HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2020 October 15.

Published in final edited form as:

Stat Med. 2019 October 15; 38(23): 4625–4641. doi:10.1002/sim.8322.

Robust semiparametric gene-environment interaction analysis using sparse boosting

Mengyun Wu[1,2], **Shuangge Ma**[*,2]

[1]School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China

[2]Department of Biostatistics, Yale University, New Haven, CT, USA

Abstract

For the pathogenesis of complex diseases, gene-environment (G-E) interactions have been shown to have important implications. G-E interaction analysis can be challenging with the need to jointly analyze a large number of main effects and interactions and to respect the "main effects, interactions" hierarchical constraint. Extensive methodological developments on G-E interaction analysis have been conducted in recent literature. Despite considerable successes, most of the existing studies are still limited as they cannot accommodate long-tailed distributions/data contamination, make the restricted assumption of linear effects, and cannot effectively accommodate missingness in E variables. To directly tackle these problems, a semiparametric model is assumed to accommodate nonlinear effects, and the Huber's loss function and Q_n estimator are adopted to accommodate long-tailed distributions/data contamination. A regression-based multiple imputation approach is developed to accommodate missingness in E variables. For model estimation and selection of relevant variables, we adopt an effective sparse boosting approach. The proposed approach is practically well motivated, has intuitive formulations, and can be effectively realized. In extensive simulations, it significantly outperforms multiple direct competitors. The analysis of TCGA data on stomach adenocarcinoma and cutaneous melanoma shows that the proposed approach makes sensible discoveries with satisfactory prediction and stability.

Keywords

Gene-environment interaction; Robustness; Semiparametric modeling; Missingness; Sparse boosting

1 | INTRODUCTION

For understanding, modeling, and treating complex diseases, gene-environment (G-E) interactions have been shown to have a fundamental role beyond the main genetic (G) and environmental (E) effects. Extensive methodological developments have been conducted[1,2].

[*]**Correspondence** Shuangge Ma, Department of Biostatistics, Yale University, New Haven, CT, USA shuangge.ma@yale.edu.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

The existing approaches can be largely classified into two families. The first family conducts the marginal analysis of one or a small number of genes at a time[3,4], whereas the second family conducts the joint analysis of a large number of genes via a single model[5,6]. The etiology, prognosis, and response to treatment of complex diseases are attributable to the combined effects of multiple genetic effects and G-E interactions. As such, joint analysis can be more sensible. With the need to accommodate high data dimensionality and select main effects and interactions that are relevant to the response variable, joint analysis can be very challenging. Another challenge comes from the need to respect the "main effects, interactions" hierarchical constraint, which only allows an interaction into the model if the corresponding main effects are also identified[7]. Violating the constraint causes trouble in estimation and interpretation. For relevant discussions, we refer to the literature[8].

Despite considerable successes, the existing approaches may still be limited in the following aspects. First, most of the existing studies conduct "standard" likelihood-based estimation and cannot accommodate long-tailed distributions/contamination in the response variable. In biomedical studies, responses with long-tailed distributions/contamination are not rare and can be caused by multiple factors, including the inherent variability of data, biased sample selection, errors in data collection/recording, and others[9]. Take the TCGA (The Cancer Genome Atlas) stomach adenocarcinoma (STAD) data, which is analyzed in this article, as an example. There are 157 deaths during follow-up, with survival times ranging from 0.10 to 72.17 months (median 11.6 months). For these subjects, we present the scatter plot of survival times in Figure S1 (Supporting Information), as well as the mean and three times standard deviation. It is observed that there are five subjects with survival times 55.39, 57.39, 59.49, 68.99 and 72.17 months, which are out of the three standard deviation range, suggesting that the data may be "contaminated" with outliers. Nonrobust approaches cannot effectively accommodate long-tailed distributions/contamination and may lead to biased estimation and marker identification. For "classic" low-dimensional biomedical studies, robust approaches have been extensively developed and shown to be powerful. For high-dimensional genetic studies, there are some but still limited developments[10,11,12]. The existing high-dimensional robust studies are mostly limited to main effects, and there is still insufficient attention to genetic interactions[13]. The second limitation of the existing approaches is that they usually assume linear effects. In low-dimensional biomedical studies, it has been shown that in many occasions, nonlinear covariate effects are present[14]. In G-E interaction analysis, it may not be realistic to consider nonlinear effects for the G effects because of the high dimensionality. However, modeling E effects in a nonlinear way can be both feasible and necessary. The third limitation is that most of the existing approaches require complete measurements. With the fast development of profiling techniques, the problem of missingness in G measurements is diminishing. However, as consistently observed in biomedical studies, missingness in E measurements is almost inevitable. It is important to effectively accommodate missingness in E measurements in G-E interaction analysis.

In this article, we conduct G-E interaction analysis that respect the "main effects, interactions" hierarchy under the joint analysis paradigm. A novel approach is developed to directly address the aforementioned limitations of the existing approaches. Specifically, the robust Huber's loss function is adopted to accommodate long-tailed distributions/

contamination in response. Compared to alternative robust approaches including the popular quantile approach, the Huber's approach is computationally more affordable, which is especially important for high-dimensional analysis. A partially linear model is assumed to accommodate nonlinear E effects. A regression-based multiple imputation approach is developed to accommodate missingness in E measurements. Compared to alternatives for example the inverse probability weighting[15], the proposed imputation approach has a more intuitive formulation. For estimation and variable selection in G-E interaction modeling as well as imputation analysis, we adopt a sparse boosting approach, which has competitive performance in high-dimensional data analysis but has not been well employed in genetic interaction analysis. Overall, this study is warranted by providing a practically useful new venue for studying G-E interactions and directly overcoming multiple limitations of the existing literature.

