Biometrical Journal **XX** (2022) XX, zzz–zzz / DOI: 10.1002/bimj.200100000

Penalized estimation of a class of single-index varying-coefficient models for integrative genomic analysis

Hoi Min Ng [1], **Binyan Jiang** [1], and **Kin Yau Wong** [1] *

[1] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong

Received zzz, revised zzz, accepted zzz

Recent technological advances have made it possible to collect high-dimensional genomic data along with clinical data on a large number of subjects. In the studies of chronic diseases such as cancer, it is of great interest to integrate clinical and genomic data to build a comprehensive understanding of the disease mechanisms. Despite extensive studies on integrative analysis, it remains an ongoing challenge to model the interaction effects between clinical and genomic variables, due to high-dimensionality of the data and heterogeneity across data types. In this paper, we propose an integrative approach that models interaction effects using a single-index varying-coefficient model, where the effects of genomic features can be modified by clinical variables. We propose a penalized approach for separate selection of main and interaction effects. Notably, the proposed methods can be applied to right-censored survival outcomes based on a Cox proportional hazards model. We demonstrate the advantages of the proposed methods through extensive simulation studies and provide applications to a motivating cancer genomic study.

Key words: Adaptive lasso; Group penalty; Interaction; Semiparametric models; Splines

1 Introduction

The major goals of cancer genomics include identification of risk factors associated with disease outcomes, such as time to tumor progression or death since initial diagnosis or treatment. In conventional cancer studies, clinical factors such as age, gender, and tumor stage are routinely studied and used as prognostic factors. Recent advances in high-throughput technologies facilitate the generation of high-dimensional genomic data, which provide useful insights into the molecular pathways underlying cancer development. For example, in The Cancer Genome Atlas (TCGA), clinical and omics data, including copy number alteration, DNA methylation, mutation, and the expressions of mRNA, microRNA, and protein, were collected from more than 11000 cancer patients across 33 tumor types. Also, in the Molecular Taxonomy of Breast Cancer International Consortium (Curtis *et al.*, 2012), copy number alteration, mutation, and mRNA expression data were collected from about 2000 breast cancer patients. Such massive omics data enable researchers to gain deeper understanding of the biological mechanisms involved in cancer progression. Many studies have shown that the integrative analysis of clinical and genomic data confers greater prognostic power than the analysis of clinical data alone (Li, 2006; Shedden *et al.*, 2008; Bøvelstad *et al.*, 2009; Fan *et al.*, 2011; Zhao *et al.*, 2015).

Methods for integrative analysis of clinical and genomic data have been extensively investigated in recent decades. A straightforward integration strategy is to combine clinical and genomic data into a single data set, on which conventional analyses are performed. Some studies demonstrated that direct combination of clinical factors and gene expressions improves outcome prediction over the use of either data type alone (Bøvelstad *et al.*, 2009; Fan *et al.*, 2011; Zhao *et al.*, 2015). Alternatively, one may take into account the difference in prognostic power of the data types through some weighting approach. Gevaert

*Corresponding author: e-mail: kin-yau.wong@polyu.edu.hk

et al. (2006) developed a Bayesian network approach that builds separate models for clinical and microarray data and used a weighted approach to combine the model predictions. Daemen *et al.* (2007) proposed a weighted kernel-based method to integrate clinical and microarray data for classification. Both studies demonstrated that models that account for the distinction between clinical and genomic data tend to yield better prediction accuracy over models that treat these data types equally.

Integrative methods for multiple genomic data types have also been studied. Lanckriet *et al.* (2004), Daemen *et al.* (2009), and Seoane *et al.* (2014) proposed weighted kernel-based approaches to integrate multiple heterogeneous data types. Boulesteix *et al.* (2017) and Wong *et al.* (2019) proposed penalization regression methods on multiple data types while accounting for their differences in prognostic power. Wang *et al.* (2013) and Zhu *et al.* (2016) incorporated prior knowledge of the regulatory relationship among different types of genomic variables for the regression of disease outcomes on genomic variables. These methods, though accounting for differences among different data types, do not allow for interaction effects. Nevins *et al.* (2003) and Pittman *et al.* (2004) developed tree-based classification methods to evaluate the effects of clinical and genomic data on (binary) disease outcomes, allowing for potential interactions among multiple risk factors. However, the estimated model does not have simple interpretations, and the methods may not accommodate a large number of variables. In a recent study, Kerin and Marchini (2020) proposed a linear environment mixed model with interaction effects between linear combination of environmental variables and genetic markers. Despite that this model characterizes the interaction effects between clinical variables and genotypes, it may not accommodate a large number of variables. Li *et al.* (2020) proposed a regularization method to select for gene-gene interaction effects on disease outcomes, but the interactions between clinical and genomic data were not considered.

The effects of genomic features on cancer progression are often modified by clinical factors. For example, Landi *et al.* (2008) demonstrated that the effects of some gene expressions on the risk of lung cancer mortality vary with tobacco consumption. Also, Chen *et al.* (2017) and Relli *et al.* (2018) showed that the molecular mechanisms of carcinogenesis exhibit a high level of heterogeneity between two subtypes of non-small-cell lung carcinoma (NSCLC), and the same set of features can have distinct effects on disease outcome across different subtypes. As the effects of genomic features can vary across different clinical characteristics, it is highly desirable to incorporate interaction effects between clinical and genomic variables in regression analyses of disease outcomes on clinical and genomic variables.

A conventional approach to model interaction effects is to include pairwise product terms of predictors into the regression model. However, this approach may not be ideal for analyzing the interactions between clinical and genomic data. First, inclusion of product terms may greatly expand the model complexity and aggravate the high-dimensionality issue. Second, the scales of (quantitative) clinical and genomic variables are generally incomparable, and modeling interaction effects using pairwise product terms may not be appropriate. A possible approach to accommodate a flexible relationship between heterogeneous features is to fit a single-index varying-coefficient model. The single-index varying-coefficient model is a combination of the varying-coefficient model (Hastie and Tibshirani, 1993) and the single-index model (Härdle *et al.*, 1993). It allows the effect of each predictor to vary flexibly with a single index, which is a linear combination of another set of predictors, known as effect modifiers. In genomic analyses, the single-index varying-coefficient model can be used to describe the modifications of the effects of genomic features by clinical factors and introduce interaction effects between each genomic feature and a set of clinical factors. It avoids the curse of dimensionality by projecting the clinical features to an index, so that the number of parameters only increases linearly with the number of genomic features. To accommodate the difference in scales between clinical and genomic features, effects of genomic features are formulated as non-parametric functions of the index.

In the literature, Xue and Pang (2013) and Zhao *et al.* (2019) considered single-index varying-coefficient models for continuous outcomes and proposed kernel-based methods to estimate the index parameters and varying covariate effects. Lin *et al.* (2016) studied a proportional hazards model with single-index varying-coefficient components for censored outcomes. These studies assumed that the covariate effects are generally non-constant, which may not hold in many applications. Penalized estimation methods have been

developed for structure identification in single-index varying-coefficient models. Feng and Xue (2013) proposed a penalized estimation method to select and estimate important index parameters and coefficient functions based on spline approximation and the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001). As an extension, Feng and Xue (2015) imposed an additional penalty to the derivatives of the coefficient functions so as to identify the zero, constant, and varying effects. With the same purpose to distinguish constant and varying covariate effects, Guan (2017) proposed an alternative penalization method using the minimax concave penalty (MCP) (Zhang, 2007) and extended the penalization framework to a class of generalized linear models in a low-dimensional setting. All existing penalization methods are developed for continuous or binary outcomes, and models for censored event time have not been considered. It is unclear whether existing methods can be extended to accommodate right-censored outcomes, especially under a semiparametric outcome model with an infinite-dimensional nuisance parameter. In this paper, we propose a penalized (sieve) maximum likelihood estimation method for variable selection and estimation for a class of single-index varying-coefficient models, which accommodates continuous and censored outcomes. We adopt a novel two-part penalty that allows for separate selection of genomic features with effects modified by clinical features and of genomic features with non-zero constant effects. The proposed penalty functions are weighted to unify the degree of shrinkage of the constant and varying effects of a predictor. A coordinate-wise algorithm for the computation of the penalized estimators is developed. Unlike existing methods, our method accommodates right-censored survival outcomes, which are common in cancer genomic studies. Also, the proposed method is based on convex penalties, which tends to be more computationally stable compared with the existing methods based on the non-convex penalties.

The rest of this paper is organized as follows. We describe the model and estimation procedures in Section 2. We assess the estimation performance of the proposed methods through simulation studies, and the results are summarized in Section 3. We demonstrate the applications of the proposed methods on two TCGA data sets in Section 4. Finally, we make some concluding remarks in Section 5. Some technical details are given in the Appendix.

