\mathbf{x}}_\omega^{(l)} = \frac{\sum_{i=1}^{n_l} \omega_i^{(l)} \mathbf{x}_{(i)}^{(l)}}{\sum_{i=1}^{n_l} \omega_i^{(l)}}, \quad \bar{\mathbf{u}}_\omega^{(l)} = \frac{\sum_{i=1}^{n_l} \omega_i^{(l)} \mathbf{u}_{(i)}^{(l)}}{\sum_{i=1}^{n_l} \omega_i^{(l)}}, \quad \bar{y}_\omega^{(l)} = \frac{\sum_{i=1}^{n_l} \omega_i^{(l)} y_{(i)}^{(l)}}{\sum_{i=1}^{n_l} \omega_i^{(l)}}.$$

We replace the observations in the l th dataset with $\sqrt{\omega_i^{(l)}}(\mathbf{x}_{(i)}^{(l)} - \bar{\mathbf{x}}_\omega^{(l)})$, $\sqrt{\omega_i^{(l)}}(\mathbf{u}_{(i)}^{(l)} - \bar{\mathbf{u}}_\omega^{(l)})$, and $\sqrt{\omega_i^{(l)}}(y_{(i)}^{(l)} - \bar{y}_\omega^{(l)})$, respectively. Moreover, the rest of the operation is the same as in Section 2.2.

3 | SIMULATION

Three datasets are simulated ($L = 3$) with sample sizes $n_1 = 180$, $n_2 = 170$, and $n_3 = 150$ (total sample size $n = 500$). The number of environmental predictors is $q = 2$, and the number of genetic predictors is $p = 50$ and 100 .

For the parametric part, $x_j^{(l)}$'s are independently generated from $N(0, 1)$. We consider five different scenarios (S1 to S5) for the coefficient vector $(\beta_1^{(l)}, \dots, \beta_p^{(l)}, \alpha_{12}^{(l)}, \dots, \alpha_{p-1,p}^{(l)})$. The values of nonzero parameters are shown in Table 1. For each scenario, there are 10 nonzero main effect coefficients and 10 nonzero interaction coefficients. Overall, the differences among three datasets in S1 are bigger than these in S2. Both S1 and S2 have smaller differences than S3, where some main effects and interactions have different signs in different datasets. For S4, some effects are canceled out in the sense that $\beta_j^{(1)} + \beta_j^{(2)} + \beta_j^{(3)} = 0$ or $\alpha_{jj'}^{(1)} + \alpha_{jj'}^{(2)} + \alpha_{jj'}^{(3)} = 0$ for some j and j' . For S5, we consider the case where the strong hierarchical constraint is violated with nonzero interaction $\mathbf{x}_1 \mathbf{x}_{11}$ but zero main effect \mathbf{x}_{11} . In addition, we consider the

TABLE 1 Simulation scenarios: the true values of nonzero parameters

| | | Main Effects | | | | | | | | | | Interactions | | | | | | | | | |
|----|-------|--------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|--------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|-----------------|-----------------|
| | | β_1 | β_2 | β_3 | β_4 | β_5 | β_6 | β_7 | β_8 | β_9 | β_{10} | α_{12} | α_{13} | α_{14} | α_{23} | α_{24} | α_{34} | α_{56} | α_{57} | α_{67} | α_{89} |
| S1 | Data1 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1.5 | 1.5 | 1 | 1 | 1 | 0.5 | 0.5 | 0.5 | 0.5 |
| | Data2 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 0.5 | 0.5 | 0.5 | 1.5 | 1.5 | 1.5 | 1.5 |
| | Data3 | 1 | 1 | 1 | 1 | 1 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 1.5 | 1 | 1 | 0.5 | 0.5 | 0.5 | 2 | 2 | 2 | 2 |
| | | Main Effects | | | | | | | | | | Interactions | | | | | | | | | |
| | | β_1 | β_2 | β_3 | β_4 | β_5 | β_6 | β_7 | β_8 | β_9 | β_{10} | α_{12} | α_{13} | α_{14} | α_{23} | α_{24} | α_{34} | α_{56} | α_{57} | α_{67} | α_{89} |
| S2 | Data1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1.5 | 1.5 | 1 | 1 | 1 | 0.5 | 0.5 | 0.5 | 0.5 |
| | Data2 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 0.5 | 0.5 | 0.5 | 1.5 | 1.5 | 1.5 | 1.5 |
| | Data3 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 1.5 | 1 | 1 | 0.5 | 0.5 | 0.5 | 1 | 1 | 1 | 1 |
| | | Main Effects | | | | | | | | | | Interactions | | | | | | | | | |
| | | β_1 | β_2 | β_3 | β_4 | β_5 | β_6 | β_7 | β_8 | β_9 | β_{10} | α_{12} | α_{13} | α_{14} | α_{23} | α_{24} | α_{34} | α_{56} | α_{57} | α_{67} | α_{89} |
| S3 | Data1 | 2 | 2 | 2 | 2 | 2 | -1 | 1 | 1 | 1 | 1 | 2 | 1.5 | 1.5 | -1 | 1 | 1 | 0.5 | 0.5 | 0.5 | 0.5 |
| | Data2 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 0.5 | 0.5 | 0.5 | 1.5 | 1.5 | 1.5 | 1.5 |
| | Data3 | 1 | 1 | 1 | 1 | -1 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 1.5 | 1 | 1 | 0.5 | 0.5 | 0.5 | -2 | 2 | 2 | 2 |
| | | Main Effects | | | | | | | | | | Interactions | | | | | | | | | |
| | | β_1 | β_2 | β_3 | β_4 | β_5 | β_6 | β_7 | β_8 | β_9 | β_{10} | α_{12} | α_{13} | α_{14} | α_{23} | α_{24} | α_{34} | α_{56} | α_{57} | α_{67} | α_{89} |
| S4 | Data1 | 2 | 2 | -2 | -2 | 2 | -1 | 1 | 1 | 1 | 1 | 2 | -1.5 | 1.5 | 1 | -1 | 1 | 0.5 | 0.5 | 0.5 | 0.5 |
| | Data2 | -1.5 | -1 | 1 | -1.5 | 1.5 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 0.5 | 0.5 | 0.5 | 1.5 | 1.5 | 1.5 | 1.5 |
| | Data3 | -0.5 | -1 | 1 | -0.5 | -1 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 1.5 | 0.5 | 1 | 0.5 | 0.5 | 0.5 | -2 | 2 | 2 | 2 |
| | | Main Effects | | | | | | | | | | Interactions | | | | | | | | | |
| | | β_1 | β_2 | β_3 | β_4 | β_5 | β_6 | β_7 | β_8 | β_9 | β_{10} | α_{12} | α_{13} | α_{14} | α_{15} | α_{16} | α_{17} | α_{18} | α_{19} | $\alpha_{1,10}$ | $\alpha_{1,11}$ |
| S5 | Data1 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1.5 | 1.5 | 1 | 1 | 1 | 0.5 | 0.5 | 0.5 | 0.5 |
| | Data2 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 0.5 | 0.5 | 0.5 | 1.5 | 1.5 | 1.5 | 1.5 |
| | Data3 | 1 | 1 | 1 | 1 | 1 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 1.5 | 1 | 1 | 0.5 | 0.5 | 0.5 | 2 | 2 | 2 | 2 |

effect of correlation structure among the covariates. In S6, we generate $\mathbf{x}^{(l)}$ from the normal distribution with zero mean 0, unit variance, and correlation $0.5^{|j-k|}$ ($j, k = 1, \dots, p$), and other settings are the same as S1.

For the nonparametric part, we consider the following scenarios. (1) For S1 to S6, the nonparametric functions for the three datasets are the same

$$f_1^{(1)}(u) = f_1^{(2)}(u) = f_1^{(3)}(u) = \sin(4\pi u),$$

$$f_2^{(1)}(u) = f_2^{(2)}(u) = f_2^{(3)}(u) = 10 \{ \exp(-3.25u) + 4 \exp(-6.5u) + 3 \exp(-9.75u) \}.$$

(2) Different datasets have different nonparametric functions

$$f_1^{(1)}(u) = \sin(4\pi u), \quad f_2^{(1)}(u) = 10 \{ \exp(-3.25u) + 4 \exp(-6.5u) + 3 \exp(-9.75u) \},$$

$$f_1^{(2)}(u) = \sin(4\pi u), \quad f_2^{(2)}(u) = 3(2u - 1)^2,$$

$$f_1^{(3)}(u) = 10 \{ \exp(-3.25u) + 4 \exp(-6.5u) + 3 \exp(-9.75u) \}, \quad f_2^{(3)}(u) = 3(2u - 1)^2,$$

and we use S7 to denote this scenario where the parametric part is the same as in S1. For all settings, $u_k^{(l)}$'s are generated independently from the uniform distribution on $[0, 1]$, and the random errors $\epsilon^{(l)}$'s are generated from $N(0, 1)$.

We compare the proposed method (denoted as M3) to two alternatives. (1) We analyze each dataset separately using the interaction analysis under semiparametric model but ignore the homogeneous sparsity structure condition of different datasets (denoted as M1); (2) We analyze the three datasets together by combining them into one big dataset (denoted as M2), resulting in the same estimates of parameters for different datasets. All methods use B-splines to approximate the nonparametric functions. To determine the number of knots, we use the method of Wang et al²¹ and perform all possible spline approximations with 0 to 7 equally spaced internal knots for each nonparametric component, and calculate the prediction error (PE) defined as

$$PE = \frac{1}{n} \sum_{m=1}^M \| \mathbf{y}^{(m)} - \hat{\mathbf{y}}^{(m)} \|_2^2.$$