2 | METHODS

2.1 | Partially linear modeling

Let y be the response variable, which can be a continuous marker, categorical disease status, or survival time. $x = \left(x_0 \;\; x_1 \;\; \cdots \;\; x_p\right)$ denotes the $(p + 1)$-dimensional vector of genes (SNPs, or other genetic functional units) with $x_0 = 1$ for intercept. $z = \left(z_1 \;\; \cdots \;\; z_q\right)$ is the q-dimensional vector of E factors. Determining which covariate effects to model nonlinearly is a "classic" problem and has been addressed thoroughly in the literature[16]. Here, with the dimension of E factors usually low, we simply model all continuous E effects in a nonlinear way. For the simplicity of notation, rearrange z so that the first q_1 factors are continuous, and the remaining are discrete. Consider the partially linear model:

$$y \sim \phi\left(\sum_{j=0}^{p} \beta_j x_j + \sum_{k=1}^{q_1}\sum_{j=0}^{p} g_{kj}(z_k)x_j + \sum_{k=q_1+1}^{q}\sum_{j=0}^{p} \alpha_{kj} z_k x_j\right), \quad (1)$$

where β_0 is the intercept, $g_{kj}(\cdot)$'s are unknown functions, and $\phi(\cdot)$ is the known model function, for example, the linear regression model with $\phi(t) = t$ for a continuous response, and logistic model $\phi(t) = \frac{1}{1+e^{-t}}$ for a binary response.

In this model, for $j = 1, \cdots, p$, β_j's, α_{k0}'s and $g_{k0}(\cdot)$'s denote the main effects of G factors, discrete and continuous E factors, respectively, and α_{kj}'s and $g_{kj}(\cdot)$'s represent interactions.

For estimating the nonlinear functions, we conduct basis expansion

$$g_{kj}(z_k) \approx \sum_{l=1}^{L} \gamma_{kjl} b_{kjl}(z_k), \quad (2)$$

where L is the number of basis functions, $\boldsymbol{b}_{kj}(z_k) = \left(b_{kj1}(z_k) \quad \cdots \quad b_{kjL}(z_k)\right)$ is the vector of known basis functions, and $\gamma_{kj} = \left(\gamma_{kj1} \quad \cdots \quad \gamma_{kjL}\right)$ is the vector of unknown coefficients. For identifiability, the basis functions are constrained to have means zero. In numerical study, we adopt the normalized B spline basis, which have been the choice of many published studies[17], and note that other basis functions may also be applicable. Model (1) can now be rewritten as

$$y \sim \phi\left(\beta_0 + \left(\boldsymbol{w}_1 \cdots \boldsymbol{w}_{p+q+pq}\right)\left(\boldsymbol{\theta}_1 \cdots \boldsymbol{\theta}_{p+q+pq}\right)'\right) = \phi(\beta_0 + \boldsymbol{w}\boldsymbol{\theta}'), \quad (3)$$

where $\boldsymbol{w}_j = x_j$ and $\boldsymbol{\theta}_j = \beta_j$ for $j = 1, \cdots, p$, $\boldsymbol{w}_{p+k} = \left(b_{k01}(z_k) \quad \cdots \quad b_{k0L}(z_k)\right)$ and $\boldsymbol{\theta}_{p+k} = \left(\gamma_{k01} \quad \cdots \quad \gamma_{k0L}\right)$ for $k = 1, \cdots, q_1$, $\boldsymbol{w}_{p+k} = z_k$ and $\boldsymbol{\theta}_{p+k} = \alpha_{k0}$ for $k = q_1 + 1, \cdots, q$, $\boldsymbol{w}_{p+q+(j-1)q+k} = \left(b_{kj1}(z_k)x_j \quad \cdots \quad b_{kjL}(z_k)x_j\right)$ and $\boldsymbol{\theta}_{p+q+(j-1)q+k} = \left(\gamma_{kj1} \quad \cdots \quad \gamma_{kjL}\right)$ for $j = 1, \cdots, p$ and $k = 1, \cdots, q_1$ and $\boldsymbol{w}_{p+q+(j-1)q+k} = z_k x_j$ and $\boldsymbol{\theta}_{p+q+(j-1)q+k} = \alpha_{kj}$ for $j = 1, \cdots, p$ and $k = q_1 + 1, \cdots, q$.

2.2 | Robust estimation and selection via sparse boosting

For the simplicity of notation, first consider the most popular linear regression model

$$y = \sum_{j=0}^{p} \beta_j x_j + \sum_{k=1}^{q_1} \sum_{j=0}^{p} g_{kj}(z_k)x_j + \sum_{k=q_1+1}^{q} \sum_{j=0}^{p} \alpha_{kj} z_k x_j + \varepsilon, \quad (4)$$

where ε is the random error. Accommodating survival data will be studied in detail below. Extension to categorical and count data under generalized linear models will also be discussed.

To accommodate long-tailed distributions/contamination in the response, we adopt the Huber's approach. A novel modification is made, which significantly reduces computational cost. To accommodate the high data dimensionality in estimation and to select important interactions (and main effects), we adopt the sparse boosting technique. Directly applying sparse boosting may generate results that violate the "main effects, interactions" hierarchy. To solve this problem, a modified boosting algorithm is developed. Assume n independent observations $\left\{\left(y_i, \boldsymbol{x}_i, z_i\right), i = 1, \cdots, n\right\}$. The proposed approach is summarized in Algorithm 1.