2 Model and estimation

2.1 Model, data, and sieve likelihood

Let Y be an outcome of interest, $\boldsymbol{U} \equiv (U_1, \ldots, U_q)$ and $\boldsymbol{Z} \equiv (Z_1, \ldots, Z_r)$ be two sets of predictors that may overlap, and $\boldsymbol{X} \equiv (X_0, \ldots, X_p)^{\mathrm{T}}$ be a set of predictors with $X_0 = 1$. We are interested in the effect of $(\boldsymbol{X}, \boldsymbol{Z})$ on Y, where the effect of $\boldsymbol{X}$ is allowed to depend on $\boldsymbol{U}$. In genomic studies, Y can be a disease outcome such as time to death, $\boldsymbol{X}$ can be a set of gene expressions, and $\boldsymbol{U}$ and $\boldsymbol{Z}$ can be clinical factors. We assume the following partial linear single-index varying-coefficient model:

$$Y \mid (\boldsymbol{U}, \boldsymbol{X}, \boldsymbol{Z}) \sim f\left\{ \cdot \, ; \sum_{j=0}^{p} g_j(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\beta}) X_j + \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\psi} \right\},$$

where f is a density function, $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$ are regression parameters, and $g_0, \ldots, g_p$ are unspecified smooth functions. The function g_0 characterizes the (possibly non-linear) baseline effect of the index on the outcome. For model identifiability, we set $\|\boldsymbol{\beta}\| = 1$, and if $\boldsymbol{U}$ is a subset of $\boldsymbol{Z}$, then we set the component of $\boldsymbol{\psi}$ that corresponds to the last component of $\boldsymbol{U}$ to be 0. This model assumes that the effect of each component of $\boldsymbol{X}$ is characterized by a non-parametric transformation of an index $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\beta}$. If each g_j ($j = 0, \ldots, p$) is constant, then the model contains only linear effects of $(\boldsymbol{X}, \boldsymbol{Z})$. If g_j is a linear function, then the model contains the linear effect of X_j and the interaction effect of $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\beta}$ and X_j. The proposed model accommodates many different types of outcomes. For continuous or binary outcomes, we set f to be a density from the exponential family. For survival outcomes, we set f to be the density under the Cox proportional hazards model. Under the Cox model, the conditional hazard function of survival time T given $(\boldsymbol{U}, \boldsymbol{X}, \boldsymbol{Z})$

takes the form of $h(t)\exp\{\sum_{j=0}^{p} g_j(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\beta})X_j + \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\psi}\}$ for $t \geq 0$, where $h(\cdot)$ is an unspecified baseline hazard function.

Suppose there are n observations. For uncensored outcomes, the observed data consist of $(Y_i, \boldsymbol{U}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ for $i = 1, \ldots, n$. The log-likelihood function is

$$\ell_n(\boldsymbol{\beta}, \boldsymbol{\psi}, \mathcal{G}) = \sum_{i=1}^{n} \log f\left\{Y_i; \sum_{j=0}^{p} g_j(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{\beta})X_{ij} + \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\psi}\right\},$$

where $\mathcal{G} = (g_0, \ldots, g_p)$. For a potentially right-censored outcome, let C_i be the censoring time for the ith subject, $\widetilde{Y}_i = \min(Y_i, C_i)$, and $\Delta_i = I(Y_i \leq C_i)$. The observed data consist of $(\widetilde{Y}_i, \Delta_i, \boldsymbol{U}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ for $i = 1, \ldots, n$. We set ℓ_n to be the log-partial-likelihood function, such that

$$\ell_n(\boldsymbol{\beta}, \boldsymbol{\psi}, \mathcal{G}) = \sum_{i=1}^{n} \Delta_i \left(\sum_{j=0}^{p} g_j(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{\beta})X_{ij} + \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\psi} - \log\left[\sum_{h:\widetilde{Y}_h \geq \widetilde{Y}_i} \exp\left\{ \sum_{j=0}^{p} g_j(\boldsymbol{U}_h^{\mathrm{T}}\boldsymbol{\beta})X_{hj} + \boldsymbol{Z}_h^{\mathrm{T}}\boldsymbol{\psi} \right\} \right] \right).$$

Because the likelihood involves the non-parametric functions $(g_0, \ldots, g_p)$, maximum likelihood estimation is not feasible. We propose to approximate g_j by B-spline functions. Let $(B_1, \ldots, B_d)$ be a set of B-spline functions on a pre-specified set of grid points, such that each function passes through the origin; the construction of the B-spline functions are discussed in the Appendix. For $j = 0, \ldots, p$, we approximate g_j by $\gamma_j + \sum_{k=1}^{d} \alpha_{jk} B_k$, where $(\gamma_j, \alpha_{j1}, \ldots, \alpha_{jd})$ are regression parameters. For right-censored outcomes, we set $\gamma_0 = 0$ for identifiability.

2.2 Penalized sieve likelihood

When p is large, the total number of parameters may be larger than the sample size, and penalization on $\boldsymbol{\gamma} \equiv (\gamma_0, \ldots, \gamma_p)^{\mathrm{T}}$ and $\boldsymbol{\alpha} \equiv (\boldsymbol{\alpha}_{jk})_{j=0,\ldots,p;k=1,\ldots,d}$ could be adopted for stable estimation and variable selection. We propose to estimate the parameters by maximizing the following penalized log-likelihood function:

$$p\ell_n(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) = \ell_n\left\{\boldsymbol{\beta}, \boldsymbol{\psi}, \left(\gamma_j + \sum_{k=1}^{d} \alpha_{jk} B_k\right)_{j=0,\ldots,p}\right\} - \sum_{j=1}^{p} \rho_1(\gamma_j; \lambda_1) - \sum_{j=1}^{p} \rho_2(\boldsymbol{\alpha}_j; \lambda_2),$$

where ρ_1 and ρ_2 are penalty functions, λ_1 and λ_2 are tuning parameters, and $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jd})^{\mathrm{T}}$ for $j = 1, \ldots, p$. This formulation allows separate selection of constant and non-constant effects of X_j by separate penalization on γ_j and $\boldsymbol{\alpha}_j$. Let $\widehat{\boldsymbol{\beta}}$, $\widehat{\gamma}_j$, and $\widehat{\boldsymbol{\alpha}}_j$ denote the penalized estimator of $\boldsymbol{\beta}$, γ_j, and $\boldsymbol{\alpha}_j$, respectively ($j = 0, \ldots, p$). For $j = 1, \ldots, p$, if $\widehat{\gamma}_j = 0$ and $\widehat{\boldsymbol{\alpha}}_j = \mathbf{0}$, then X_j does not have an effect on the outcome in the estimated model. If only $\widehat{\boldsymbol{\alpha}}_j = \mathbf{0}$, then X_j has a constant effect of $\widehat{\gamma}_j$. If $\widehat{\boldsymbol{\alpha}}_j$ is non-zero, then X_j has a non-constant effect indexed by $\boldsymbol{U}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$.

Many choices of penalty functions, such as the (group) lasso (Tibshirani, 1996; Yuan and Lin, 2006), SCAD (Fan and Li, 2001; Breheny and Huang, 2009), and MCP (Zhang, 2007; Breheny and Huang, 2009) are possible. In this paper, we propose to set $\rho_1(\gamma_j; \lambda_1) = \lambda_1 w_j |\gamma_j|$ and $\rho_2(\boldsymbol{\alpha}_j; \lambda_2) = \lambda_2 w_j (\boldsymbol{\alpha}_j^{\mathrm{T}} \boldsymbol{K}_j \boldsymbol{\alpha}_j)^{1/2}$, where w_j is a weight for the jth predictor, and $\boldsymbol{K}_j$ is some symmetric $(d \times d)$-positive definite matrix; the first penalty is similar to the adaptive lasso penalty (Zou, 2006), and the second penalty is a weighted version of the group lasso.

There are two advantages of this adaptive-lasso type penalty over other commonly used penalty functions. First, the adaptive-lasso type penalty can reduce the estimation bias while retaining the convexity of the regularization term. Although lasso is computationally efficient, the shrinkage introduced by lasso may result in substantial biases for large regression coefficients. Non-convex penalties, such as SCAD and MCP, are designed to diminish this bias while achieving selection consistency. However, this class of

penalty functions tends to possess multiple local optima. To preserve stability, we suggest using a weighted convex penalty function. Second, the weighted penalty utilizes the information that the constant and non-constant effects correspond to the same predictor. Similar adaptive group lasso penalties have been adopted by Wei and Huang (2010) and Wei *et al.* (2011) through setting the weight for each group to be inversely proportional to the norm of the corresponding group lasso estimates. Although conventional choices of penalty functions for ρ_1 and ρ_2 with separate weights can produce sparse estimation of the constant and non-constant effects, they fail to take into account the fact that γ_j and $\boldsymbol{\alpha}_j$ $(j = 1, \ldots, p)$ correspond to the same predictor X_j. The weight w_j is introduced to capture the overall signal strength of g_j and unify the degree of shrinkage of γ_j and $\boldsymbol{\alpha}_j$. In particular, we set $w_j = (\widetilde{\gamma}_j^2 + \|\widetilde{\boldsymbol{\alpha}}_j\|^2)^{-1/2}$, where $\widetilde{\gamma}_j$ and $\widetilde{\boldsymbol{\alpha}}_j$ are estimates of γ_j and $\boldsymbol{\alpha}_j$ obtained from maximizing the penalized log-likelihood with $w_j = 1$ for $j = 1, \ldots, p$. If the initial estimates $\widetilde{\gamma}_j$ and $\widetilde{\boldsymbol{\alpha}}_j$ are accurate in that variables with stronger signal receive smaller weights, then the weighted estimators would yield better variable selection and estimation accuracy than unweighted estimators.