TABLE 2 Simulation results for $p = 50$

| Scenario | Method | FP | | FN | | MSE(SE) | | PE(SE) | RMSE(SE) |
|----------|--------|------|-------|------|-------|-------------|--------------|-------------|-------------|
| | | Main | Inter | Main | Inter | Main | Inter | | |
| S1 | M1 | 3.80 | 5.85 | 0.00 | 0.00 | 2.98(1.06) | 8.47(5.81) | 11.04(2.54) | 3.14(0.09) |
| | M2 | 1.40 | 4.25 | 0.42 | 0.30 | 4.65(0.67) | 8.23(0.84) | 12.31(0.99) | 2.15(0.14) |
| | M3 | 0.05 | 1.35 | 0.05 | 0.27 | 0.97(0.28) | 1.41(0.49) | 5.65(0.30) | 1.64(0.12) |
| S2 | M1 | 2.93 | 4.19 | 0.00 | 0.00 | 1.16(0.30) | 2.17(0.63) | 5.34(0.40) | 1.63(0.06) |
| | M2 | 0.33 | 0.33 | 0.00 | 0.00 | 1.42(0.25) | 2.26(0.24) | 5.82(0.52) | 1.14(0.10) |
| | M3 | 0.07 | 1.07 | 0.00 | 0.00 | 0.45(0.13) | 0.53(0.14) | 2.84(0.17) | 0.77(0.03) |
| S3 | M1 | 3.99 | 6.73 | 0.08 | 0.10 | 5.47(1.32) | 17.95(4.21) | 12.69(1.71) | 3.43(0.20) |
| | M2 | 1.77 | 4.85 | 0.77 | 0.70 | 11.69(0.67) | 18.82(2.04) | 17.95(2.50) | 2.38(0.28) |
| | M3 | 0.04 | 2.17 | 0.48 | 0.52 | 3.89(1.30) | 10.38(2.41) | 8.98(1.92) | 1.60(0.10) |
| S4 | M1 | 4.02 | 5.78 | 0.00 | 0.05 | 9.95(1.56) | 11.85(5.97) | 15.41(4.18) | 3.30(0.29) |
| | M2 | 0.80 | 1.23 | 4.85 | 5.88 | 14.63(2.66) | 22.15(5.53) | 26.19(3.54) | 2.43(0.323) |
| | M3 | 0.25 | 0.89 | 0.10 | 0.06 | 3.45(1.04) | 11.38(3.24) | 8.85(2.73) | 1.62(0.32) |
| S5 | M1 | 5.69 | 7.85 | 0.00 | 0.25 | 3.69(1.16) | 20.09(15.37) | 14.10(4.18) | 3.28(0.17) |
| | M2 | 2.57 | 3.18 | 0.02 | 0.36 | 4.67(0.52) | 13.51(1.54) | 13.67(1.38) | 2.15(0.23) |
| | M3 | 0.38 | 2.43 | 0.11 | 0.35 | 1.54(0.86) | 8.80(1.11) | 7.23(0.50) | 1.58(0.11) |
| S6 | M1 | 3.67 | 4.00 | 0.00 | 0.00 | 3.66(1.43) | 9.39(7.08) | 11.53(2.20) | 3.27(0.16) |
| | M2 | 0.25 | 1.00 | 0.00 | 0.02 | 4.73(0.44) | 8.17(0.73) | 12.81(1.89) | 2.23(0.33) |
| | M3 | 0.03 | 0.68 | 0.10 | 0.18 | 1.04(0.87) | 2.16(0.93) | 5.78(1.09) | 1.51(0.09) |
| S7 | M1 | 3.14 | 5.36 | 0.00 | 0.00 | 4.59(1.93) | 9.11(4.48) | 10.49(2.86) | 3.20(0.26) |
| | M2 | 2.99 | 3.70 | 0.16 | 0.64 | 6.36(0.56) | 9.38(0.75) | 11.77(1.06) | 3.38(0.07) |
| | M2+I | 3.03 | 3.71 | 0.00 | 0.00 | 6.36(0.58) | 9.37(0.76) | 11.73(1.04) | 3.38(0.07) |
| | M3 | 0.04 | 1.32 | 0.00 | 0.20 | 0.97(0.18) | 1.87(0.26) | 3.48(0.34) | 0.78(0.09) |

Abbreviations: FP: number of false positive variables; FN: number of false negative variables; MSE: mean square error for parametric part; PE: prediction error; RMSE: root of MSE for nonparametric part.

Then, we choose the optimal combination which has the smallest PE. We get the most frequently chosen combination in 100 runs for each method, and these combinations are used in the following estimation procedures.

To evaluate the methods, we consider (1) identification accuracy, which includes the numbers of false positive variables (FP) and false negative variables (FN) for main effects and interactions; (2) estimation accuracy, which includes mean squared error (MSE) for parametric part and root of MSE (RMSE) for nonparametric part, which is defined as

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_k \|\hat{f}_k(\mathbf{u}_k) - f_k(\mathbf{u}_k)\|^2},$$

and (3) prediction performance, measured by PE.

We summarize the results for $p = 50$ and $p = 100$ in Tables 2 and 3, respectively. In each cell, the mean and standard deviation (in parentheses) values are based on 100 replications. The main conclusions are as follows.

1. M1 produces the largest FP but the smallest FN in all scenarios. M2 improves the performance a little bit in terms of FP, yet leads to larger FN. This is because M1 ignores the commonalities while M2 does not consider the differences in coefficients among different datasets. Overall, M3 has a superior identification performance, with the smallest FP and comparable FN with M1 in most scenarios.
2. In estimation, M3 outperforms M1 and M2 with the smallest MSEs and RMSEs. Although M1 has smaller MSEs for the main effect coefficients than M2, it may perform quite badly in estimating the interaction coefficients, especially for the case of $p = 100$. Besides, M2 shows better fitting performance for the nonparametric functions than M1 for scenarios S1 to S6. However, for S7, M1 yields a smaller RMSE than M2 when $p = 50$; this is because M1 analyzes the datasets separately, while M2 combines all datasets. M2 enjoys the advantage of fewer parameters in S1 to S6, but this advantage leads to its inferior performance in S7. To address the differences across datasets in S7 when applying M2, we consider the combined analysis using the indices of the datasets as covariates (denoted as M2+I). Here, we always keep the index variables in the model without selection. The results show that M2+I has slightly better performance than M2, but its performance is worse than that of the proposed method.
3. M3 has the best prediction performance in all scenarios. When $p = 50$, M1 yields a smaller PE than M2, but it produces the largest PE when $p = 100$ in most scenarios. This is because the dimension of the parametric parts increases from 1275 to 5050 with p changing from 50 to 100, but M1 suffers in the higher dimension case because it

TABLE 3 Simulation results for $p = 100$

| Scenario | Method | FP | | FN | | MSE(SE) | | PE(SE) | RMSE(SE) |
|----------|--------|------|-------|------|-------|-------------|--------------|--------------|------------|
| | | Main | Inter | Main | Inter | Main | Inter | | |
| S1 | M1 | 7.31 | 10.01 | 0.00 | 0.00 | 4.68(1.84) | 13.98(7.53) | 15.01(5.44) | 3.34(0.22) |
| | M2 | 5.31 | 7.42 | 0.00 | 0.00 | 4.62(0.40) | 8.03(0.56) | 12.56(1.35) | 2.26(0.24) |
| | M3 | 0.04 | 0.81 | 0.12 | 0.23 | 1.19(0.59) | 2.19(0.85) | 6.00(1.59) | 1.56(0.10) |
| S2 | M1 | 5.49 | 8.99 | 0.00 | 0.00 | 2.22(1.05) | 8.43(2.56) | 10.68(2.55) | 3.49(0.57) |
| | M2 | 4.63 | 6.54 | 0.00 | 0.00 | 2.59(0.66) | 5.34(0.39) | 8.53(1.68) | 2.01(0.31) |
| | M3 | 0.04 | 0.69 | 0.08 | 0.11 | 1.02(0.56) | 1.59(0.55) | 4.11(0.96) | 1.48(0.18) |
| S3 | M1 | 7.74 | 13.00 | 0.00 | 0.00 | 5.50(1.65) | 26.35(10.34) | 19.74(6.75) | 3.30(0.18) |
| | M2 | 3.63 | 8.63 | 0.63 | 0.63 | 11.71(0.91) | 19.11(5.46) | 17.39(2.07) | 2.15(0.37) |
| | M3 | 0.00 | 2.53 | 0.00 | 0.47 | 3.83(2.21) | 9.39(6.07) | 8.88(1.77) | 1.58(0.17) |
| S4 | M1 | 8.76 | 10.55 | 0.00 | 0.00 | 16.94(1.10) | 59.98(37.39) | 36.71(23.45) | 3.30(0.17) |
| | M2 | 4.28 | 5.41 | 2.17 | 3.62 | 38.13(1.14) | 73.85(61.28) | 44.44(39.74) | 2.36(0.43) |
| | M3 | 4.17 | 1.59 | 0.14 | 1.10 | 17.67(1.22) | 36.36(7.43) | 25.34(3.98) | 1.60(0.24) |
| S5 | M1 | 8.22 | 11.43 | 0.00 | 0.00 | 4.68(1.84) | 13.98(7.53) | 20.14(7.33) | 3.67(0.41) |
| | M2 | 6.35 | 9.44 | 0.00 | 0.00 | 5.43(0.94) | 10.76(0.89) | 15.89(2.63) | 2.57(0.21) |
| | M3 | 0.14 | 1.19 | 0.02 | 0.15 | 2.32(0.98) | 3.67(1.21) | 8.21(1.78) | 1.47(0.20) |
| S6 | M1 | 6.85 | 9.16 | 0.00 | 0.00 | 3.48(1.30) | 16.27(8.84) | 15.53(6.03) | 3.33(0.21) |
| | M2 | 0.42 | 0.95 | 0.00 | 0.00 | 4.54(0.47) | 8.21(0.65) | 12.42(0.88) | 2.19(0.23) |
| | M3 | 0.05 | 1.05 | 0.26 | 0.58 | 1.68(0.95) | 4.01(1.21) | 6.80(3.18) | 1.62(0.28) |
| S7 | M1 | 7.20 | 13.35 | 0.00 | 0.00 | 6.52(2.62) | 14.85(8.77) | 11.08(3.72) | 3.20(0.26) |
| | M2 | 6.55 | 9.96 | 0.00 | 0.00 | 6.91(0.66) | 9.83(0.99) | 12.18(1.08) | 3.09(0.08) |
| | M2+I | 6.50 | 9.08 | 0.00 | 0.00 | 6.91(0.63) | 9.82(0.96) | 12.14(1.08) | 3.09(0.08) |
| | M3 | 0.25 | 0.28 | 0.00 | 0.03 | 1.02(0.27) | 1.92(0.43) | 3.56(0.35) | 1.77(0.09) |

Abbreviations: FP: number of false positive variables; FN: number of false negative variables; MSE: mean square error for parametric part; PE: prediction error; RMSE: root of MSE for nonparametric part.

uses less data and analyzes the datasets separately. It is interesting to note that the exception is S4, where M2 performs quite badly because some effects are canceled out and it cannot identify the effects by combining all datasets together.

- The results of S1 to S4 show that there is a decrease in identification, estimation, as well as prediction accuracy as the differences in coefficients among datasets increase for all methods. Moreover, for S5, where the data settings are the same as S1 except that the hierarchical assumption is violated for one interaction, the performance of the three methods are slightly worse.
- The correlation structure has no obvious effects on the performance of the methods.

In addition, we plot the nonparametric fittings and 95% pointwise confidence intervals based on bootstrap. Here, we only provide the plots of S1 and S7 for $p = 50$ in Figures 2 and 3, respectively. The figures for other scenarios can be found in the Appendix. In each figure, the true function is represented by the black solid curve, and the estimated functions based on M1, M2, and M3 are represented by blue dotted-dashed curve, green dotted curve, and red-dashed curve, respectively. The corresponding 95% confidence intervals are represented in shades. We observe from Figure 2 that the estimated functions using the three methods are very close to the true functions. In most regions, the confidence intervals of M2 are narrower than these of M1 and M3. Figure 3 shows that M2 produces substantially biased estimates. This is expected as M2 misspecifies the different functions in different datasets as the same function. The estimated curves of M1 and M3 are very close to the true curves but the confidence intervals of M1 are wider. Overall, M3 outperforms M1 and M2 in estimating the nonparametric functions.

Lastly, we conduct another simulation to evaluate the effect of the strong hierarchy on the algorithms. We use the same setting as S1. We use the three methods to analyze the datasets, but do variable selection on γ directly without the reparametrization of interaction parameters. The results are presented in Table 4. It can be seen that the methods without constraints generally perform not as well as the methods with constraints, especially in identifying interactions. M3 is still superior to M1 and M2 in identification, estimation and prediction accuracy, but the values of indices show an increase as compared in S1.