The most prominent consideration of our analysis is on robustness, for which multiple steps have been taken. First, as opposed to the nonrobust least squared loss, a robust loss function is taken in (5). The standard Huber's loss

Algorithm 1 :

G-E interaction analysis via sparse boosting

Step 1: Initialization. Set $m = 0$, and $\boldsymbol{F}^{(m)} = \left(F_1^{(m)} \ \cdots \ F_n^{(m)}\right)' = \mathbf{0}$, where $F_i^{(m)}$ is the estimated effect for subject i at iteration m. With the variable arrangement in (3), set $\mathcal{T}^{(m)} = \left\{1, \cdots, p+q\right\}$ to include all main effects, where $\mathcal{T}^{(m)} \subset \left\{1, \cdots, p+q+pq\right\}$ is the set of variables which can potentially enter the regression model (5) at iteration $m+1$. Set $\mathcal{S}(m) = \emptyset$, where $\mathcal{S}^{(m)} \subset \left\{1, \cdots, p+q+pq\right\}$ is the set of variables with nonzero effects at iteration m and may include both main and interaction effects.

Step 2: Fit and update. $m = m+1$.

Compute

$$\left(\hat{\beta}_{j0}, \hat{\boldsymbol{\theta}}_j\right) = \ \text{argmin} \sum_{i=1}^{n} \tilde{w}_i^{(m)} \left(r_i^{(m)} - \beta_{j0} - \boldsymbol{w}_{ij}\boldsymbol{\theta}_j'\right)^2, \ \ j \in \mathcal{T}^{(m-1)}, \quad (5)$$

where $r_i^{(m)} = y_i - F_i^{(m-1)}$ and $\tilde{w}_i^{(m)} = \psi_c\left(r_i^{(m)}\right) \cdot \psi_c(\cdot)$ is Huber's ψ-function (details below).

For variable selection, compute

$$S_m = \ \underset{j \in \mathcal{T}^{(m-1)}}{\text{argmin}} \left\{\log\left(R_m(j)\right) + df_m(j)\log(n)/n\right\}, \quad (6)$$

where $R_m(j)$ is the robust measure of lack-of-fit (details below), and $df_m(j)$ is the degree of freedom defined as the number of variables that have been selected by iteration m.

Update $\boldsymbol{F}^{(m)} = \boldsymbol{F}^{(m-1)} + \nu\hat{\boldsymbol{r}}^{(m)}$, where ν is the step size, and $\hat{\boldsymbol{r}}^{(m)} = \left(\hat{\beta}_{S_m,0} + \boldsymbol{w}_{1,S_m}\hat{\boldsymbol{\theta}}'_{S_m} \cdots \hat{\beta}_{S_m,0} + \boldsymbol{w}_{n,S_m}\hat{\boldsymbol{\theta}}'_{S_m}\right)'$.

Update $\mathcal{S}^{(m)} = \mathcal{S}^{(m-1)} \cup S_m$.

Update $\mathcal{T}^{(m)}$ by the following rule. If $S_m \notin \mathcal{S}^{(m-1)}$ and $S_m \in \left\{1, \cdots, p\right\}$, that is, the selected variable S_m is a new main G effect, add the corresponding interactions into $\mathcal{T}^{(m)}$ for all main E factors that are already included in $\mathcal{S}^{(m)}$. If $S_m \notin \mathcal{S}^{(m-1)}$ and $S_m \in \left\{p+1, \cdots, p+q\right\}$, that is, the selected variable S_m is a new main E effect, add the corresponding interactions into $\mathcal{T}^{(m)}$ for all main G factors that are already included in $\mathcal{S}^{(m)}$.

Step 3: Iteration and stopping. Repeat Step 2 for M iterations. Estimate the stopping iteration by

$$m_{\text{stop}} = \ \underset{m=1,\ldots,M}{\text{argmin}} \left(\log\left(\sum_{i=1}^{n} \rho_c\left(y_i - F_i^{(m)}\right)\right) + df_m\log(n)/n\right), \quad (7)$$

where $\rho_c(\cdot)$ is the Huber's loss function (details below).

Variables selected and their estimates at iteration m_{stop} are the final results.

function takes the form

$$\rho_c(t) = \begin{cases} t^2 & \text{if } |t| \le c \\ 2c|t| - c^2 & \text{if } |t| > c, \end{cases}$$

where $c > 0$ is a tuning constant[18]. Following published literature[19], we set $c = 1.345\text{MAD}\left(y_i - F_i^{(m-1)}, i = 1, \cdots, n\right)$, where MAD(·) is the median absolute deviation adjusted by a factor of 1.4826. Our preliminary investigation suggests that directly applying this loss function (which demands an iteratively reweighted least squared algorithm and has been referred to as the M-estimation) is computationally very expensive. To solve this problem, what we propose, in a sense, is the first iteration of the M-estimation with the Huber's ψ-function