2.3 Estimation

We propose to compute the estimates of $(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\gamma}, \boldsymbol{\alpha})$ using an alternating algorithm. In particular, we initialize $\boldsymbol{\beta}$ as some unit vector such that $\|\boldsymbol{\beta}\| = 1$ and update the parameter estimates of $(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\psi})$ and $\boldsymbol{\beta}$ alternatively until convergence. For any fixed (λ_1, λ_2) and w_j $(j = 1, \ldots, p)$, the algorithm is as follows:

Step 1: Initialize $\widehat{\boldsymbol{\beta}}^{(0)}$ as some unit vector such that $\big\|\widehat{\boldsymbol{\beta}}^{(0)}\big\| = 1$. Set $m = 1$.

Step 2: Update $(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\psi})$ by

$$\big(\widehat{\boldsymbol{\gamma}}^{(m)}, \widehat{\boldsymbol{\alpha}}^{(m)}, \widehat{\boldsymbol{\psi}}^{(m)}\big) \equiv \underset{(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\psi})}{\arg\max}\, p\ell_n\big(\widehat{\boldsymbol{\beta}}^{(m-1)}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\psi}\big).$$

For fixed $\boldsymbol{\beta}$, the objective function $p\ell_n(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\gamma}, \boldsymbol{\alpha})$ is essentially a penalized log-likelihood function for a conventional regression model under a group lasso penalty, and $(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\psi})$ can be updated using existing algorithms for the group lasso (Breheny and Huang, 2009).

Step 3: Update $\boldsymbol{\beta}$ by

$$\widehat{\boldsymbol{\beta}}^{(m)} \equiv \underset{\|\boldsymbol{\beta}\|=1}{\arg\max}\, \ell_n\left\{\boldsymbol{\beta}, \widehat{\boldsymbol{\psi}}^{(m)}, \left(\widehat{\gamma}_j^{(m)} + \sum_{k=1}^{d} \widehat{\alpha}_{jk}^{(m)} B_k\right)_{j=0,\ldots,p}\right\}.$$

For fixed $(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\psi})$, $\boldsymbol{\beta}$ can be updated using the Lagrange multiplier method, and $\widehat{\beta}$ is defined as the solution of the equations:

$$\begin{cases} \frac{\partial}{\partial \boldsymbol{\beta}} \ell_n\big\{\boldsymbol{\beta}, \boldsymbol{\psi}, \big(\gamma_j + \sum_{k=1}^{d} \alpha_{jk} B_k\big)_{j=0,\ldots,p}\big\} + c\boldsymbol{\beta} = \mathbf{0} \\ \|\boldsymbol{\beta}\|^2 - 1 = 0, \end{cases}$$

where c is the Lagrange multiplier. We can solve for $\boldsymbol{\beta}$ and c simultaneously using the Newton-Raphson algorithm.

Step 4: Set $m = m + 1$. Repeat Steps 2–4 until convergence.

The performance of the proposed methods depends critically on the choice of tuning parameters. We propose to select the tuning parameters λ_1 and λ_2 using a version of the Bayesian information criterion (BIC), defined as

$$-2\ell_n(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\psi}}, \widehat{\mathcal{G}}) + (df + q + r)\log(n^*),$$

where $\widehat{\mathcal{G}} = (\widehat{\gamma}_j + \sum_{k=1}^{d} \widehat{\alpha}_{jk} B_k)_{j=0,\ldots,p}$, *df* is the effective degrees of freedom, and n^* is the effective sample size. Specifically, $n^* = n$ for uncensored outcomes, and n^* is the number of uncensored observations for right-censored outcomes. Following Breheny and Huang (2009), we define the effective degrees of freedom as

$$df = \sum_{j=1}^{p} \left(\frac{\widehat{\gamma}_j}{\widehat{\gamma}_j^*} + \sum_{k=1}^{d} \frac{\widehat{\alpha}_{jk}}{\widehat{\alpha}_{jk}^*} \right),$$

where $(\widehat{\gamma}_j, \widehat{\alpha}_{jk})$ denote the estimated value of (γ_j, α_{jk}), $\widehat{\gamma}_j^*$ denote the maximizer of the unpenalized log-likelihood function with respect to γ_j with other parameters fixed at the estimated value with penalty, and $\widehat{\alpha}_{jk}^*$ denote the maximizer of the unpenalized log-likelihood function with respect to α_{jk} with other parameters fixed at the estimated value with penalty. (In the coordinate descent algorithm, the penalized estimates are computed by applying a soft-thresholding operator on the unpenalized estimates, so γ_j^* and α_{jk}^* are obtained as by-products of the algorithm with no additional computation.)

In practice, we select the tuning parameters over a two-dimensional grid of λ_1 and λ_2. To construct an $(m_1 \times m_2)$ grid, for $\boldsymbol{\alpha} = \mathbf{0}$ and $\boldsymbol{\psi}$ at their maximum likelihood estimates, we find the smallest value of λ_1 for which $\boldsymbol{\gamma} = \mathbf{0}$. Then, starting from this smallest value, we construct a sequence of m_1 logarithmically spaced values for λ_1. Next, for each value of λ_1, we compute the smallest value of λ_2 such that $\boldsymbol{\alpha} = \mathbf{0}$ with $(\boldsymbol{\gamma}, \boldsymbol{\psi})$ fixed at the lasso estimate under the λ_1 value. We can then construct a sequence of m_2 logarithmically spaced values for λ_2 for each λ_1 value. Note that at each set of (λ_1, λ_2) values, the algorithm presented in Steps 1–4 is performed, and a modified BIC value is obtained. The set of (λ_1, λ_2) values that yields the smallest modified BIC is selected.

We adopt the BIC, instead of other popular criteria such as cross-validation and AIC, due to its computational efficiency and model selection performance. Performing cross-validation over a two-dimensional grid, especially with multiple initial values for $\boldsymbol{\beta}$, is highly computationally expensive. On the other hand, AIC tends to select many more variables due to its relatively weak penalty on model complexity. Based on the results not presented, AIC selects many more noise variables and also yields higher estimation error than BIC under the simulation settings considered in Section 3.

In conventional group lasso problems, the predictor matrix of the jth group, denoted by $\boldsymbol{V}_j$, is typically transformed such that $\boldsymbol{V}_j^{\mathrm{T}} \boldsymbol{V}_j$ is a diagonal matrix with equal diagonal elements. This is equivalent to setting $\boldsymbol{K}_j$ to be (a scaled version of) $\boldsymbol{V}_j^{\mathrm{T}} \boldsymbol{V}_j$. In the current problem, however, the "predictor matrix," which consists of rows $(X_{ij}, B_1(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{\beta})X_{ij}, \ldots, B_d(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{\beta})X_{ij})$ $(i = 1, \ldots, n)$, depends on the unknown parameter $\boldsymbol{\beta}$. One estimation strategy is to set $\boldsymbol{K}_j$ based on the predictor matrix evaluated at some initial estimator of $\boldsymbol{\beta}$, such as that obtained under $\boldsymbol{K}_j = \boldsymbol{I}$. Another strategy is to update $\boldsymbol{K}_j$ with $\boldsymbol{\beta}$ after each iteration; this can be thought of as setting $\boldsymbol{K}_j$ based on the converged value of $\boldsymbol{\beta}$. Another difficulty that arises from the unknown $\boldsymbol{\beta}$ is that the converged estimates may vary with the initial value of $\boldsymbol{\beta}$. We propose to consider multiple initial values and select the final estimates that yield the smallest modified BIC. In simulation studies, we considered 5 initial values of $\boldsymbol{\beta}$ and updated $\boldsymbol{K}$ along with $\boldsymbol{\beta}$ at each iteration, and the algorithm converged at almost all replicates. For the continuous outcome, with either choice of w_j, the algorithm converged for at least one set of initial values in all replicates. For the right-censored outcome, with $w_j = 1$, we recorded 0, 0, 1, and 2 non-convergence cases under $p = 20$, 50, 100, and 300, respectively; with w_j being the adaptive weight, the algorithm converged for at least one set of initial values in all replicates.