As for computation cost, the calculation time of M2 is the shortest because it has the least number of parameters. However, it performs the worst in prediction and estimation. Although M1 saves about 70% computation time, M3 is also attractive because it reduces MSE by about 85% for both the main effects and interactions, and reduces PE and RMSE about 65% and 45%, respectively.

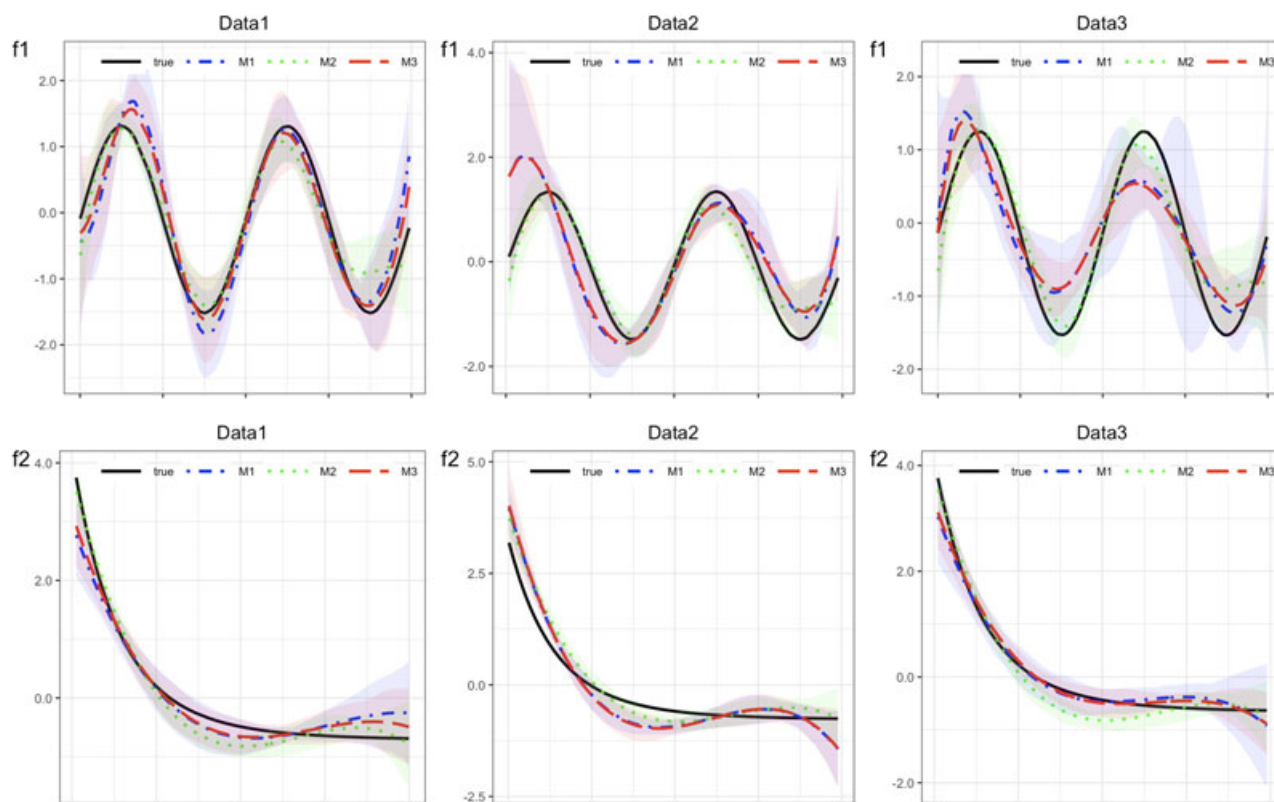


FIGURE 2 Estimated results for nonparametric functions under S1 with $p = 50$ [Colour figure can be viewed at wileyonlinelibrary.com]

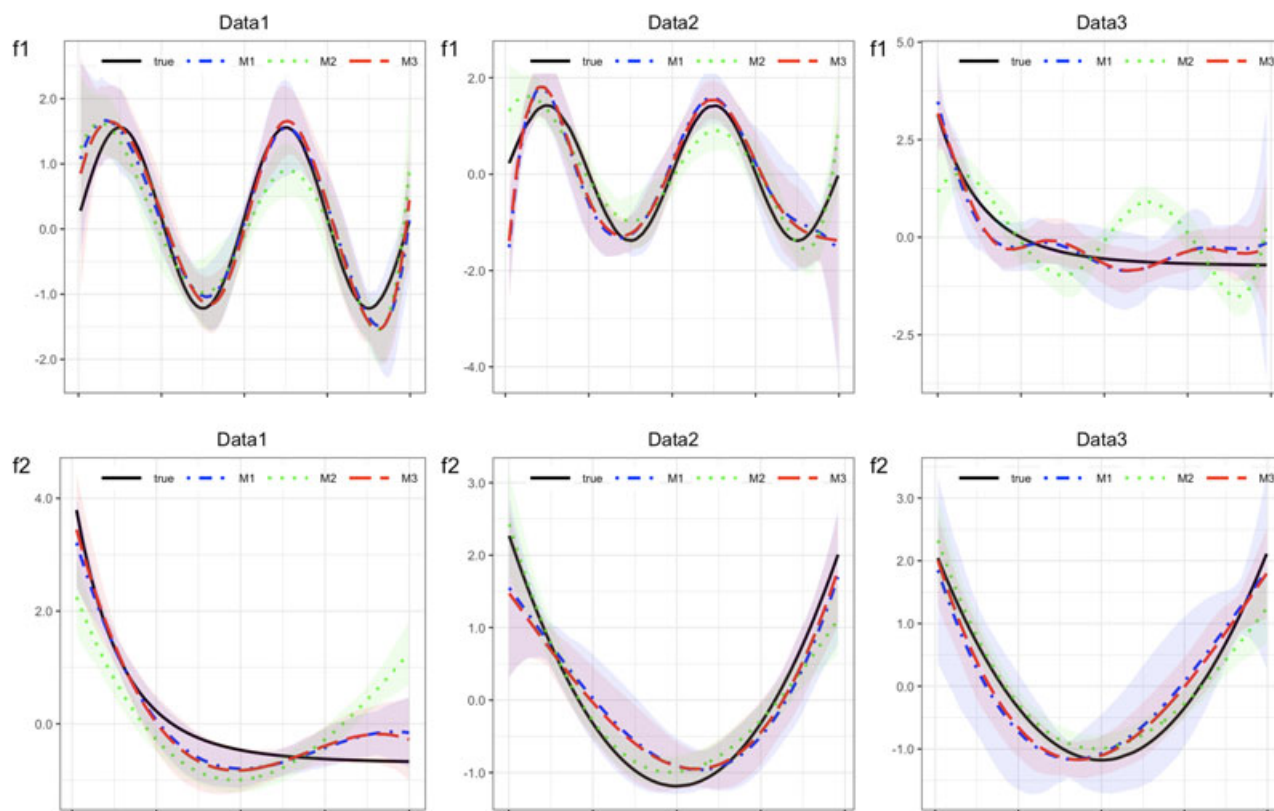


FIGURE 3 The estimated results for nonparametric functions under S7 with $p = 50$ [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 4 Simulation results for the methods without strong hierarchical constraint

| <i>p</i> | Method | FP | | FN | | MSE(SE) | | PE(SE) | RMSE(SE) |
|----------------|--------|------|-------|------|-------|------------|-------------|-------------|------------|
| | | Main | Inter | Main | Inter | Main | Inter | | |
| <i>p</i> = 50 | M1 | 4.83 | 6.79 | 0.00 | 0.29 | 4.37(1.45) | 10.83(6.35) | 17.95(7.35) | 3.53(0.63) |
| | M2 | 2.24 | 4.16 | 0.00 | 0.09 | 6.49(1.35) | 10.24(4.84) | 20.09(5.34) | 2.56(0.73) |
| | M3 | 0.87 | 1.42 | 0.11 | 0.42 | 2.79(1.03) | 7.95(2.35) | 10.41(2.84) | 1.53(0.15) |
| <i>p</i> = 100 | M1 | 7.18 | 12.48 | 1.08 | 1.19 | 5.34(1.84) | 15.88(5.35) | 18.24(6.24) | 3.65(0.32) |
| | M2 | 4.23 | 8.23 | 1.98 | 2.32 | 6.16(0.59) | 9.94(0.87) | 14.13(1.89) | 2.48(0.14) |
| | M3 | 0.29 | 2.35 | 1.42 | 1.82 | 1.98(0.81) | 3.32(1.04) | 7.84(1.63) | 1.43(0.22) |

Abbreviations: FP: number of false positive variables; FN: number of false negative variables; MSE: mean square error for parametric part; PE: prediction error; RMSE: root of MSE for nonparametric part.

4 | DATA ANALYSIS

4.1 | Analysis of cutaneous melanoma data

We illustrate the proposed method by analyzing the SKCM data from TCGA. Skin Cutaneous Melanoma is a type of malignant tumor derived from melanocytes, which is common in skin and is also found in the mucosa, choroid, and other parts of the eyes. Melanoma is the most malignant tumor of skin and prone to distant metastasis. In previous research, several genes have been discovered as involved in the progression of cutaneous melanoma.²⁶ However, most previous studies fail to capture the similarity in sparsity structures and simultaneously accommodate differences across different cancer stages.

There are 340 samples, with 78 in stage I, 129 in stage II, and 133 in stages III and IV. The response variable of interest is Breslow thickness, which is an important prognostic marker. A total of 18 353 genes are measured in all three datasets. To improve the stability of analysis, we construct marginal regression with each gene under combined analysis and select the top 100 genes with the smallest *p*-value for downstream analysis. Moreover, age is included as the nonparametric component. The response variable is log transformed. Normalization is conducted for each dataset separately to standardize each genetic measurement and age to have zero mean and unit variance.