$$\psi_c(t) = t \cdot \min\left\{1, \frac{c}{|t|}\right\}.$$

In (5), the weight $\tilde{w}_i^{(m)}$'s downweigh the influence of observations with large residuals, leading to robustness to long-tailed distributions/contamination. This modification can significantly reduce computational cost, and our numerical investigation suggests that it can lead to results similar to the M-estimation (details omitted) and has also been suggested in the literature[20]. The second step we take to achieve robustness is in $R_m(j)$. In published studies[22], the sum of squared residuals is adopted as the selection criterion. With the least squared estimator $\beta_{j0} = \bar{r}^{(m)} - \bar{\boldsymbol{w}}_j \boldsymbol{\theta}'_j$, we have

$$\sum_{i=1}^{n} \left(r_i^{(m)} - \beta_{j0} - \boldsymbol{w}_{ij}\boldsymbol{\theta}'_j\right)^2 = n\left(\frac{1}{n}\sum_{i=1}^{n}\left(r_i^{(m)} - \bar{r}^{(m)}\right)^2 - \boldsymbol{\theta}_j\left[\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{w}_{ij} - \bar{\boldsymbol{w}}_j\right)'\left(\boldsymbol{w}_{ij} - \bar{\boldsymbol{w}}_j\right)\right]\boldsymbol{\theta}'_j\right), \quad (8)$$

where $\bar{r}^{(m)} = \frac{1}{n}\sum_{i=1}^{n} r_i^{(m)}$ and $\bar{\boldsymbol{w}}_j = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{w}_{ij}$. We see that the first term is not robust to long-tailed distributions/contamination. To solve this problem, we adopt Q_n[20], a robust scale estimator, which is defined as

$$Q_n\left(r_1, \cdots, r_n\right) = 2.2219\left\{|r_i - r_j|; i < j\right\}_{(k)},$$

with $k = \binom{[n/2]+1}{2}$, and $\{t_1, \cdots, t_n\}_{(k)}$ denoting the kth order statistic of t_i's. Overall, the proposed robust lack-of-fit measure is

$$R_m(j) = Q_n\left(r_1^{(m)}, \cdots, r_n^{(m)}\right) - \boldsymbol{\theta}_j\left[\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{w}_{ij} - \bar{\boldsymbol{w}}_j\right)'\left(\boldsymbol{w}_{ij} - \bar{\boldsymbol{w}}_j\right)\right]\boldsymbol{\theta}'_j. \quad (9)$$

Here, we develop (9) for $R_m(j)$ instead of using $\sum_{i=1}^{n} \tilde{w}_i^{(m)} \left(r_i^{(m)} - \beta_{j0} - w_{ij}\theta_j\right)^2$ in (5) directly. This is because the weights for the same observation are different in different iterations, which may result in an inaccurate measure of the balance between the lack-of-fit and complexity term. The third step we take to achieve robustness is in the selection of m_{stop}, where we adopt the Huber's prediction error – which tends to fit "normal samples" better than "outliers" – as the stopping criterion.

The proposed estimation/selection algorithm fits in the sparse boosting paradigm. It assembles multiple weak learners to achieve a strong learner. When determining in each iteration which variable to be included in the model, both model fitting and model complexity are considered. As pointed out in the literature, sparse boosting tends to generate smaller models than ordinary boosting, which is especially desirable with high-dimensional data. When implementing sparse boosting, we set the step size $\nu = 0.1$ and use BIC for measuring model complexity following published studies[21], where the choice of ν is suggested to be not critical as long as it is small[22]. Beyond robustness, the most significant difference/advancement of the proposed sparse boosting algorithm is that it only searches over those interactions with corresponding main effects already selected in the model. This strategy ensures that the strong hierarchy[7] is respected. There are also other strategies to accommodate hierarchy, such as first searching in the whole space with all main and interaction effects directly, and then adding back the corresponding main effects if specific interactions are present in the final model. The proposed strategy may be advantageous to those in some of the existing studies as its search space is dramatically smaller, which significantly reduces computational cost. This strategy has been partly motivated by the progressive penalization approach[23], which has been developed for genetic interaction analysis using the penalization technique and shown to generate results comparable to searching the whole space (which is much more expensive). Extension of the proposed approach to respect the weak hierarchy is simple and omitted here.

Accommodating other types of response variables—The method described above can be modified to accommodate other types of responses. Consider for example a survival response T under the accelerated failure time (AFT) model:

$$\log(T) = \sum_{j=0}^{p} \beta_j x_j + \sum_{k=1}^{q_1} \sum_{j=0}^{p} g_{kj}(z_k) x_j + \sum_{k=q_1+1}^{q} \sum_{j=0}^{p} \alpha_{kj} z_k x_j + \varepsilon. \quad (10)$$

Under right censoring, denote C as the censoring time, $y = \log(\min(T, C))$, and $\delta = I(T \le C)$. To accommodate censoring, we consider a weighted approach[24], which has computational cost considerably lower than the alternatives. Assume that data $\{(\boldsymbol{x}_i, \boldsymbol{z}_i, y_i, \delta_i), i = 1, \ldots, n\}$ have been sorted according to y_i's from the smallest to the largest. The Kaplan-Meier (KM) weights $\left\{\tilde{w}_i^{(KM)}\right\}_{i=1}^{n}$ can be computed as

$\tilde{w}_1^{(KM)} = \frac{\delta_1}{n}$, $\tilde{w}_i^{(KM)} = \frac{\delta_i}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1}\right)^{\delta_j}$, $i = 2, \cdots, n$. Subject i needs to be

reweighted by $\widetilde{w}_i^{(KM)}$, otherwise, Algorithm 1 mostly remains unchanged. For example, in (5), the weight $\widetilde{w}_i^{(m)}$ needs to be replaced by $\widetilde{w}_i^{(m)} = \psi_c\left(r_i^{(m)}\right)\widetilde{w}_i^{(KM)}$. In numerical study, we examine survival data under the AFT model along with continuous data under the linear regression model.