2.4 Partial dependence functions

The single-index varying-coefficient model characterizes the effect of X_j $(j = 1, \ldots, p)$ on the outcome through the coefficient function $g_j(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\beta})$, which takes the linear combination of $\boldsymbol{U}$ as input. One way to represent the effect of X_j is to plot g_j over different index values. Also, one may be interested in the influence of a particular effect modifier U_k on the effect of X_j, which depends on other effect modifiers. To

characterize the overall influence of a single effect modifier U_k on g_j, we introduce the partial dependence function. Let U_{ik} be the index variable of the i-th observation, $\boldsymbol{U}_{i\backslash k}$ be the subvector of $\boldsymbol{U}_i$ with the k-th element removed, and β_k and $\boldsymbol{\beta}_{\backslash k}$ be the parameters corresponding to U_{ik} and $\boldsymbol{U}_{i\backslash k}$, respectively. We define the partial dependence function for the influence of U_k on the effect of X_j (evaluated at $U_k = u$) as

$$\overline{g}_{jk}(u) = \frac{\sum_{i=1}^{n} K_h(U_{ik} - u)\widehat{g}_j(U_{ik}\beta_k + \boldsymbol{U}_{i\backslash k}^{\mathrm{T}}\boldsymbol{\beta}_{\backslash k})}{\sum_{i=1}^{n} K_h(U_{ik} - u)},$$

for $j = 1, \ldots, p$ and $k = 1, \ldots, q$, where $K_h(x)$ is a kernel function with bandwidth h. The function $\overline{g}_{jk}(u)$ is a kernel-based estimate of $\mathrm{E}\{\widehat{g}(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\beta}) \mid U_k = u\}$, where the expectation is taken with respect to components of $\boldsymbol{U}$ besides U_k. We can plot the partial dependence function to visualize the overall influence of U_k on the effect of X_j.

3 Simulation studies

We set the dimension of $\boldsymbol{U}$ to be 4 and generated components of $\boldsymbol{U}$ as i.i.d. standard normal variables. We set $\boldsymbol{Z} = \boldsymbol{U}$ and generated $\boldsymbol{X}$ from the p-variate standard normal distribution. We set $\boldsymbol{\beta} = (0.4, -0.4, 0.2, -0.8)^{\mathrm{T}}$ and $\boldsymbol{\psi} = (0.2, -0.2, 0.5, -0.5)^{\mathrm{T}}$. We set $g_1, \ldots, g_{20}$ to be non-zero coefficient functions, where $g_1, \ldots, g_{10}$ are constant and $g_{11}, \ldots, g_{20}$ are varying; the functions are plotted in Figure 1. We set g_0 and $g_{21}, \ldots, g_p$ to be constant at 0. We considered a continuous outcome variable and a right-censored outcome variable. For the continuous outcome, we set $f(y; \mu) = (2\pi)^{-1/2}\exp\{-(y-\mu)^2/2\}$ so that conditional on $(\boldsymbol{U}, \boldsymbol{X}, \boldsymbol{Z})$, Y follows the normal distribution with unit variance. For the right-censored outcome, we set $f(y; \mu) = h(y)\exp(\mu)\exp\big\{-\exp(\mu)\int_0^y h(t)\,\mathrm{d}t\big\}$, where h is the baseline hazard function with $h(t) = t$. The censoring time was generated from an exponential distribution with the mean chosen to yield a censoring rate of about 30%. In each setting, we considered a sample size of 500 and $p = 20$, 50, 100, and 300.

We compare the proposed methods with conventional regression models with or without interaction terms. For the proposed methods, we simply set the degree of the B-spline functions to be 2 and the knots at $-\max_i \|\boldsymbol{U}_i\|_2$, 0, and $\max_i \|\boldsymbol{U}_i\|_2$. We considered the proposed weighted approach and an unweighted approach with $w_j = 1$ $(j = 1, \ldots, p)$. We also considered the lasso regression on the linear predictors $(\boldsymbol{X}, \boldsymbol{Z})$ and the lasso regression on $\boldsymbol{X}$, $\boldsymbol{Z}$, and pairwise interactions between components of $\boldsymbol{X}$ and $\boldsymbol{U}$; in both cases, coefficients of $\boldsymbol{Z}$ were not penalized. In the following discussion, we refer to the proposed method as "SIVC", the lasso regression with main effects as "MAIN", and the lasso regression with both main and interaction effects as "INT". In addition, we considered adaptive lasso for the models with or without interactions, where the weights are the inverse of the absolute value of the corresponding lasso estimates. The tuning parameters for all methods were selected using the modified BIC defined in Section 2.3. For the proposed methods, we set $m_1 = m_2 = 20$, so a (20×20) grid was considered.

We evaluate the performance of each method in terms of variable selection and prediction. For variable selection, we report the sensitivity and the false discovery rate (FDR). Sensitivity is the proportion of correctly identified signal variables among all true signal variables. FDR is the proportion of noise variables that are incorrectly identified as signal variables among all selected variables. For the SIVC, a variable X_j is selected if either γ_j or $\boldsymbol{\alpha}_j$ is estimated as non-zero $(j = 1, \ldots, p)$. For SIVC and INT, we also report the sensitivity and FDR with respect to the selection of non-constant effects, where for the SIVC, the non-constant effect of X_j is selected if $\widehat{\boldsymbol{\alpha}}_j \neq \boldsymbol{0}$, and for the INT, the non-constant effect is selected if the coefficient of the product of X_j and any component of $\boldsymbol{U}$ is non-zero. In addition, we report the total numbers of the selected variables and the number of variables identified to have non-constant effects.

For prediction, we report the mean-squared error (MSE), defined as $\mathrm{E}(\widehat{\eta} - \eta_0)^2$, where $\eta_0 = \eta(\boldsymbol{\beta}_0, \mathcal{G}_0, \boldsymbol{\psi}_0)$, $\eta(\boldsymbol{\beta}, \mathcal{G}, \boldsymbol{\psi}) \equiv \sum_{j=1}^{p} g_j(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\beta})X_j + \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\psi}$, and $(\boldsymbol{\beta}_0, \mathcal{G}_0, \boldsymbol{\psi}_0)$ denote the true parameter values. For the SIVC, $\widehat{\eta} = \eta(\widehat{\boldsymbol{\beta}}, \widehat{\mathcal{G}}, \widehat{\boldsymbol{\psi}})$, where $(\widehat{\boldsymbol{\beta}}, \widehat{\mathcal{G}}, \widehat{\boldsymbol{\psi}})$ denote the estimated parameter values. For MAIN and INT,

$\widehat{\eta} = \sum_j \widehat{b}_j X_j + \sum_k \widehat{c}_k Z_k + \sum_{j,l} \widehat{d}_{jl} X_j U_l$ and $\widehat{\eta} = \sum_j \widetilde{b}_j X_j + \sum_k \widetilde{c}_k Z_k$, respectively, where $\widehat{b}_j$, $\widehat{c}_k$, $\widehat{d}_{jl}$, $\widetilde{b}_j$, and $\widetilde{c}_k$ are the corresponding estimated regression parameters. For the right-censored outcome, we also compute the concordance index (C-index) (Harrell *et al.*, 1982), defined as $\mathrm{P}(\eta_i > \eta_j \mid \widetilde{Y}_i < \widetilde{Y}_j)$ for two generic independent subjects indexed by i and j. C-index typically takes values between 0.5 and 1, where a value of 0.5 indicates no discrimination and a value of 1 indicates perfect discrimination. We compute the MSE and C-index on a set of independently generated data set of size 5000. For the SIVC, we also report the absolute inner product $|\boldsymbol{\beta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}|$ to assess the estimation accuracy of $\widehat{\boldsymbol{\beta}}$. Because $\boldsymbol{\beta}$ is of unit L_2 norm, the inner product $\boldsymbol{\beta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$ is equal to the (scaled) negative squared error $-\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$ up to a constant. The simulation results for the continuous and right-censored outcomes based on 100 replicates are summarized in Tables 1 and 2, respectively. Figure 1 shows the average estimated values of $g_1, \ldots, g_{20}$ for the continuous outcome under $p = 300$. The simulation results under other settings are plotted in Figures S1–7 in the supplementary materials.

All analyses were run on an Intel Xeon Processor E7 2.8 GHz. In each replicate, we record the average computation time for the convergence cases along the solution path. For the continuous outcome, the average computation time of the proposed unweighted (weighted) methods based on 100 replicates are 17.90 (17.79), 23.16 (17.97), 37.74 (24.07), and 88.55 (42.61) seconds under $p = 20$, 50, 100, and 300, respectively. For the right-censored outcome, the average computation time of the proposed unweighted (weighted) methods under the right-censored outcome are 31.90 (21.51), 40.81 (36.01), 61.73 (36.76), and 96.12 (66.62) seconds under $p = 20$, 50, 100, and 300, respectively.

Table 1 Simulation results for the continuous outcome.