The selection results are provided in Table 5. The proposed method identifies 12 main effects and six interactions, which have both common and different effects from the alternatives. We also present the detailed estimation results in Table 6. It can be seen that M1 selects different predictors for different stages, while M2 yields the same estimates for the selected terms in different stages. Different from M1 and M2, the proposed method selects the same terms for different stages, but yields different coefficients. For example, gene *sine oculis homeobox homolog 1* (*SIX1*) is selected in all stages by the three methods. As indicated in Graziano,²⁷ *SIX1* encodes a homeoprotein transcription factor and is an important developmental regulator during embryogenesis. A profound nuclear to cytoplasmic shift of *SIX1* is shown to be accompanied with melanoma progression and correlated with poor 5-year survival. Moreover, higher cytoplasmic *SIX1* is associated with increased tumor thickness, lower nuclear *SIX1* with ulceration and histological satellitosis, and both with advanced AJCC stages and nodular melanoma. Gene *TMEM156*, which is also selected by the three methods, is upregulated in human cancer specimens as determined by immunostaining of human normal and cancer tissue arrays.²⁸ *Semaphorin 3A* (*SEMA3A*) has negative coefficients in all datasets. The *SEMA3A* acts as a potent suppressor of tumor angiogenesis in various cancer models. Moreover, it has been demonstrated that *SEMA3A* suppresses tumor growth and metastasis using multiple in vitro and in vivo approaches in mice melanoma models.²⁹ However, M1 fails to select it in the first stage. Besides, gene *EIF2S2*, which is selected by the proposed method but not by M2, has been shown to be significantly correlated with pigmentation phenotypes, and is significantly differentially expressed in the skin epidermis.³⁰

TABLE 5 Skin cutaneous melanoma data analysis: numbers of selected and overlapping effects. In each cell: dataset 1/dataset 2/dataset 3

| | | M1 | M2 | M3 |
|--------------|----|----------|-------|----------|
| | | 18/13/13 | 3/8/6 | 5/4/5 |
| | | 13/13/13 | 5/5/5 | 12/12/12 |
| | | 4/7/7 | 0/0/0 | 0/0/2 |
| Interactions | M1 | 4/7/7 | 0/0/0 | 0/0/2 |
| | M2 | 5/5/5 | 1/1/1 | 6/6/6 |
| | M3 | 6/6/6 | | |

TABLE 6 Analysis of skin cutaneous melanoma data: estimated coefficients for main effects and interactions

| | M1 | | M2 | | M3 | | |
|------------------|--------|--------|--------|-----------|--------|--------|--------|
| | Data1 | Data2 | Data3 | Data1/2/3 | Data1 | Data2 | Data3 |
| AEN | | 0.001 | 0.021 | 0.029 | -0.031 | 0.143 | 0.182 |
| C15ORF54 | | | | | 0.083 | 0.127 | 0.176 |
| CA7 | | 0.083 | | | | | |
| CBARP | | | 0.025 | | | | |
| CCNB3 | 0.001 | | 0.008 | | | | |
| CGB2 | | | | 0.013 | -0.014 | 0.204 | -0.234 |
| COX7A2L | 0.058 | | | 0.002 | | | |
| CPNE6 | | | | | 0.036 | 0.174 | 0.060 |
| CRELD2 | | | | 0.067 | | | |
| DENND6A | -0.022 | | | | | | |
| DIP2C | | | | | 0.024 | -0.016 | -0.099 |
| DNAH17 | | 0.012 | | 0.042 | | | |
| DPY30 | | | -0.061 | | | | |
| DUX4L6 | -0.001 | | | | | | |
| EIF2S2 | -0.002 | | | | 0.102 | 0.099 | -0.021 |
| GJA10 | -0.054 | | | | -0.123 | 0.030 | 0.154 |
| GPAT2 | 0.007 | | | | | | |
| HNRNPC | | | -0.029 | -0.048 | | | |
| KCNQ3 | | -0.028 | | -0.010 | | | |
| LCN15 | -0.083 | | | | | | |
| LINC00479 | | -0.004 | | | | | |
| LIPI | | -0.034 | | | | | |
| MYH11 | -0.002 | -0.033 | | | | | |
| OR5C1 | 0.024 | | | | | | |
| PRAMEF10 | | | 0.022 | | | | |
| PVALB | | | | | -0.103 | -0.539 | -0.144 |
| RFPL3S | -0.054 | | | | | | |
| RPL18A | -0.018 | | | | | | |
| RPL39L | | | 0.033 | | | | |
| SEMA3A | | -0.018 | -0.165 | -0.108 | -0.060 | -0.121 | -0.306 |
| SERP2 | | 0.004 | 0.024 | 0.054 | | | |
| SIX1 | 0.046 | -0.075 | -0.004 | -0.076 | 0.114 | -0.141 | -0.024 |
| SLC7A8 | | | | 0.005 | | | |
| SPATA31C1 | 0.013 | | -0.010 | | 0.149 | -0.140 | -0.342 |
| SPATA31D3 | -0.052 | -0.001 | | | | | |
| TBC1D3C | | 0.008 | | 0.024 | | | |
| TFPI2 | -0.027 | | | | | | |
| TMA16 | | | -0.027 | | | | |
| TMEM156 | -0.005 | -0.138 | -0.011 | -0.084 | 0.033 | -0.220 | -0.280 |
| TRIM41 | -0.009 | | | | | | |
| AEN × CGB2 | | | | -0.005 | -0.078 | -0.068 | 0.545 |
| AEN × CRELD2 | | | | 0.032 | | | |
| AEN × TMEM156 | | | 0.001 | | 0.038 | -0.075 | 0.138 |
| CA7 × SEMA3A | | 0.010 | | | | | |
| CBARP × SERP2 | | | 0.114 | | | | |
| CPNE6 × PVALB | | | | | -0.074 | 0.266 | 0.073 |
| DIP2C × GJA10 | | | | | -0.032 | 0.041 | 0.167 |
| DNAH17 × HNRNPC | | | | -0.004 | | | |
| DNAH17 × KVNQ3 | | -0.086 | | | | | |
| DNAH17 × TMEM156 | | -0.051 | | | | | |
| EIF2S2 × PVALB | | | | | 0.155 | 0.673 | 0.138 |
| EIF2S2 × GPAT2 | 0.001 | | | | | | |
| EIF2S2 × RPL18A | 0.001 | | | | | | |
| GJA10 × RPL18A | 0.013 | | | | | | |

(Continues)

TABLE 6 (Continued)

| | | | | |
|---------------------|--------|-------|--------|--------|
| HNRNPC × SERP2 | -0.038 | | | |
| HNRNPC × TMA16 | 0.009 | | | |
| KCNQ3 × LIPI | -0.043 | | | |
| KCNQ3 × TMEM156 | 0.024 | | | |
| LIPI × SEMA3A | -0.020 | | | |
| PRAMEF10 × SEMA3A | 0.033 | | | |
| RPL18A × SPATA31D3 | 0.014 | | | |
| SEMA3A × TMEM156 | 0.004 | | | |
| SERP2 × TMA16 | 0.056 | | | |
| SERP2 × SIX1 | -0.124 | | | |
| SERP2 × TMEM156 | -0.122 | | | |
| SPATA31C1 × TMEM156 | 0.002 | 0.132 | -0.015 | -0.358 |

Abbreviations: SEMA3A, semaphorin 3A; SIX1, sineoculis homeobox homolog 1.

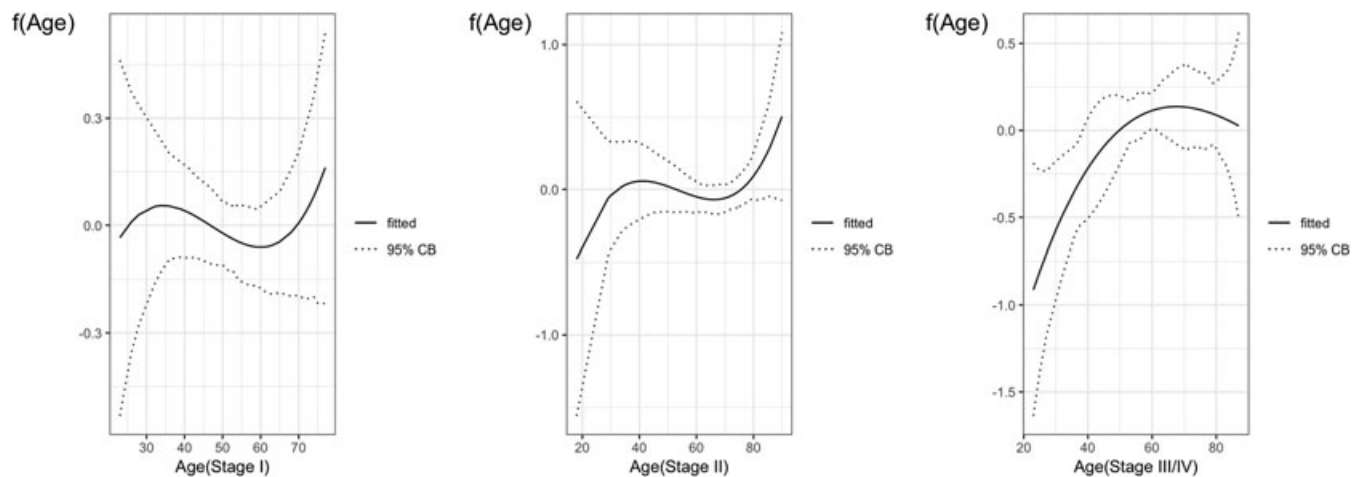


FIGURE 4 Skin cutaneous melanoma data: The estimated curves for age. In each plot, the solid line is the estimate and the dashed lines are the 95% confidence interval

Besides, we plot the estimates of nonparametric functions in Figure 4. Clearly, age has different effects on Breslow thickness in different stages. Notice that the confidence intervals are wider around the two end points, which is common in nonparametric estimation due to the small sample sizes. Therefore, we mainly focus on the middle segments of the curves. The estimated curves for stage I and II show a downward trend for patients with age between 40 and 60, but an upward trend for age after 60. According to the Cutaneous Oncology Program, under which all patients had stage I or stage II disease, a greater percentage of patients with melanoma over the age of 65 had ulcerated lesions and a greater percentage of thick lesions at diagnosis.³¹ Besides, the estimated curve for stage III/IV looks like a quadratic function and indicates that Breslow thickness increases along with age before 70 and declines only slightly with age after 70. This is similar to the previous studies which demonstrated that Breslow thickness increases with aging.³²

To further assess the identification and prediction accuracy of the methods, we randomly split the data into two parts, ie, three-quarters of the subjects form the training dataset and the rest form the testing dataset. The average PEs over 100 repetitions are computed for the three methods, which are M1: 5.366 (6.852), M2: 3.084 (2.870), and M3: 1.636 (1.460), respectively. The proposed method has the best prediction performance. In addition, we calculate observed occurrence index (OOI), the probabilities of being identified in the 100 repetitions for the selected terms using the full data, to assess stability. Figure 5 presents the top 20 markers with the largest OOIs and their average values. The OOIs of the proposed method are comparable to those of M2, and much higher than those of M1. Moreover, the selection of the main effects is more stable than that of the interaction effects, which may be caused by the strong hierarchy constraint. The OOI of one interaction can not be larger than the OOIs of the corresponding two main effects. Consequently, the selection of the interactions is affected by the selection of the main effects, and the stability of selection for the interactions is worse than that of the main effects.