For categorical and count data under generalized linear models, extensions of the Huber's loss have been developed in the literature[25]. It is expected that such losses can be used in the proposed analysis.

2.3 | Accommodating missingness in E variables

Consider the more practical setting with missingness in E variables. To simplify notation, first consider the scenario where z_1 is continuous and has missingness, and $z_2, \cdots, z_q$ have complete measurements for all subjects. We adopt the multiple imputation technique, which is one of the most popular missing data techniques and has demonstrated competitive performance in many published studies [26]. The proposed multiple imputation approach differs significantly from the existing ones and proceeds as follows.

Modeling covariate relationships—An "informative" imputation demands that the missing values in z_1 depend on the observed values in other variables. Note that the proposed analysis includes independence between z_1 and other E variables as a special case. In the first step, we model z_1 as a function of $z_2, \cdots, z_q$ Specifically, consider the model

$$z_1 = \eta_0 + \sum_{k=2}^{q_1} g_k(z_k) + \sum_{k=q_1+1}^{q} \eta_k z_k + \xi \approx \eta_0 + \left(u_2 \ \cdots \ u_q\right)\left(\eta_2 \ \cdots \ \eta_q\right)' + \xi, \quad (11)$$

where η_0 is the intercept, g_k's are unknown functions (with zero means for identifiability), η_k's are unknown regression coefficients, ξ is the random error with density function $f(\xi)$. Following the rationale and strategy described in the last section, we adopt a basis function expansion approach. For $k = 2, \ldots, q_1$, $\boldsymbol{u}_k = \left(b_{k1}(z_k) \ \cdots \ b_{kL}(z_k)\right)$ is a vector composed of normalized B spline basis functions, L is the number of basis functions, and $\boldsymbol{\eta}_k = \left(\gamma_{k1} \ \cdots \ \gamma_{kL}\right)$ is the vector of unknown coefficients. For $k = q_1 + 1, \cdots, q$, $\boldsymbol{u}_k = z_k$, and $\boldsymbol{\eta}_k = \eta_k$.

Assume that data have been rearranged so that the first n_c subjects have complete measurements, and the remaining $n - n_c$ subjects have z_1 values missing. For estimating the unknown regression coefficients (functions), we propose a sparse boosting approach, which is summarized in Algorithm 2. The notations have similar implications as in Algorithm 1.

Algorithm 2

Estimation of covariate relationships via sparse boosting

Step 1: Initialization. Set $m = 0$ and $\boldsymbol{F}^{(0)} = \left(F_1^{(0)} \ \cdots \ F_{n_c}^{(0)}\right)' = \mathbf{0}$.

Step 2: Fit and update. $m = m + 1$.

Compute

$$\left(\hat{\eta}_{k0}, \hat{\boldsymbol{\eta}}_k\right) = \ \underset{}{\text{argmin}} \sum_{i=1}^{n_c} \left(r_i^{(m)} - \eta_{k0} - \boldsymbol{u}_{ik}\boldsymbol{\eta}_k'\right)^2, \ \ k = 2, \cdots, q, \quad (12)$$

where $r_i^{(m)} = z_{i1} - F_i^{(m-1)}$.

For variable selection, compute

$$S_m = \ \underset{2 \le k \le q}{\text{argmin}} \left\{ \log\left(\sum_{i=1}^{n_c} \left(r_i^{(m)} - \eta_{k0} - \boldsymbol{u}_{ik}\boldsymbol{\eta}_k'\right)^2 \right) + df_m(k)\log\left(n_c\right)/n_c \right\}. \quad (13)$$

Update $\boldsymbol{F}^{(m)} = \boldsymbol{F}^{(m-1)} + \nu\hat{\boldsymbol{r}}^{(m)}$, where $\nu = 0.1$, and $\hat{\boldsymbol{r}}^{(m)} = \left(\hat{\eta}_{S_m,0} + \boldsymbol{u}_{1,S_m}\hat{\boldsymbol{\eta}}_{S_m}' \cdots \hat{\eta}_{S_m,0} + \boldsymbol{u}_{n_c,S_m}\hat{\boldsymbol{\eta}}_{S_m}'\right)'$.

Step 3: Iteration and stopping. Repeat Step 2 for M iterations. Estimate the stopping iteration by

$$m_{\text{stop}} = \ \underset{1 \le m \le M}{\text{argmin}} \left\{ \log\left(\sum_{i=1}^{n_c} \left(z_{i1} - F_i^{(m)}\right)^2 \right) + df_m \log\left(n_c\right)/n_c \right\}. \quad (14)$$

Variables selected and their estimates at iteration m_{stop} are the final results.

We note that for most imputation approaches, explicitly modeling the covariate relationships is not needed. The proposed approach with modeling may have multiple advantages: it more lucidly describes the associations among covariates, which may have independent scientific implications. More importantly, modeling combined with the sparse boosting estimation make it possible to screen out variables not related to z_1 and conduct imputation using only relevant variables. Most of the existing imputation approaches do not have this much desired feature, conduct imputation using all variables (which likely include noises), and hence are less effective. It is noted that some imputation approaches, for example those based on parametric joint distributions, can be viewed as inexplicitly assuming regression models. There are also imputation approaches that have been claimed to be flexible by not assuming specific distributions/models. We note that the proposed model is semiparametric and flexible enough to capture complex relationships among variables. When z_1 has a distribution other than continuous, alternative models (for example, generalized linear

models) can be assumed, and the proposed approach proceeds with minor modifications. Different from the previous section, we adopt nonrobust loss/criteria in sparse boosting, as long-tailed distributions/contamination are not observed in E variables in our data analysis. If needed, robust loss/criteria can be adopted as in the previous section.