		$p = 20$			$p = 50$			$p = 100$			$p = 300$		
		SIVC	MAIN	INT	SIVC	MAIN	INT	SIVC	MAIN	INT	SIVC	MAIN	INT
Unweighted													
SEN	Overall	0.990	0.801	0.979	0.978	0.776	0.967	0.962	0.755	0.952	0.940	0.732	0.927
	Non-constant	0.999	-	0.949	0.986	-	0.917	0.983	-	0.900	0.942	-	0.861
FDR	Overall	0	0	0	0.376	0.295	0.504	0.543	0.475	0.707	0.740	0.728	0.871
	Non-constant	0.251	-	0.426	0.434	-	0.718	0.530	-	0.834	0.669	-	0.926
NS	Overall	19.80	16.01	19.57	31.55	22.21	39.13	42.45	29.11	65.17	73.09	54.33	145.00
	Non-constant	13.46	-	16.65	17.62	-	32.84	21.15	-	54.66	28.95	-	117.63
$\lvert\boldsymbol{\beta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}\rvert$		0.997	-	-	0.996	-	-	0.996	-	-	0.996	-	-
MSE		0.378	1.583	0.990	0.569	1.754	1.160	0.698	1.764	1.203	0.862	1.847	1.488
Weighted													
SEN	Overall	0.948	0.657	0.908	0.915	0.649	0.900	0.898	0.639	0.897	0.888	0.647	0.875
	Non-constant	0.978	-	0.848	0.951	-	0.826	0.914	-	0.834	0.907	-	0.794
FDR	Overall	0	0	0	0.155	0.096	0.324	0.248	0.200	0.540	0.399	0.504	0.786
	Non-constant	0.069	-	0.243	0.154	-	0.552	0.223	-	0.709	0.326	-	0.872
NS	Overall	18.95	13.14	18.15	21.77	14.47	26.90	24.16	16.19	39.43	30.08	26.55	83.15
	Non-constant	10.56	-	11.36	11.33	-	18.83	11.95	-	29.19	13.62	-	63.50
$\lvert\boldsymbol{\beta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}\rvert$		0.998	-	-	0.997	-	-	0.997	-	-	0.997	-	-
MSE		0.313	1.602	1.008	0.408	1.724	1.135	0.544	1.730	1.198	0.588	1.820	1.638

NOTE: "SEN" represents sensitivity; "NS" represents number of selected variables; "Overall" gives values of corresponding measures concerning all components of $\boldsymbol{X}$; "Non-constant" gives values of corresponding measures concerning components of $\boldsymbol{X}$ with non-constant effects on the outcome.

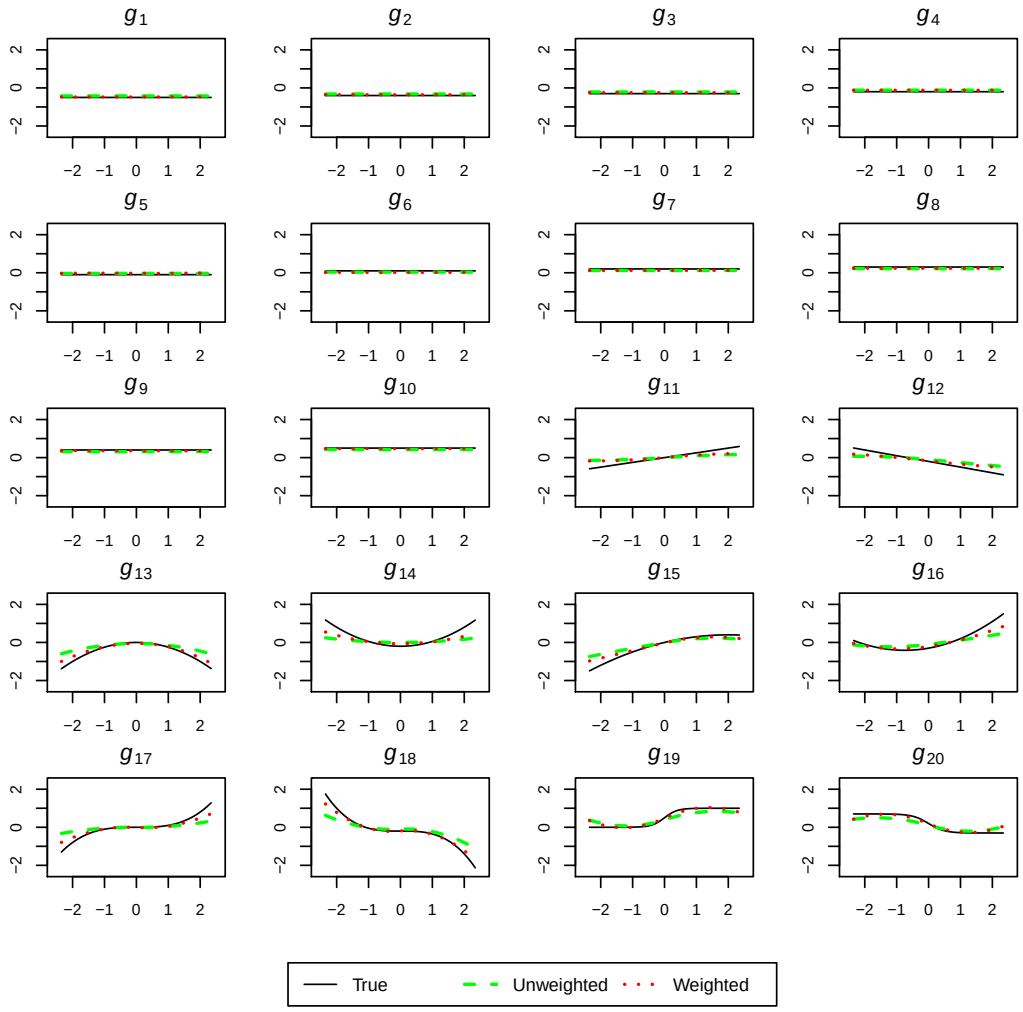

Figure 1 Estimated coefficients for the continuous outcome under $p = 300$.

In terms of prediction, both the weighted and unweighted versions of the SIVC correctly identify the interaction structure between $\boldsymbol{X}$ and $\boldsymbol{U}$ and yield higher prediction accuracy than other methods. In particular, they yield lower MSE in all settings and higher C-index for the right-censored outcome. In addition, the estimated value of $\boldsymbol{\beta}$ is close to the true value, indicating that the proposed methods can correctly identify the composition of the index. The weighted estimators are generally accurate, whereas the unweighted estimators tend to be biased towards zero due to the uniform shrinkage imposed on all parameters. The INT generally yields smaller MSE than MAIN, suggesting that a varying-coefficient model can be approximated by a conventional regression model with pairwise interaction terms. Nevertheless, possibly due to the complexity of the interaction model, the performance of the INT is substantially worse than that of SIVC.

Table 2 Simulation results for the right-censored outcome.

		$p = 20$			$p = 50$			$p = 100$			$p = 300$		
		SIVC	MAIN	INT	SIVC	MAIN	INT	SIVC	MAIN	INT	SIVC	MAIN	INT
Unweighted													
SEN	Overall	0.960	0.795	0.949	0.912	0.737	0.914	0.862	0.689	0.874	0.807	0.662	0.818
	Non-constant	0.909	-	0.895	0.807	-	0.852	0.723	-	0.794	0.626	-	0.711
FDR	Overall	0	0	0	0.315	0.297	0.446	0.482	0.480	0.629	0.656	0.692	0.801
	Non-constant	0.167	-	0.392	0.276	-	0.661	0.389	-	0.776	0.483	-	0.884
NS	Overall	19.20	15.90	18.97	26.91	21.11	33.28	33.58	26.78	48.24	47.75	43.51	83.04
	Non-constant	11.05	-	14.91	11.41	-	25.51	12.23	-	36.66	12.70	-	62.38
$\lvert\boldsymbol{\beta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}\rvert$		0.993	-	-	0.980	-	-	0.970	-	-	0.953	-	-
MSE		0.842	1.873	1.419	1.345	2.142	1.701	1.492	2.134	1.746	1.737	2.221	1.913
C-index		0.772	0.716	0.743	0.758	0.716	0.738	0.747	0.708	0.726	0.734	0.704	0.715
Weighted													
SEN	Overall	0.878	0.631	0.864	0.834	0.615	0.832	0.773	0.581	0.787	0.728	0.582	0.749
	Non-constant	0.807	-	0.780	0.710	-	0.732	0.611	-	0.687	0.530	-	0.609
FDR	Overall	0	0	0	0.138	0.139	0.303	0.246	0.273	0.489	0.422	0.519	0.686
	Non-constant	0.069	-	0.238	0.143	-	0.518	0.220	-	0.658	0.310	-	0.817
NS	Overall	17.56	12.62	17.27	19.49	14.39	24.13	20.81	16.19	31.22	26.07	24.66	49.01
	Non-constant	8.76	-	10.40	8.40	-	15.60	7.88	-	21.03	7.54	-	34.20
$\lvert\boldsymbol{\beta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}\rvert$		0.994	-	-	0.991	-	-	0.984	-	-	0.965	-	-
MSE		0.696	1.822	1.209	0.934	1.983	1.414	1.146	1.947	1.528	1.339	2.011	1.872
C-index		0.773	0.714	0.743	0.766	0.716	0.739	0.755	0.709	0.726	0.744	0.703	0.710

NOTE: See NOTE to Table 1.