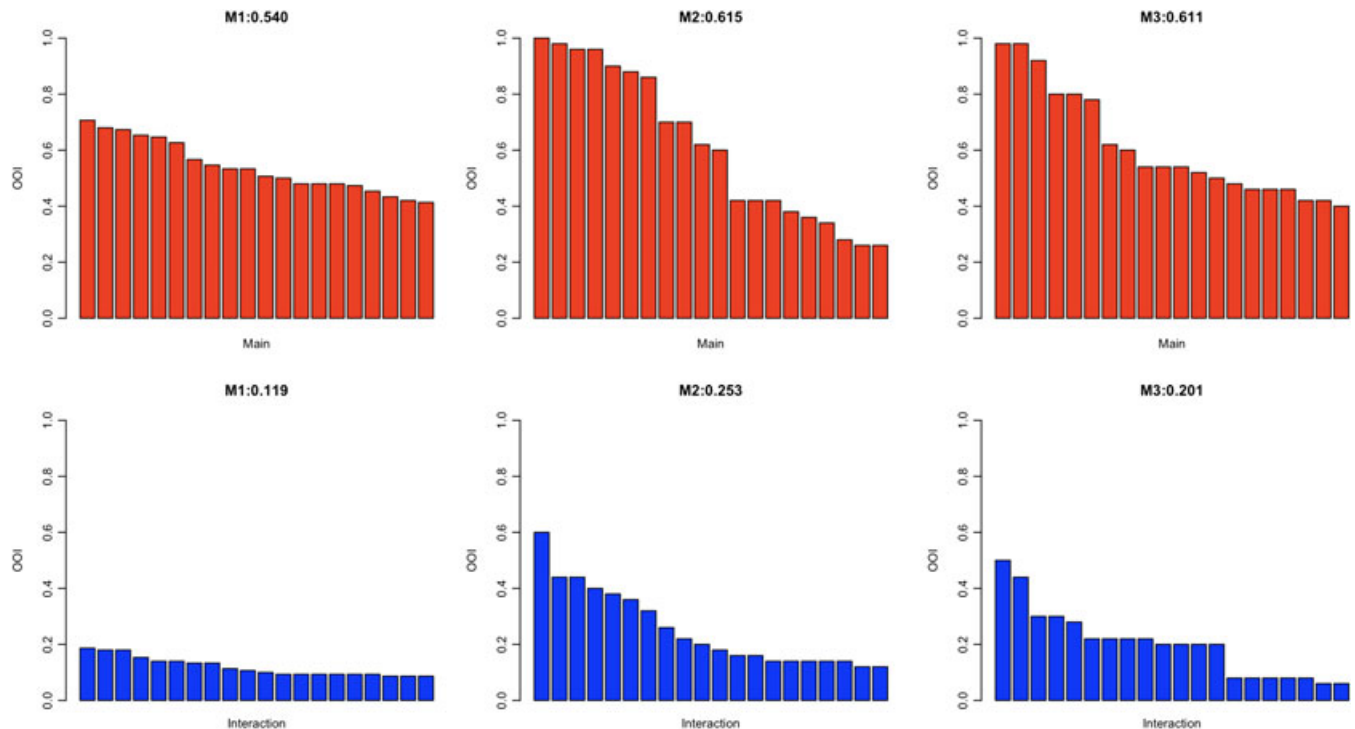


FIGURE 5 Skin cutaneous melanoma analysis: observed occurrence index (OOI). Top/red: main effects. Bottom/blue: interactions
[Colour figure can be viewed at wileyonlinelibrary.com]

4.2 | Analysis of lung cancer data

Nonsmall-cell lung cancer (NSCLC) is the most common type of lung cancer. There are many gene profiling studies searching for markers associated with the prognosis of lung cancer. Most studies either focus on one subtype of NSCLC or collect and analyze a few subtypes without considering the differences across subtypes. In TCGA, there are two lung cancer datasets on lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) respectively, which are the major NSCLC histological subtypes.^{33,34} Although LUAD and LUSC share similar mRNAs in the predicted networks, the gene and mRNA expression differences between them have also been acknowledged.³⁵

A total of 196 patients of LUAD or LUSC in stages III and IV are included in this study. Specifically, there are 106 patients from the LUAD dataset, among whom 61 died during follow-up, and 90 patients from the LUSC dataset with 49 died during follow-up. The medians of observed survival time of LUAD and LUSC are 18.2 months (ranging from 0.13 to 129.4 months) and 20.1 months (ranging from 0.03 to 132.3 months), respectively. A total of 18 277 genes are measured for all patients in the two lung cancer datasets. Using the idea of Huang et al.,³⁶ we preprocess by a marginal screening based on noncensored data and select the top 100 genes which have the highest correlations with the survival time. Standardization is conducted prior to analysis. Age is included to characterize the influence on the survival time in the form of nonparametric functions. Different from the analysis of SKCM data, we use the AFT model for the lung cancer datasets, and the response variable of interest is the survival time (measured in month).

TABLE 7 Lung cancer data analysis: numbers of selected and overlapping effects. In each cell: lung adenocarcinoma/lung squamous cell carcinoma

| | | M1 | M2 | M3 |
|--------------|--|-------|-------|-------|
| | | 13/14 | 5/5 | 7/5 |
| | | M2 | 12/12 | 9/9 |
| | | M3 | | 11/11 |
| Interactions | | M1 | M2 | M3 |
| | | 10/9 | 0/0 | 2/0 |
| | | M2 | 4/4 | 2/2 |
| | | M3 | | 10/10 |

TABLE 8 Analysis of lung cancer data: estimated coefficients for main effects and interactions

| | M1 | | M2 | M3 | |
|--------------------|--------|--------|-----------|--------|--------|
| | LUAD | LUSC | LUAD/LUSC | LUAD | LUSC |
| AASS | | -0.335 | | | |
| APOBEC3D | | | 0.055 | 0.008 | -0.010 |
| BTLA | 0.101 | | | | |
| CD302 | | -0.219 | | | |
| CHI3L2 | -0.178 | | | | |
| CLEC9A | | | 0.0192 | | |
| CLECL1 | | 0.117 | | | |
| CR2 | 0.094 | | | | |
| CYLD | 0.525 | | 0.039 | 0.026 | -0.036 |
| CYP20A1 | 0.098 | 0.287 | 0.075 | 0.022 | 0.051 |
| DHH | -0.157 | | | | |
| FAM177B | | 0.232 | | | |
| FAM238A | | | 0.078 | 0.005 | 0.029 |
| GPR183 | | -0.140 | | | |
| IL6 | | 0.162 | | | |
| LY6G5C | | 0.312 | 0.054 | 0.002 | 0.031 |
| MYLIP | | 0.358 | 0.097 | -0.009 | 0.038 |
| NINJ2 | 0.139 | 0.267 | 0.003 | 0.006 | 0.039 |
| NT5C1A | 0.100 | | | | |
| NXF3 | | 0.113 | | | |
| PDE6G | 0.128 | | 0.013 | -0.004 | 0.014 |
| PIK3IP1 | | | 0.009 | | |
| PLA2G4C | | 0.133 | 0.054 | | |
| RAMP3 | | -0.299 | | | |
| RASSF5 | -0.232 | | | | |
| RBP5 | 0.122 | | | 0.028 | -0.016 |
| SPACA5 | 0.203 | | 0.194 | 0.013 | 0.018 |
| TPSB2 | 0.137 | 0.137 | | 0.023 | 0.063 |
| AASS × MYLIP | | 0.122 | | | |
| APOBEC3D × FAM238A | | | 0.020 | | |
| APOBEC3D × SPACA5 | | | | 0.002 | 0.007 |
| APOBEC3D × MYLIP | | | 0.039 | 0.017 | 0.010 |
| APOBEC3D × NINJ2 | | | -0.004 | -0.037 | 0.057 |
| APOBEC3D × TPSB2 | | | | -0.007 | -0.019 |
| CD302 × IL6 | | -0.182 | | | |
| CD302 × CYP20A1 | | 0.124 | | | |
| CD302 × MYLIP | | 0.552 | | | |
| CHI3L2 × SPACA5 | 0.041 | | | | |
| CLECL1 × MYLIP | | -0.214 | | | |
| CR2 × RASSF5 | 0.013 | | | | |
| CYLD × CYP20A1 | -0.094 | | | -0.009 | 0.059 |
| CYLD × RBP5 | | | | -0.043 | -0.031 |
| CYLD × SPACA5 | -0.271 | | | -0.013 | 0.032 |
| CYLD × PDE6G | -0.061 | | | | |
| CYP20A1 × FAM238A | | | 0.006 | | |
| CYP20A1 × RBP5 | 0.140 | | | | |
| CYP20A1 × MYLIP | | | | -0.006 | 0.016 |
| CYP20A1 × RASSF5 | -0.005 | | | | |
| DHH × PDE6G | 0.107 | | | | |
| DHH × TPSB2 | -0.167 | | | | |
| FAM238A × MYLIP | | | | -0.004 | 0.035 |
| IL6 × RAMP3 | | 0.473 | | | |
| IL6 × NINJ2 | | -0.130 | | | |
| MYLIP × RAMP3 | | 0.463 | | | |
| PDE6G × SPACA5 | | | | -0.008 | -0.048 |
| SPACA5 × TPSB2 | -0.099 | | | | |
| TPSB2 × MYLIP | | -0.630 | | | |

Abbreviations: LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma.

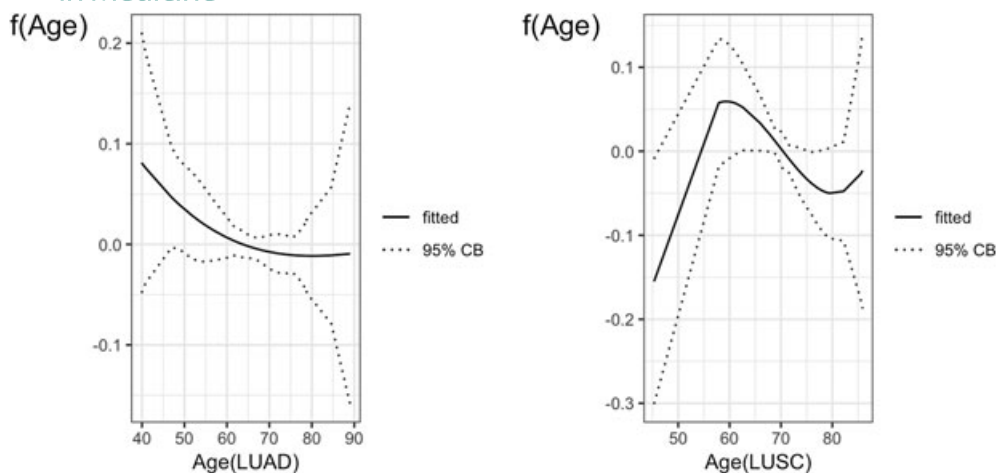


FIGURE 6 Lung cancer data: The estimated curves for age. In each plot, the solid line is the estimate and the dashed lines are the 95% confidence interval. LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma

The proposed method and two alternative methods are applied. The summary results are provided in Table 7. The proposed method identifies 11 main effects and 10 interactions, among which nine main effects and two interactions are also selected by M2. The selection results of M1 differ more from those of M3. The detailed estimation results can be found in Table 8. Note that CYP20A1 is selected in both datasets by the three methods. Cytochromes P450, a superfamily of enzymes, has been found to partially determine the metabolic capacity of lung tissues and influence the susceptibility to lung diseases, such as lung cancer and chronic obstructive pulmonary disease.³⁷ Another gene selected by all methods is NINJ2. APOBEC3D is selected by the proposed method and M2, but not M1. APOBEC enzymes play a key role in natural and adaptive immunity, which promote cancer development and clonal evolution of cancer by inducing collateral genomic damage due to their DNA deamination activity.³⁸ It has been suggested that APOBEC-mediated mutagenesis is universal throughout cancer genomes, and a significant presence of the APOBEC mutation pattern has been found in breast, cervical, bladder, head, neck, and lung cancers.³⁹ Gene TPSB2 is selected by M1 and M3, but missed by M2. Tryptase $\beta 2$, a TPSB2 gene product, stimulates the release of granulocyte chemoattractants and induces the expression of interleukin- 1β during inflammation, which is shown to be involved in the recruitment of inflammatory cells to mast cell activation sites.⁴⁰ Mast cells are commonly seen in a variety of tumors.⁴¹

Figure 6 shows the estimated nonparametric functions. It can be seen that, with the increase of age, the survival time for LUAD patients has a downward trend. However, there is a different trend for LUSC patients. Specifically, the survival time increases with age for age less than 59, and decreases with the age after 59.