Estimating the random error distribution—In some imputation studies, the random error distribution $f(\xi)$ is assumed to be known. In this study, to be more flexible, we propose the following estimation approach, which has been partly motivated by the cross-fitted method for variance estimation[27]: (a) Randomly partition subjects with complete measurements into two subsets D_1 and D_2 with an equal sample size; (b) Apply the estimation method described in Algorithm 2 to D_1; (c) Use the D_1 estimate, make prediction for subjects in D_2, and compute predicted errors; (d) Repeat (a)-(c) 10 times. Conduct a nonparametric estimation of $f(\xi)$ using the predicted errors and denote it as.

Overall strategy—Overall, consider G-E interaction analysis with missing values in z_1. The proposed analysis consists of the following steps:

(a) Estimate the relationship between z_1 and $z_2,\ldots,z_q$ using the method described in Algorithm 2. Estimate $f(\xi)$ using the method described above.

(b) Generate complete datasets based on model (11) and random errors generated from $\hat{f}(\xi)$. For complete dataset $m(= 1,...,\widetilde{M})$, apply the method described in Algorithm 1. Denote $\beta_0^{(m)}, \boldsymbol{\theta}_1^{(m)}, \cdots, \boldsymbol{\theta}_{p+q+pq}^{(m)}$ as the estimate and $\mathcal{S}^{(m)}$ as the set of selected variables.

(c) Combine and generate the final results. In our analysis, selection is at least as important as estimation. Simply taking average across the $\widetilde{M}$ results may generate unsatisfactory selection. We apply the following, which has been motivated by the stability selection [28]:

The final intercept estimate: $\beta_0 = \frac{1}{\widetilde{M}}\sum_{m=1}^{\widetilde{M}} \beta_0^{(m)}$;

The final set of selected variables: $\mathcal{S}_\tau = \left\{j: \frac{1}{\widetilde{M}}\sum_{m=1}^{\widetilde{M}} I\left(j \in \mathcal{S}^{(m)}\right) \geq \tau\right\}$, where $\tau \in (0, 1)$ is selected using the same approach as in stability selection.

The final estimates of regression coefficients: $\boldsymbol{\theta}_j = \begin{cases} \frac{1}{\widetilde{M}}\sum_{m=1}^{\widetilde{M}} \boldsymbol{\theta}_j^{(m)} & \text{if } j \in \mathcal{S}_\tau, \\ 0 & \text{else.} \end{cases}$

Accommodating missingness in multiple E variables—When multiple E variables have missing measurements, we propose adopting an incremental imputation approach. This approach differs from the multiple imputation by chained equations[29] and can effectively avoid problems caused by random initialization. Variables with missingness are first re-ordered based on their missing rates, from the smallest to the largest. Imputation is then conducted one variable at a time, staring from the first. In each step, variables that have been imputed in the previous steps are regarded as non-missing and added to model (11), along

with variables without missingness. This approach uses more information than the one that uses only subjects with complete measurements.

3 | SIMULATION

3.1 | Settings

Performance of the proposed analysis is evaluated using extensive simulations. Summary of the sixteen simulation scenarios is presented in Table S1 (Supporting Information). Under all scenarios, we set $q = 10$ and $p = 1{,}000$. There are thus a total of 1,010 main effects and 10,000 interactions. The following simulation settings are considered, comprehensively covering a wide spectrum of data/model conditions. (a) We consider both continuous and categorical G factors, mimicking gene expression and SNP data, respectively. The continuous G variables are generated from a multivariate normal distribution with marginal means 0 and an auto-regressive correlation structure where the correlation between the jth and kth variables is $0.3^{|j-k|}$. To generate the categorical G variables, we dichotomize the above continuous variables at the 1st and 3rd quartiles and generate 3-level measurements. (b) There are six continuous and four discrete E factors. $z_2, \ldots, z_6$ are simulated from $U(0, 1)$. $z_7, \ldots, z_{10}$ are simulated from a binomial distribution with a success probability of 0.6. To test performance of the imputation approach, z_1 is computed from model (11) with $g_2(z_2) = 2\sin(2\pi z_2)$, $g_3(z_3) = 2\exp(2z_3 - 1) - 2.35$, $g_4(z_4) = -12z_4(1 + z_4) + 10$, and $\eta_7 = \eta_8 = 0.5$. Other variables are irrelevant for z_1. Then, z_1 is re-scaled to the range of 0 to 1. The random error ξ follows a standard normal distribution. (c) We consider two types of response variables and models. The first is a continuous response under model (4). In addition, we also consider censored survival data under the AFT model (10). Here the censoring times are generated from an exponential distribution, where the parameter is adjusted to make the censoring rate around 20%. (d) There are three main E effects (one linear and two nonlinear), eight main G effects, and eleven G-E interactions (three linear and eight nonlinear). The strong hierarchy is satisfied. Specifically, β_1, β_2,…,β_7, β_8, $\alpha_{7,0}$, $\alpha_{7,5}$, $\alpha_{7,6}$ and $\alpha_{7,7}$ are generated from $U(1,1.5)$, and $g_{1,0}(z_1) = 6\sin(2\pi z_1) - 0.06$, $g_{2,0}(z_2) = 6\exp(2z_2 - 1) - 7.05$, $g_{1,1}(z_1) = g_{1,2}(z_1) = g_{1,3}(z_1) = g_{1,4}(z_1) = -2z_1(1 + z_1) + 2$, and $g_{2,1}(z_2) = g_{2,2}(z_2) = g_{2,3}(z_2) = g_{2,4}(z_2) = -4z_2^3 + 1$. The rest effects are zero. In total, there are twelve linear and ten nonlinear effects. (e) Consider four error distributions: $N(0,1)$ (Error 1), $90\%N(0,1) + 10\%Cauchy(0, 5)$ (Error 2), $90\%N(0,1) + 10\%LogNormal(0, 3)$ (Error 3), and $90\%N(0,1) + 10\%Slash$ (i.e. $N(0,1)/U(0,1)$ Error 4). (f) For the continuous and survival responses, set the sample size equal to 150 and 250, respectively. (g) Consider four missingness settings. Under M1 and M2, one E variable (z_1) has missing measurements, whereas under M3 and M4, two E variables (z_1 and z_2) have missing measurements. The missingness mechanism is MAR (missing at random). The missingness probabilities satisfy logistic regression models, whose parameters are adjusted so that the overall missing rates are about 20% for M1 and M3 and 40% for M2 and M4.