In terms of variable selection, both the SIVC and INT have substantially higher sensitivity than lasso with main effects alone. The FDR is lower under the SIVC than INT, indicating that the SIVC tends to yield more interpretable models. The FDR for the proposed methods is higher than those for the MAIN under some settings, possibly because MAIN generally selects much fewer variables. For all methods, the weighted estimators yield substantially lower FDR than the unweighted estimators. By setting higher penalty for noise variables and lower penalty for signal variables, the weighted method yields higher variable selection accuracy.

We have conducted additional simulation studies to investigate the impact of model misspecification. First, there could be no interaction effect between the covariates in real applications. To evaluate the proposed methods when the extra model complexity is not necessary, we conducted an additional simulation study under a main effect model. We showed in Section S3.1 that the extra model complexity does not substantially worsen the estimation performance under a simple true model. Second, prior knowledge about the form of main effect may not be available in practice, and the underlying main effect may be non-linear. The major issue with misspecification of main effects of $\boldsymbol{U}$ is that interaction terms may explain the variability introduced by the non-linear main effects and tend to be selected, even if no such interaction effects are present. We conducted an additional simulation study to investigate the impact of misspecification of main effects. In the study presented in Section S3.2, the effect of misspecification of main effects on the variable selection performance appears to be minimal. In addition, we conducted an additional simulation study to evaluate the performance of each method under larger noise. We showed in Section S3.3 that larger noise results in lower sensitivity and less accurate prediction. Nevertheless, the proposed methods yield superior prediction performance to the lasso approaches.

4 Real data analysis

4.1 TCGA NSCLC data

We demonstrate the application of the proposed methods using a set of NSCLC patients from TCGA. The data set consists of two subtypes of lung cancer, namely lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). We are interested in the potential risk factors associated with pulmonary function, measured by the percentage of expiratory volume in one second (FEV1); a higher FEV1 represents larger lung capacity, and patients with severely impaired lung function have an increased risk of mortality (Hole *et al.*, 1996). In particular, we investigated the effects of gene expressions and clinical variables on FEV1, allowing for interactions between the two types of variables. We fit the proposed model with $\boldsymbol{U}$ consisting of age, number of packs of cigarettes smoked per year or pack-year smoked (PYS), cancer subtype ("0" for LUSC and "1" for LUAD), tumor stage ("0" for stage I and "1" for stage II or above), and gender ("0" for female and "1" for male). This formulation allows the effects of genomic factors to be modified by clinical variables. We set $\boldsymbol{Z} = \boldsymbol{U}$ to allow linear effects of clinical variables on FEV1. After discarding genes with zero expressions for 30% or more subjects, the data set consists of 17148 gene expressions. We set $\boldsymbol{X}$ to consist of 300 gene expressions that have the most significant marginal association with FEV1 (adjusted for clinical variables). After removing subjects with missing data, the sample size is 353, with 185 and 168 LUAD and LUSC patients, respectively. Following the simulation studies, we set the degree of the B-spline functions to be 2 and the knots at $-\max_i \|\boldsymbol{U}_i\|_2$, 0, and $\max_i \|\boldsymbol{U}_i\|_2$. We adopted the weighted penalty approach to fit the single-index varying-coefficient model. For comparison, we also perform the (weighted) lasso on the main effect model. In all methods, we standardized all variables to have zero mean and unit variance. The selected gene expressions and their estimated coefficients are shown in Table S1 in the supplementary materials.

The proposed method identified 17 gene expressions to be associated with FEV1, among which 13 of them are selected by lasso. Among the selected gene expressions, EIF4A3 was known to be involved in the development of NSCLC, and KCNK2 and N4BP1 were known as prognostic factors in some cancer types (Innamaa *et al.*, 2013; Xu *et al.*, 2017; Lin *et al.*, 2018; Li *et al.*, 2019). The effects of CDK11A and LRRC29 were identified to vary with the clinical variables; CDK11A has previously been shown to be associated with many cancer types (Zhou *et al.*, 2016). The estimated index parameters $\boldsymbol{\beta}$ for age, PYS, gender, tumor stage, and cancer subtype are 0.198, 0.636, 0.157, -0.549, and -0.479, respectively. The index is dominated by PYS, tumor stage, and cancer subtype, suggesting that the effects of CDK11A and LRRC29 mainly depend on these three clinical factors. Figure 2 displays the estimated values of g_0 and the g functions for CDK11A and LRRC29.

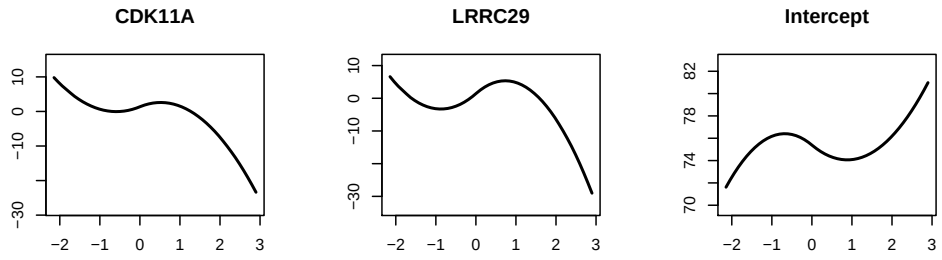

Figure 2 Estimated coefficients for NSCLC analysis.

To gain further insight into the results, we construct partial dependence plots for the influence of age and PYS on the effects of CDK11A. We adopt a Gaussian kernel function with $K_h(x) = K(x/h)/h$ and $K(x) = (2\pi)^{-1/2}\exp\{-x^2/2\}$, and the bandwidth h is selected using the rule of thumb bandwidth

estimator (Silverman, 1986). Figure 3 displays the partial dependence plots for the varying coefficient of CDK11A. We also plot the partial dependence functions for different tumor stage groups. Overall, the effect of CDK11A tends to be more negative as PYS or age increases. In particular, PYS tends to strengthen the negative effect of CDK11A on FEV1 except for values near the upper boundary, whereas age weakens the positive effect of CDK11A on FEV1. As shown in the partial dependence plots under separate tumor stage groups, the influences of age and PYS on the effect of CDK11A are stronger for patients with stage I than for patients with stage II or above.

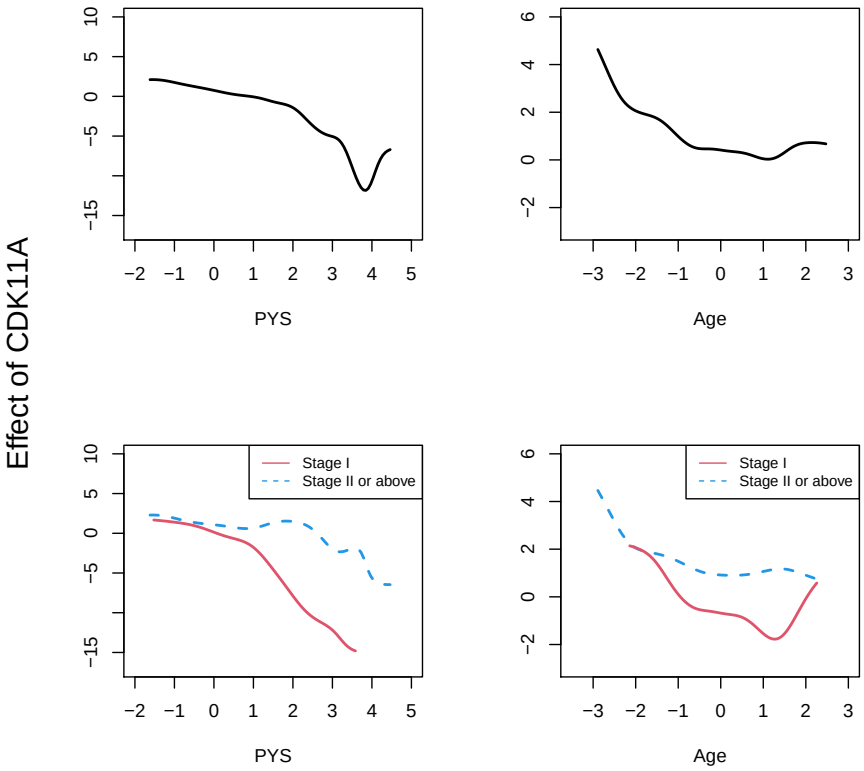

Figure 3 Partial dependence plots for the varying coefficient of CDK11A. The upper two plots show the partial dependence functions for all subjects combined. The lower two plots show the partial dependence functions for subjects with stage I and subjects with stage II or above separately.