The identification and prediction accuracy is also evaluated. We use log-rank test statistic to measure prediction accuracy as the response variable is right censored,⁹ and a larger value of log-rank test statistic indicates better prediction accuracy. We randomly split the data into training and testing datasets with sample size ratio 3:1. The average log-rank test statistics over 100 repetitions for M1, M2, and M3 are 24.8 (9.32), 31.3 (5.01), and 32.4 (6.07), respectively, which shows that the proposed method has better prediction accuracy than M2 and M1. Furthermore, we compute OOI to measure the identification stability and the results are plotted in Figure 7. For main effects, the proposed method and M2 have similar OOI while M1 produces the smallest OOI. For interactions, the proposed method and M1 have the same OOI, which is much larger than that of M2. Overall, the proposed method outperforms M1 and M2 in terms of identification stability.

5 | DISCUSSION

With multiple genetic datasets, integrative analysis provides an effective way of improving estimation and prediction while allowing differences across datasets. This study conducts integrative analysis with semiparametric models under the homogeneity structure, where the genetic measurements and environmental measurements are included as the parametric and nonparametric parts, respectively. The gene-gene interactions are also considered. We use B-splines to approximate the nonparametric functions while a reparametrization of interaction parameters is implemented to guarantee the strong hierarchy assumption. Moreover, we impose a group structure on the coefficients matrix and use a group Lasso penalty

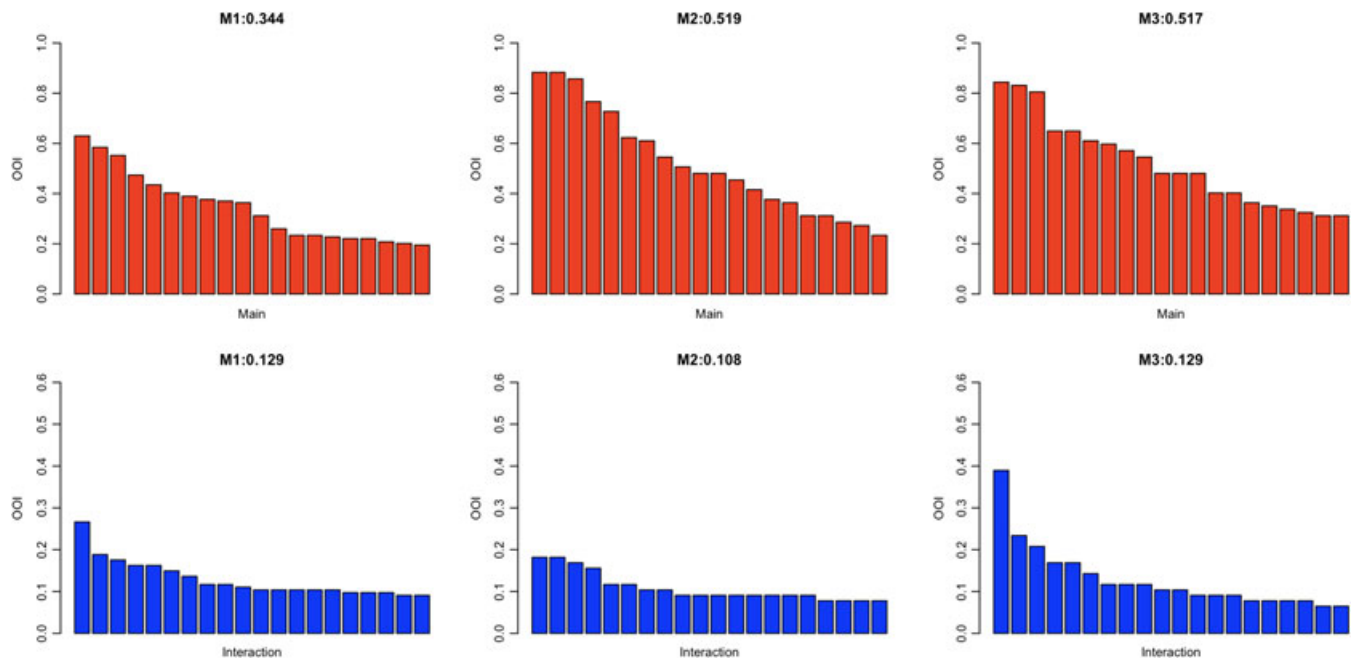


FIGURE 7 Lung cancer data analysis: observed occurrence index (OOI). Top/red: main effects. Bottom/blue: interactions [Colour figure can be viewed at wileyonlinelibrary.com]

to do variable selection, which identifies the same important genetic predictors for different datasets but allows different estimates of coefficients. An iterative algorithm is proposed to estimate the parameters of main effects, interactions, and spline regressions, which show superiority over other existing methods in simulations and real data analysis. Actually, the algorithm developed in this paper can be straightforwardly adopted to account for the hierarchical structure of higher-order interactions by reparameterization. It is interesting to conduct further studies on the hierarchical selection of higher-order interactions.

In this study, we have considered the continuous response variable under the linear model and the AFT model with right-censored data. The proposed method can be extended to handle truncated data after minor modifications. It would also be interesting to extend it to more complex data, such as longitudinal or clustered data. For future work, one may consider the selection of the nonparametric parts when there are many environmental factors. In addition, the gene-gene interactions have been discussed in this paper under the strong hierarchical constraint, but in practice, more complex structures such as gene regulatory networks still need to be examined. Moreover, the gene-environment interactions may be of interest in practice, and the varying-coefficient model for the nonlinear gene-environment interactions has been discussed.^{42–44} However, the selection and estimation for nonlinear gene-environment interactions in the framework of integrative analysis is still an open problem deserving investigation. The hypothesis testing of the homogeneity structure also deserves further attention. Besides, some theoretical justifications would be useful, including the theoretical properties of estimation and inference.

ACKNOWLEDGEMENTS

The authors thank the editor, associate editor, and referees for their insightful comments and suggestions that have led to a significant improvement of this paper. The authors also thank Ms. Ziyuan Luo for the productive discussion. This work was supported by the National Institutes of Health (CA204120, CA121974, CA196530), the National Natural Science Foundation of China (11701561, 71771211, 81774206), the MOE Project of Key Research Institute of Humanities and Social Sciences at Universities (16JJD910002), and the fund for building world-class universities (disciplines) of Renmin University of China.

ORCID

Yang Li  <https://orcid.org/0000-0002-6287-5094>

Shuangge Ma  <https://orcid.org/0000-0001-9001-4999>

REFERENCES

- Ghosh D, Chinnaiyan AM. Classification and selection of biomarkers in genomic data using LASSO. *J Biomed Biotechnol*. 2005;2005(2):147-154.
- Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol*. 1996;58(1):267-288.
- Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc*. 2006;101(476):1418-1429.
- Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc*. 2001;96(456):1348-1360.
- Zhang C-H. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat*. 2010;38(2):894-942.
- Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *J R Stat Soc Series B Stat Methodol*. 2006;68(1):49-67.
- Sun Y, Jiang Y, Li Y, Ma S. Identification of cancer omics commonality and difference via community fusion. *Statist Med*. 2018;38(7):1200-1212. <https://doi.org/10.1002/sim.8027>
- Wu M, Huang J, Ma S. Identifying gene-gene interactions using penalized tensor regression. *Statist Med*. 2018;37(4):598-610.
- Huang Y, Liu J, Yi H, Shia B-C, Ma S. Promoting similarity of model sparsity structures in integrative analysis of cancer genetic data. *Statist Med*. 2017;36(3):509-559.
- Huang Y, Zhang Q, Zhang S, Huang J, Ma S. Promoting similarity of sparsity structures in integrative analysis with penalization. *J Am Stat Assoc*. 2017;112(517):342-350.
- Liu J, Ma S, Huang J. Integrative analysis of cancer diagnosis studies with composite penalization. *Scand J Stat*. 2014;41(1):87-103.
- Habermann JK, Paulsen U, Roblick UJ, et al. Stage-specific alterations of the genome, transcriptome, and proteome during colorectal carcinogenesis. *Genes Chromosom Cancer*. 2007;46(1):10-26.
- Cheng Y, Lu J, Chen G, et al. Stage-specific prognostic biomarkers in melanoma. *Oncotarget*. 2015;6(6):4180-4189.
- Palaniappan A, Ramar K, Ramalingam S. Computational identification of novel stage-specific biomarkers in colorectal cancer progression. *PLoS ONE*. 2016;11(5):e0156665. <https://doi.org/10.1371/journal.pone.0156665>
- Ochoa MC, Marti A, Azcona C, et al. Gene-gene interaction between *PPAR γ 2* and *ADR β 3* increases obesity risk in children and adolescents. *Int J Obes*. 2004;28:S37-S41.
- Bien J, Taylor J, Tibshirani R. A lasso for hierarchical interactions. *Ann Stat*. 2013;41(3):1111-1141.
- Zhao P, Rocha G, Yu B. The composite absolute penalties family for grouped and hierarchical variable selection. *Ann Stat*. 2009;37(6A):3468-3497.
- Choi NH, Li W, Zhu J. Variable selection with the strong heredity constraint and its oracle property. *J Am Stat Assoc*. 2010;105(489):354-364.
- Du P, Cheng G, Liang H. Semiparametric regression models with additive nonparametric components and high dimensional parametric components. *Comput Stat Data Anal*. 2012;56(6):2006-2017.
- Liu X, Wang L, Liang H. Estimation and variable selection for semiparametric additive partial linear models. *Stat Sin*. 2011;21(3):1225-1248.
- Wang L, Liu X, Liang H, Carroll RJ. Estimation and variable selection for generalized additive partial linear models. *Ann Stat*. 2011;39(4):1827-1851.
- Du P, Ma S, Liang H. Penalized variable selection procedure for Cox models with semiparametric relative risk. *Ann Stat*. 2010;38(4):2092-2117.
- Ma S, Du P. Variable selection in partly linear regression model with diverging dimensions for right censored data. *Stat Sin*. 2012;22(3):1003-1020.
- Wu C, Cui Y, Ma S. Integrative analysis of gene-environment interactions under a multi-response partially linear varying coefficient model. *Statist Med*. 2014;33(28):4988-4998.
- Breheny P, Huang J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Stat Comput*. 2015;25(2):173-187.
- Fountain JW, Bale SJ, Housman DE, Dracopoli NC. Genetics of melanoma. *Cancer Surv*. 1990;9(4):654-671.
- Graziano L. *A Functional and Prognostic Study of the Transcription Factor SIX1 in Melanoma* [master's thesis]. Vancouver, Canada: The University of British Columbia; 2017. <https://open.library.ubc.ca/collections/ubctheses/24/items/1.0362582>
- Cheishvili D, Stefanska B, Yi C, et al. A common promoter hypomethylation signature in invasive breast, liver and prostate cancer cell lines reveals novel targets involved in cancer invasiveness. *Oncotarget*. 2015;6(32):33253-33268.
- Chakraborty G, Kumar S, Mishra R, Patil TV, Kundu GC. Semaphorin 3A suppresses tumor growth and metastasis in mice melanoma model. *PLoS ONE*. 2012;7(3):e33633. <https://doi.org/10.1371/journal.pone.0033633>
- Jacobs LC. *Genetic Determinants of Skin Color, Aging, and Cancer* [thesis]. Rotterdam, The Netherlands: Erasmus University Rotterdam; 2015.
- Austin PF, Cruse CW, Lyman G, Schroer K, Glass F, Reintgen DS. Age as a prognostic factor in the malignant melanoma population. *Ann Surg Oncol*. 1994;1(6):487-494.
- Chao C, Martin RCG II, Ross MI, et al. Correlation between prognostic factors and increasing age in melanoma. *Ann Surg Oncol*. 2004;11(3):259-264.
- de Bruin EC, McGranahan N, Mitter R, et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science*. 2014;346(6206):251-256.
- Wrangle J, Wang W, Koch A, et al. Alterations of immune response of non-small cell lung cancer with azacytidine. *Oncotarget*. 2013;4(11):2067-2079.