To better gauge performance of the proposed analysis, we also conduct extensive comparisons. For the analysis of data without missingness, besides the approach proposed in

Section 2.2 (referred to as "A1"), the following alternatives are considered: (A2) This approach is the same as the proposed, except that all E effects are assumed to be linear; (A3) This approach has the same modeling framework as the proposed. However, it adopts nonrobust loss in estimation, selection procedure, and stopping criterion; (A4) This approach is the same as the proposed, except that the hierarchical structure is not reinforced; (A5) This approach conducts joint analysis assuming linear E effects, adopts nonrobust loss, and uses the MCP penalization for selection[30]. It does not account for the hierarchical structure.

For data with missingness, we consider the proposed approach (referred to as "SBS-A1", where SBS indicates that this approach uses sparse boosting under semiparametric modeling to accommodate missingness) as well as the following alternatives: (CC-A1) This approach conducts complete-case analysis and uses the proposed approach (A1) for estimation and selection; (ML-A2) and (SBP-A2) Both approaches adopt A2 described above for interaction analysis. ML-A2 accommodates missingness using multiple imputation based on linear regression realized by R package *mice*. SBP-A2 accommodates missingness using a sparse boosting approach similar to the proposed with linear weak learners; (MRF-A1) This approach adopts A1 for interaction analysis. It accommodates missingness using multiple imputation based on nonparametric random forests realized by R package *mice*; (SBS-A3) and (SBS-A4) Both approaches adopt the proposed sparse boosting approach (SBS) for multiple imputation, and A3 and A4 for interaction analysis, respectively. For these approaches, the assumed linear or nonlinear effects for E factors are consistent in both imputation and interaction analyses. We set $\widetilde{M} = 10$ in multiple imputation and use the same stability selection approach to generate the final estimates. For the semiparametric models, there are two tuning parameters: the degree of B spline basis and number of interior knots. They can be selected data-dependently, which can be computationally expensive. In simulation, we fix degree=2 and number of knots=2, which generate satisfactory results.

We note that there are other G-E interaction analysis approaches and missing-data approaches that are potentially applicable to the simulated data. The above alternatives are chosen as they have similar frameworks as the proposed and competitive performance among the existing approaches. Specifically, comparing with A2 and A3 can directly establish the merits of the semiparametric modeling and robustness of the proposed approach, respectively. The effectiveness of accommodating hierarchy and adopting the sparse boosting framework are studied by comparing with A4 and A5, respectively. The importance of accommodating missing values using semiparametric modeling with sparse boosting is explored by comparing with CC, ML, MRF, and SBP.

3.2 | Computational cost

Simulation suggests that the proposed approach is computationally affordable. For the analysis of one dataset without missingness, the analysis can be accomplished within five minutes using a laptop with standard configurations. For a dataset with one variable having missingness, the average computational time of the proposed imputation step is 6.33 seconds, compared to 0.41 seconds (ML), 0.91 seconds (MRF), and 2.89 seconds (SBP). If there are two variables with missingness, the proposed imputation step takes about 1.15 minutes, compared to 0.68 seconds (ML), 1.65 seconds (MRF), and 47.55 seconds (SBP).

The proposed approach takes slightly more time, as it involves nonlinear effects, variable selection, as well as the incremental technique, but is still computationally affordable. For the analysis of multiple imputed datasets, the proposed procedure can be realized in a highly parallel manner to reduce computer time. We have developed R code implementing the proposed approach and made it publicly available at www.github.com/shuanggema. To facilitate usage, we have also provided demo for two example datasets with and without missingness, respectively.