4.2 TCGA LGG data

We also applied the proposed methods to identify potential risk factors associated with the survival of patients diagnosed with lower-grade glioma (LGG) in TCGA. The data set consists of grade II and grade III tumors. Instead of integrating clinical and a single type of genomic variables, we investigated the effects of protein expressions, gene expressions, and clinical variables on time to death since initial diagnosis, allowing for interactions between protein and gene expressions. After discarding genes with zero expressions for 30% or more subjects, the data set consists of 17238 gene expressions. We set the overall survival time to be the outcome of interest, which is potentially right-censored. We reduced the dimension of gene expressions using principal component analysis and set $\boldsymbol{U}$ to be the first 7 principal components, which

account for over 50% of the total variability. The set of linear predictors $\boldsymbol{Z}$ consists of $\boldsymbol{U}$, age, histological grade ("0" for grade II and "1" for grade III), and gender ("0" for female and "1" male). The set of predictors $\boldsymbol{X}$ includes the expressions of 209 proteins or phospho-proteins. After removing subjects with missing data, the sample size is 423. The median time to censoring or death is 630 days, and the censoring rate is 76.83%.

The selected protein expressions and their estimated coefficients are shown in Table S2 in the supplementary materials. Among the 11 protein expressions identified to have non-zero main effects, only one protein expression is selected by the proposed method and lasso approach. There is a little overlap between the two methods in the set of selected protein expressions. The results indicate that by allowing for potential interactions between gene expressions and protein expressions, the proposed method reveals a set of risk factors associated with the overall survival of LGG patients. The proposed method identified 6 important protein expressions to be associated with the overall survival. Some of the selected proteins, including HSP70 and Cyclin B1, have previously been shown to be associated with the survival of glioma patients (Chen *et al.*, 2008; Beaman *et al.*, 2014). The effect of Cyclin B1 was identified to vary with the gene expressions. Figure 4 displays the estimated values of g_0 and the g function for Cyclin B1.

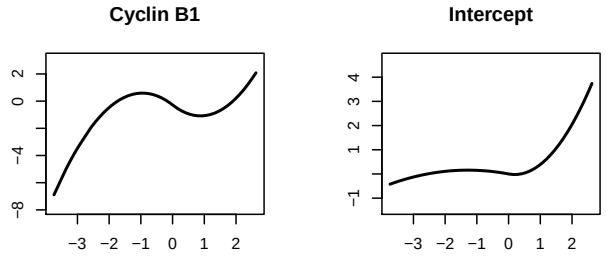

Figure 4 Estimated coefficients for LGG analysis.

5 Discussion

In this paper, we propose a single-index varying-coefficient model for the integration of clinical and genomic variables, where the effects of genomic variables are allowed to vary with clinical variables. The effects of genomic variables are set as non-parametric functions of (a projection of) the clinical variables to accommodate intrinsically different scales of measurements between clinical and genomic variables. The effect modifiers are summarized into an index and thus we can visualize the covariate effects of the genomic features across different index values. Unlike most existing estimation methods for varying-coefficient models, our penalized approach separately selects for predictors with constant effects and those with varying effects. Numerical studies illustrate that the proposed methods effectively distinguish zero, constant, and non-constant effects and yield accurate predictions. Also, we consider a semiparametric proportional hazards model for right-censored outcomes, whereas existing studies only considered complete observations.

In the proposed model, we assume that the main effects of $\boldsymbol{Z}$ are linear. To incorporate non-linear (main) effects, we can include high-order or interaction terms of the covariates to $\boldsymbol{Z}$. For example, as demonstrated in the simulation studies in Section S3.2, quadratic terms can be included as main effects to reduce the MSE. To accommodate more general model structures, such as semiparametric additive models, we can include spline basis functions of covariates into $\boldsymbol{Z}$. This extension results in more parameters to be estimated and can be considered for sufficiently large sample sizes.

Existing machine-learning approaches such as random forests can also capture potential interaction effects between variables. Compared to such machine-learning methods, the proposed methods have two major advantages in terms of interpretation. First, apart from identifying variables that have any effects on the outcome, the proposed methods perform selection on variables with interaction effects. For variables that have main effects but no interaction effects with the effect modifiers, their effects can be interpreted as in standard regression models. Second, the model has simple interpretations at given values of $\boldsymbol{U}$. For example, in the analysis of the LUAD data, the estimated model encodes linear effects of individual gene expressions on the outcome for a subject with specific clinical variable values. By contrast, models from general machine-learning methods do not produce models with similar interpretations.

Some existing works impose a (strong) hierarchical structure for fitting sparse interaction models (Bien *et al.*, 2013; Lim and Hastie, 2015), which requires that a pairwise interaction term between two covariates can be selected only when the main effects of both covariates are selected. In the proposed method, having a hierarchical structure is equivalent to requiring each non-constant varying-coefficient function not to pass through the origin. This structure, however, is not plausible, as a coefficient function can be zero at a fixed value of $\boldsymbol{U}$ but non-zero elsewhere.

There are several possible directions for future research. First, we may be interested in the interaction between two types of high-dimensional predictors, in which case the predictor vector $\boldsymbol{U}$ is high-dimensional. One possible approach is to project $\boldsymbol{U}$ to a low-dimensional space prior to fitting the proposed model. For example, as in the analysis of the LGG data, the projection can be performed by principal component analysis. However, the projected features may not have simple interpretations. Another possible approach is to perform variable selection on $\boldsymbol{U}$ by introducing an extra penalty on $\boldsymbol{\beta}$ (Peng and Huang, 2011; Feng and Xue, 2013, 2015; Radchenko, 2015). This approach would involve substantial computational difficulty due to the introduction of an extra penalty term. Second, it is of interest to consider more than two data types. A possibility is to introduce extra indices that correspond to the extra data types so that the effect of a variable may be a function of multiple indices. This approach, however, faces enormous computational challenges because it involves multivariate non-parametric functions. Third, it would be more flexible to adopt different linear combinations of $\boldsymbol{U}$ across covariates. The proposed approach assumes that a common index $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{U}$ can be used to describe the varying covariate effects of different covariates. This single-index approach can be extended to allow for multiple indices across covariates, with different coefficients $\boldsymbol{\beta}_j$ for different covariates ($j = 1, \ldots, p$). However, this approach is computationally expensive or even infeasible when the number of covariates is large.

Furthermore, it is of interest to perform statistical inference on the effects of the selected features. A simple approach for post-selection inference is based on data splitting, where the model is selected using a subset of study subjects, and estimation and inference are performed using the remaining subjects on the selected model. However, this approach is generally inefficient, because only subsets of subjects are used for variable selection, estimation, and inference. An alternative approach to post-selection inference is based on uniformly valid confidence intervals (Berk *et al.*, 2013). There is a growing literature of similar approaches for linear regression models, but extensions to varying-coefficient (semiparametric) models are highly challenging.

Acknowledgements The work of KY Wong was supported by the Hong Kong Research Grants Council grant PolyU 253042/18P. The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Conflict of Interest

The authors have declared no conflict of interest.

Appendix

Construction of basis functions

We discuss the construction of 2-degree basis functions that pass through the origin and are continuously differentiable; basis functions of a general degree can be constructed analogously. Let $(k_1, \ldots, k_d)$ be an ordered set of grid points, where the number of grid points d is odd and is larger than 2, and $k_{(d+1)/2} = 0$. Let $d' = (d+1)/2$, $(\widetilde{L}_1, \ldots, \widetilde{L}_{d'})$ be a set of 2-degree B-spline functions on $(0, -k_{d'-1}, \ldots, -k_1)$, and $(R_1, \ldots, R_{d'})$ be a set of 2-degree B-spline functions on $(0, k_{d'+1}, \ldots, k_d)$. All B-spline functions do not have an intercept, such that $\widetilde{L}_1(0) = \cdots = \widetilde{L}_{d'}(0) = R_1(0) = \cdots = R_{d'}(0) = 0$. Let $L_j = \widetilde{L}_j(-x)$ for $j = 1, \ldots, d'$. The set of continuously differentiable spline functions spanned by these B-spline functions is therefore

$$\left\{ f = \sum_{j=1}^{d'} c_j L_j + \sum_{j=1}^{d'} c_{j+d'} R_j : (c_1, \ldots, c_{2d'}) \in \mathbb{R}^{2d'}, \sum_{j=1}^{d'} c_j L_j^{(1)}(0) = \sum_{j=1}^{d'} c_{j+d'} R_j^{(1)}(0) \right\},$$

where $h^{(1)}$ denotes the first derivative of the function h. We can then construct the basis function as

$$\left(L_1 + \frac{k_{d'+1}}{k_{d'-1}} R_1, L_2, \ldots, L_{d'}, R_2, \ldots, R_{d'} \right).$$

References

Beaman, G. M., Dennison, S. R., Chatfield, L. K., and Phoenix, D. A. (2014). Reliability of HSP70 (HSPA) expression as a prognostic marker in glioma. *Molecular and Cellular Biochemistry*, 393:301–307.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics*, 41:802–837.

Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *Annals of Statistics*, 41(3):1111–1141.

Boulesteix, A.-L., De Bin, R., Jiang, X., and Fuchs, M. (2017). IPF-LASSO: Integrative-penalized regression with penalty factors for prediction based on multi-omics data. *Computational and Mathematical Methods in Medicine*, 2017(7691937).

Bøvelstad, H. M., Nygård, S., and Borgan, Ø. (2009). Survival prediction from clinico-genomic models-a comparative study. *BMC Bioinformatics*, 10(413).

Breheny, P. and Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and Its Interface*, 2:369–380.

Chen, H., Huang, Q., Dong, J., Zhai, D.-Z., Wang, A.-D., and Lan, Q. (2008). Overexpression of CDC2/CyclinB1 in gliomas, and CDC2 depletion inhibits proliferation of human glioma cells in vitro and in vivo. *BMC Cancer*, 8(29).

Chen, M., Liu, X., Du, J., Wang, X.-J., and Xia, L. (2017). Differentiated regulation of immune-response related genes between LUAD and LUSC subtypes of lung cancers. *Oncotarget*, 8:133–144.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., *et al.* (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486:346–352.

Daemen, A., Gevaert, O., and De Moor, B. (2007). Integration of clinical and microarray data with kernel methods. In *Proceedings of the IEEE Engineering in Medicine and Biology Society*, pages 5411–5415. IEEE.

Daemen, A., Gevaert, O., Ojeda, F., Debucquoy, A., Suykens, J. A., Sempoux, C., Machiels, J.-P., Haustermans, K., and De Moor, B. (2009). A kernel-based integration of genome-wide data for clinical decision support. *Genome Medicine*, 1(39).

Fan, C., Prat, A., Parker, J. S., Liu, Y., Carey, L. A., Troester, M. A., and Perou, C. M. (2011). Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures. *BMC Medical Genomics*, 4(3).

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360.

Feng, S. and Xue, L. (2013). Variable selection for single-index varying-coefficient model. *Frontiers of Mathematics in China*, 8:541–565.

Feng, S. and Xue, L. (2015). Model detection and estimation for single-index varying coefficient model. *Journal of Multivariate Analysis*, 139:227–244.

Gevaert, O., Smet, F. D., Timmerman, D., Moreau, Y., and Moor, B. D. (2006). Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics*, 22:e184–e190.

Guan, S. (2017). *Variable Selection in Varying Multi-index Coefficient Models with Applications to Gene-environmental Interactions*. PhD dissertation, Michigan State University.

Härdle, W., Hall, P., and Ichimura, H. (1993). Optimal smoothing in single-index models. *The Annals of Statistics*, 21:157–178.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *Journal of American Medical Association*, 247:2543–2546.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B*, 55:757–779.

Hole, D., Watt, G., Davey-Smith, G., Hart, C., Gillis, C., and Hawthorne, V. (1996). Impaired lung function and mortality risk in men and women: findings from the Renfrew and Paisley prospective population study. *BMJ*, 313:711–715.

Innamaa, A., Jackson, L., Asher, V., Van Schalkwyk, G., Warren, A., Keightley, A., Hay, D., Bali, A., Sowter, H., and Khan, R. (2013). Expression and effects of modulation of the K2P potassium channels TREK-1 (KCNK2) and TREK-2 (KCNK10) in the normal human ovary and epithelial ovarian cancer. *Clinical and Translational Oncology*, 15:910–918.

Kerin, M. and Marchini, J. (2020). Inferring Gene-by-Environment Interactions with a Bayesian Whole-Genome Regression Model. *The American Journal of Human Genetics*, 107(4):698–713.

Lanckriet, G. R., De Bie, T., Cristianini, N., Jordan, M. I., and Noble, W. S. (2004). A statistical framework for genomic data fusion. *Bioinformatics*, 20:2626–2635.

Landi, M. T., Dracheva, T., Rotunno, M., Figueroa, J. D., Liu, H., Dasgupta, A., Mann, F. E., Fukuoka, J., Hames, M., Bergen, A. W., *et al.* (2008). Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PloS One*, 3(2).

Li, L. (2006). Survival prediction of diffuse large-B-cell lymphoma based on both clinical and gene expression information. *Bioinformatics*, 22:466–471.

Li, W.-C., Xiong, Z.-Y., Huang, P.-Z., Liao, Y.-J., Li, Q.-X., Yao, Z.-C., Liao, Y.-D., Xu, S.-L., Zhou, H., Wang, Q.-L., *et al.* (2019). KCNK levels are prognostic and diagnostic markers for hepatocellular carcinoma. *Aging (Albany NY)*, 11:8169–8182.

Li, Y., Wang, F., Li, R., and Sun, Y. (2020). Semiparametric integrative interaction analysis for non-small-cell lung cancer. *Statistical Methods in Medical Research*, 29:2865–2880.

Lim, M. and Hastie, T. (2015). Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654.

Lin, H., Tan, M. T., and Li, Y. (2016). A semiparametrically efficient estimator of single-index varying coefficient Cox proportional hazards models. *Statistica Sinica*, 26:779–807.

Lin, Y., Zhang, J., Cai, J., Liang, R., Chen, G., Qin, G., Han, X., Yuan, C., Liu, Z., Li, Y., *et al.* (2018). Systematic analysis of gene expression alteration and co-expression network of eukaryotic initiation factor 4A-3 in cancer. *Journal of Cancer*, 9:4568–4577.

Nevins, J. R., Huang, E. S., Dressman, H., Pittman, J., Huang, A. T., and West, M. (2003). Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction. *Human Molecular Genetics*, 12:R153–R157.

Peng, H. and Huang, T. (2011). Penalized least squares for single index models. *Journal of Statistical Planning and Inference*, 141:1362–1379.

Pittman, J., Huang, E., Dressman, H., Horng, C.-F., Cheng, S. H., Tsou, M.-H., Chen, C.-M., Bild, A., Iversen, E. S., Huang, A. T., *et al.* (2004). Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proceedings of the National Academy of Sciences*, 101:8431–8436.

Radchenko, P. (2015). High dimensional single index models. *Journal of Multivariate Analysis*, 139:266–282.

Relli, V., Trerotola, M., Guerra, E., and Alberti, S. (2018). Distinct lung cancer subtypes associate to distinct drivers of tumor progression. *Oncotarget*, 9:35528–35540.

Seoane, J. A., Day, I. N., Gaunt, T. R., and Campbell, C. (2014). A pathway-based data integration framework for prediction of disease progression. *Bioinformatics*, 30:838–845.

Shedden, K., Taylor, J. M., Enkemann, S. A., Tsao, M.-S., Yeatman, T. J., Gerald, W. L., Eschrich, S., Jurisica, I., Giordano, T. J., Misek, D. E., *et al.* (2008). Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nature Medicine*, 14:822–827.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58:267–288.

Wang, W., Baladandayuthapani, V., Morris, J. S., Broom, B. M., Manyam, G., and Do, K.-A. (2013). iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*, 29:149–159.

Wei, F. and Huang, J. (2010). Consistent group selection in high-dimensional linear regression. *Bernoulli (Andover)*, 16(4):1369–1384.

Wei, F., Huang, J., and Li, H. (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Statistica Sinica*, 21(4):1515–1540.

Wong, K. Y., Fan, C., Tanioka, M., Parker, J. S., Nobel, A. B., Zeng, D., Lin, D.-Y., and Perou, C. M. (2019). I-Boost: an integrative boosting approach for predicting survival time with multiple genomics platforms. *Genome Biology*, 20(52).

Xu, J., Jiang, N., Shi, H., Zhao, S., Yao, S., and Shen, H. (2017). miR-28-5p promotes the development and progression of ovarian cancer through inhibition of N4BP1. *International Journal of Oncology*, 50:1383–1391.

Xue, L. and Pang, Z. (2013). Statistical inference for a single-index varying-coefficient model. *Statistics and Computing*, 23:589–599.

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

Zhang, C. H. (2007). Penalized linear unbiased selection. Technical report, Department of Statistics and Bioinformatics, Rutgers University.

Zhao, Q., Shi, X., Xie, Y., Huang, J., Shia, B., and Ma, S. (2015). Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. *Briefings in Bioinformatics*, 16:291–303.

Zhao, Y., Xue, L., and Feng, S. (2019). Estimation for a partially linear single-index varying-coefficient model. *Communications in Statistics - Simulation and Computation*. doi: 10.1080/03610918.2019.1680691.

Zhou, Y., Shen, J. K., Hornicek, F. J., Kan, Q., and Duan, Z. (2016). The emerging roles and therapeutic potential of cyclin-dependent kinase 11 (CDK11) in human cancer. *Oncotarget*, 7:40846–40859.

Zhu, R., Zhao, Q., Zhao, H., and Ma, S. (2016). Integrating multidimensional omics data for cancer outcome. *Biostatistics*, 17:605–618.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429.