35. Sun F, Yang X, Jin Y, et al. Bioinformatics analyses of the differences between lung adenocarcinoma and squamous cell carcinoma using The Cancer Genome Atlas expression data. *Mol Med Rep.* 2017;16(1):609-616.
36. Huang J, Ma S, Xie H. Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics.* 2006;62(3):813-820.
37. Leclerc J, Tournel G, Ngangue EC-N, et al. Profiling gene expression of whole cytochrome P450 superfamily in human bronchial and peripheral lung tissues: differential expression in non-small cell lung cancers. *Biochimie.* 2010;92(3):292-306.
38. Rebhendl S, Huemer M, Greil R, Geisberger R. AID/APOBEC deaminases and cancer. *Oncoscience.* 2015;2(4):320-333.
39. Roberts SA, Lawrence MS, Klimczak LJ, et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat Genet.* 2013;45:970-976.
40. Wiedl T. *Activity-Based Proteomics: Biomarker Identification in Human Lung Adenocarcinoma* [thesis]. Zurich, Switzerland: ETH Zurich; 2011.
41. Khazaie K, Blatner NR, Khan MW, et al. The significant role of mast cells in cancer. *Cancer Metastasis Rev.* 2011;30(1):45-60.
42. Wu C, Shi X, Cui Y, Ma S. A penalized robust semiparametric approach for gene-environment interactions. *Statist Med.* 2015;34(30):4016-4030.
43. Ma S, Yang L, Romero R, Cui Y. Varying coefficient model for gene-environment interaction: a non-linear look. *Bioinformatics.* 2011;27(15):2119-2126.
44. Wu C, Zhong P-S, Cui Y. Additive varying-coefficient model for nonlinear gene-environment interactions. *Stat Appl Genet Mol Biol.* 2018;17(2). <https://doi.org/10.1515/sagmb-2017-0008>

How to cite this article: Li Y, Li R, Lin C, Qin Y, Ma S. Penalized integrative semiparametric interaction analysis for multiple genetic datasets. *Statistics in Medicine.* 2019;38:3221-3242. <https://doi.org/10.1002/sim.8172>

APPENDIX A

ADDITIONAL NONPARAMETRIC FITTED FIGURES OF THE SIMULATION

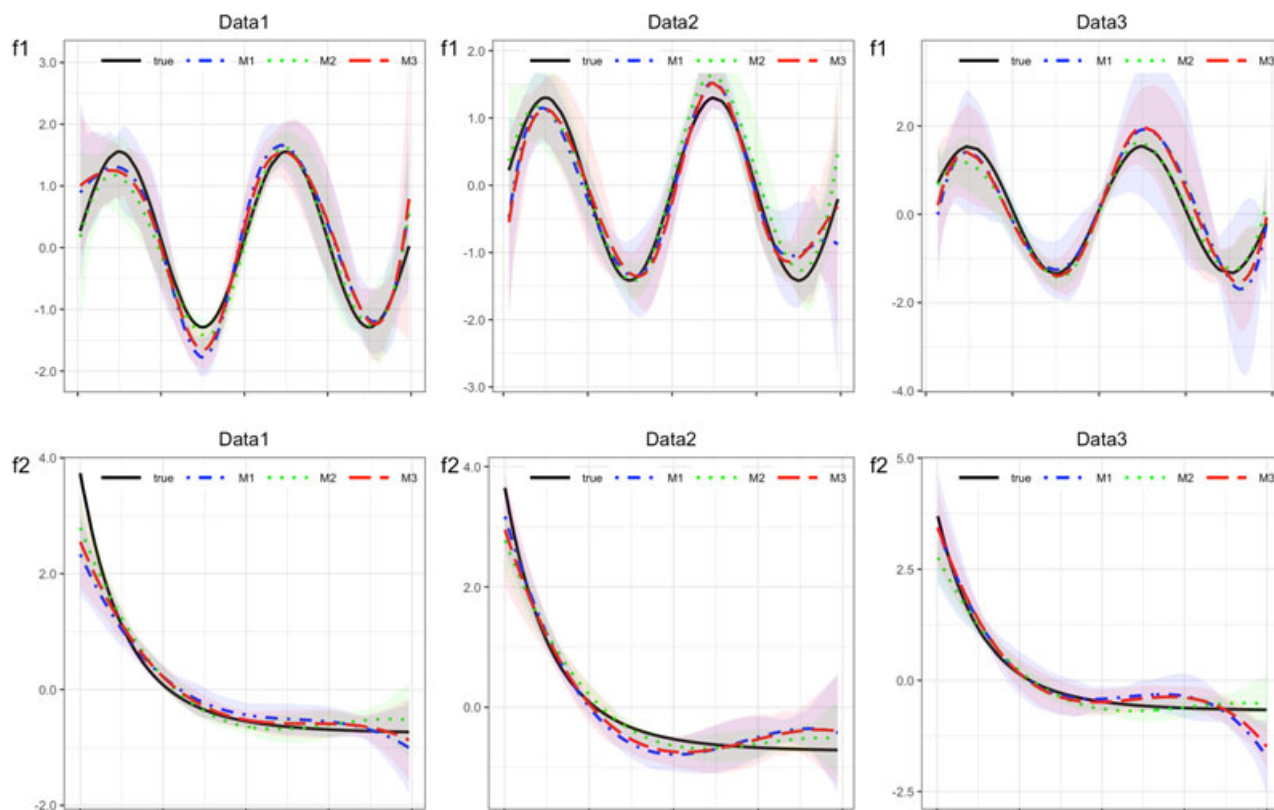


FIGURE A1 The estimated results for nonparametric functions under S2 with $p = 50$ [Colour figure can be viewed at wileyonlinelibrary.com]

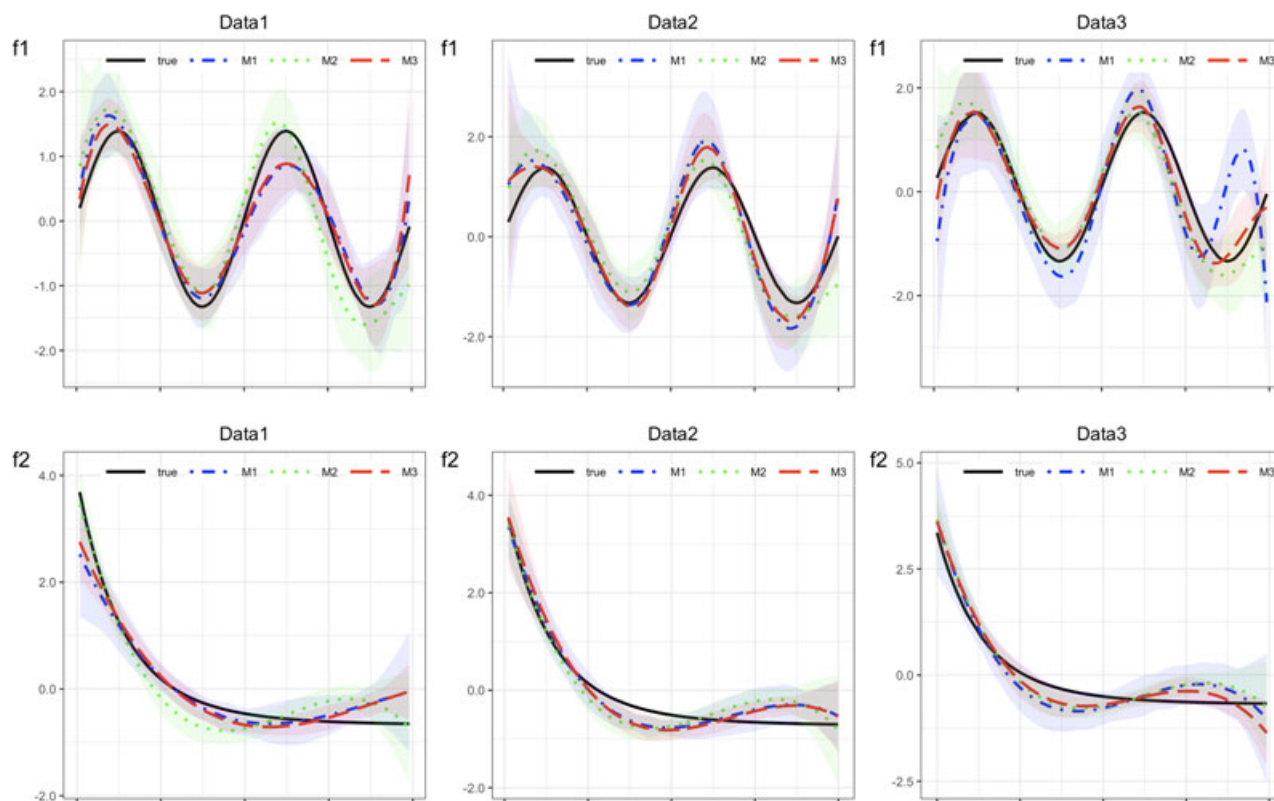


FIGURE A2 The estimated results for nonparametric functions under S3 with $p = 50$ [Colour figure can be viewed at wileyonlinelibrary.com]

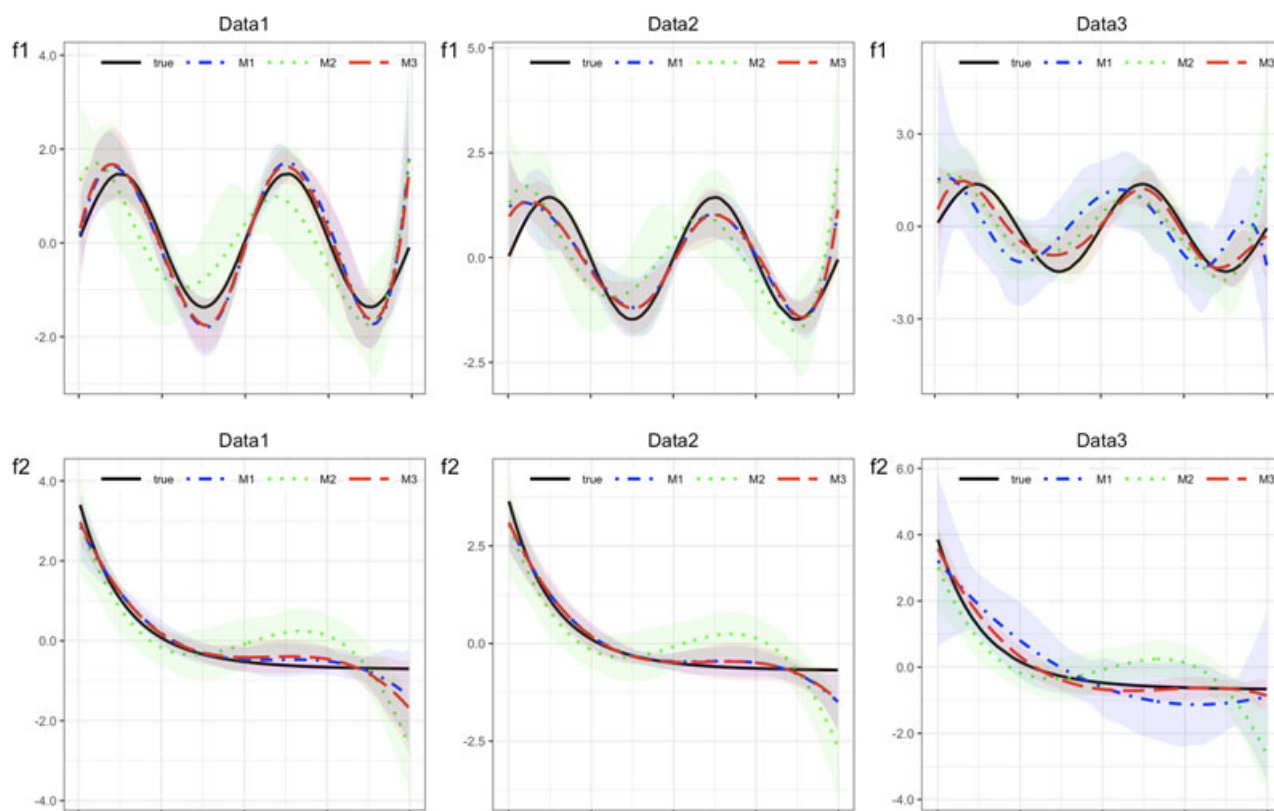


FIGURE A3 The estimated results for nonparametric functions under S4 with $p = 50$ [Colour figure can be viewed at wileyonlinelibrary.com]

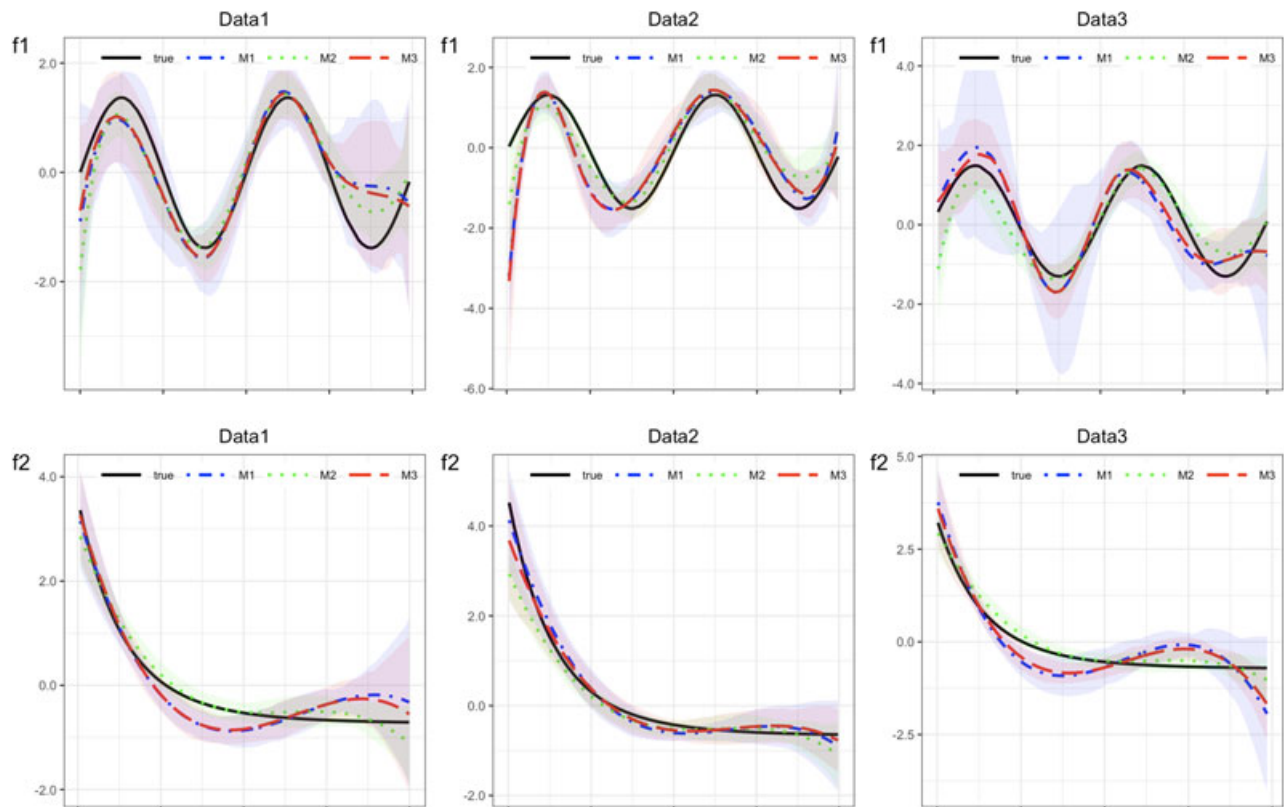


FIGURE A4 The estimated results for nonparametric functions under S5 with $p = 50$ [Colour figure can be viewed at wileyonlinelibrary.com]

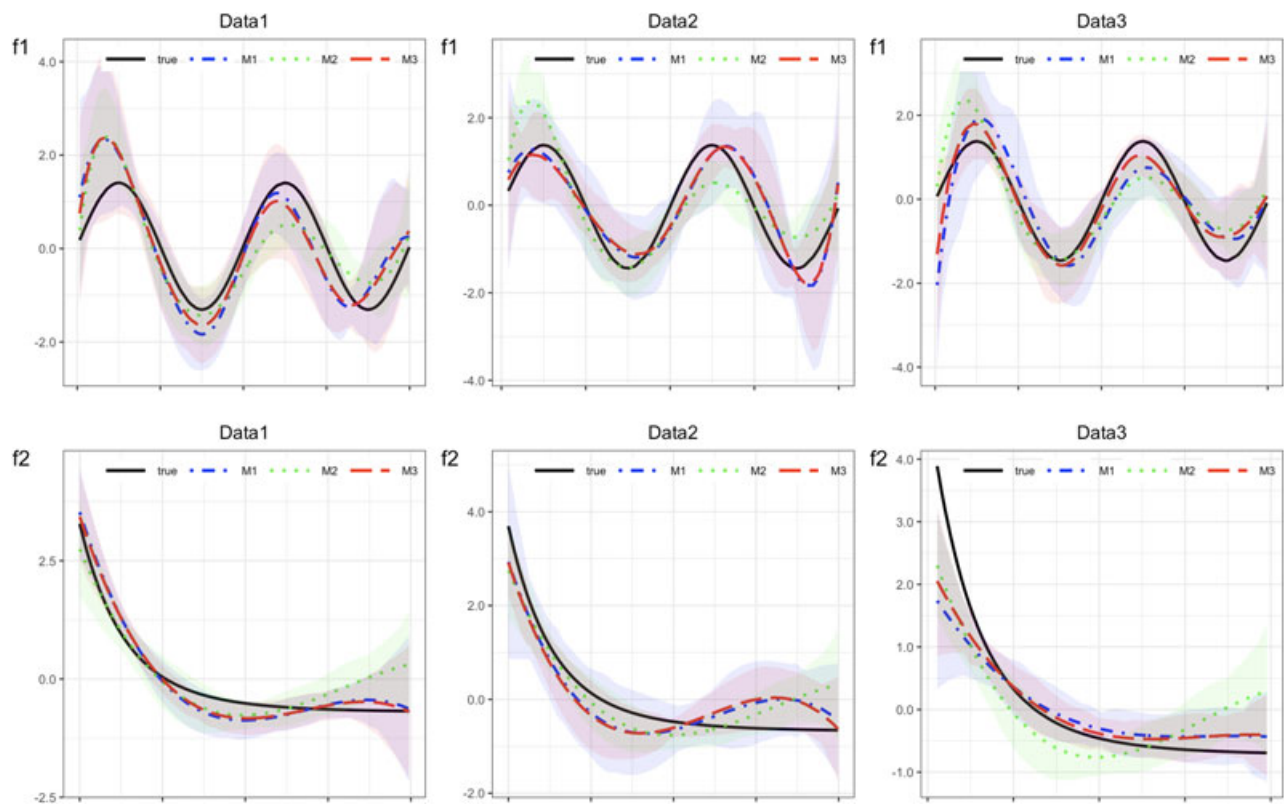


FIGURE A5 The estimated results for nonparametric functions under S6 with $p = 50$ [Colour figure can be viewed at wileyonlinelibrary.com]

APPENDIX B

SIMULATION RESULTS BASED ON THE SKCM DATA

In the simulation study of Section 3, we consider the AR1 correlation structure for convenience. Here, we conduct an additional simulation based on the SKCM data analyzed in Section 4.1. Specifically, we use the screened top 100 genes as main effects, and the true values of nonzero parameters are set in exactly the same way as S7. The results are presented in Table B1, and we can draw similar conclusion as those in Section 3. Although M3 is not the best in selecting the main terms, it outperforms the other two methods in terms of estimation and prediction accuracy.

TABLE B1 Simulation results based on the skin cutaneous melanoma data

| Method | FP | | FN | | MSE(SE) | | PE(SE) | RMSE(SE) |
|--------|-------|-------|------|-------|-------------|--------------|-------------|------------|
| | Main | Inter | Main | Inter | Main | Inter | | |
| M1 | 28.25 | 23.50 | 0.10 | 6.55 | 13.33(3.28) | 20.23(10.48) | 12.72(4.04) | 3.59(0.28) |
| M2 | 9.80 | 18.20 | 1.85 | 5.10 | 8.64(1.46) | 27.23(5.48) | 26.65(5.71) | 3.47(0.25) |
| M3 | 10.50 | 17.15 | 0.55 | 4.75 | 6.93(1.34) | 17.31(6.41) | 8.12(2.08) | 2.79(0.27) |

Abbreviations: FP: number of false positive variables; FN: number of false negative variables; MSE: mean square error for parametric part; PE: prediction error; RMSE: root of MSE for nonparametric part.