3.3 | Results for scenarios without missingness

For each approach, we evaluate identification performance using TP.L (number of true positives for linear effects), TP.NL (number of true positives for nonlinear effects), and FP (number of false positives). Estimation performance is evaluated using mean squared error (EMSE) and mean integrated squared error (EMISE) for all the linear and nonlinear effects, respectively. In addition, an independent testing set with 100 samples is generated for each simulated dataset. Prediction performance is quantified using the prediction mean squared error (PMSE) for the continuous outcome and C-statistic (Cstat) for the survival outcome. The C-statistic is the time-integrated AUC (area under curve) under the time-dependent ROC framework and measures the overall adequacy of risk prediction for censored survival data. The adopted C-statistic estimator[31] is based on the inverse probability of censoring weights and does not assume a specific prediction model. It takes values between 0.5 and 1, with a larger value indicating better prediction. In this study, it is realized using the R function *UnoC* in the package *survAUC.* For each scenario, 200 replicates are simulated, and summary statistics are computed.

Summary results for Scenarios 1–4 and 5–8 are shown in Tables 1 and 2, respectively. The rest of the results are shown in Supporting Information. It is observed that across all simulation scenarios, the proposed approach has competitive performance. For data with a continuous outcome, when there is no contamination (Scenario 1 in Table 1), approach A3, which has a nonrobust loss function, is superior. This result is as expected since the nonrobust alternative can be more efficient for data without contamination. The proposed A1 can more accurately identify both linear and nonlinear effects while having a small number of false positives. More specifically, the proposed approach has TP.NL=8.9, compared to 2.1 (A2), 9.1 (A3), 4.5 (A4), and 4.6 (A5). When data have contamination (Scenarios 2–4 in Table 1), the proposed approach has significant advantages over the alternatives. For example under Scenario 2, the proposed approach has TP.NL=7.7, compared to 2.2 (A2), 4.5 (A3), 3.1 (A4), and 2.0 (A5). In addition, it is observed that without contamination, the proposed approach has prediction performance comparable to A3 and outperforms the robust alternatives. With contamination, the proposed approach has significantly smaller prediction errors. For example under Scenario 3 in Table 1, the proposed approach has PMSE=5.79, compared to 15.51 (A2), 181.86 (A3), 19.40 (A4), and 58.93 (A5). It also behaves better in terms of estimation measured by EMSE and EMISE. For example under Scenario 3 in Table 1, the proposed approach has EMISE=1.25, compared to 37.58 (A2), 71.65 (A3), 3.45 (A4), and 15.34 (A5). To provide a more lucid demonstration, in Figure S2 (Supporting Information), we show the estimation of the nonlinear effects under Scenario 3. It is obvious that the proposed approach provides a more accurate estimation. For data with a

survival outcome (Table 2), the overall observed patterns are similar, with the proposed approach having comparable or superior performance. The observed patterns for data with discrete G variables are also similar.

3.4 | Results for scenarios with missingness

In the analysis of data with missingness, we first examine the effectiveness of the proposed imputation approach by comparing four multiple imputation approaches: ML, MRF, SBP, and the proposed SBS. ML and MRF are based on linear regression and nonlinear random forest, respectively, without conducting variable selection. SBP conducts the selection of relevant variables but assumes linear effects. In Figure S3 (Supporting Information), we show the distributions of imputed z_1 for M1 and M2 using different approaches. P-values are computed from the Kolmogorov-Smirnov tests to examine the differences between imputed distributions and the "true" distribution (without missingness). It is observed that under both settings, the distribution of z_1 estimated using the proposed imputation approach and true distribution are not significantly different. For example, under M1, the proposed approach has p-value=0.5971, compared to 0.000 (ML), 0.0224 (MRF), and 0.0000 (SBP).

Similar to in the previous section, we examine the identification, prediction, and estimation performance of the proposed approach and alternatives. The results are summarized in Tables S4-S19 (Supporting Information). The proposed approach (SBS-A1) is observed to have competitive performance: it identifies the majority of the true positives, while having a small number of false positives, and has higher prediction and estimation accuracy. Comparing the proposed approach with CC-A1 and MRF-A1 suggests the superiority of the proposed imputation approach. The advantage of the proposed approach gets more prominent with an increase in missing rate. For example under M2 with missing rate 40% in Table S5, the proposed approach selects 7.4 true nonzero nonlinear effects, compared to 4.2 (CC-A1), 2.2 (ML-A2), 2.3 (SBP-A2), 4.8 (MRF-A1), 4.7 (SBS-A3), and 2.8 (SBS-A4). It is interesting to observe that, under M1, M2, and M3, the results of SBS-A1 are close to those of A1 with complete measurements, suggesting a very high level of efficiency of the proposed imputation approach.

4 | DATA ANALYSIS

TCGA is a collaborative effort organized by NIH. It conducts comprehensive profiling for multiple cancer types. TCGA data have a high quality. With public availability, they can serve as an ideal testbed for new methodological development. In this section, we analyze TCGA data on stomach (gastric) adenocarcinoma (STAD) and cutaneous melanoma (SKCM). We analyze the processed level 3 data, which are downloaded from TCGA Provisional using the R package *cgdsr.* For G variables, we consider mRNA gene expressions, which are collected using the IlluminaHiseq RNAseq V2 platform.

4.1 | Stomach adenocarcinoma (STAD) data

The response variable of interest is overall survival, which is right censored. The E factors analyzed include age, AJCC metastasis pathologic stage (PM), AJCC nodes pathologic stage (PN), AJCC tumor pathologic stage (PT), gender, ICD O3 histology, ICD O3 site, and

History of other malignancy, all of which have been examined in published studies. The age variable is continuous, and the other seven are discrete. Age is re-scaled to the range of 0 to 1. Re-coding of the discrete variables is described in Supporting Information. To better motivate the nonlinear modeling of age, we first conduct marginal analysis using R package *npregfast*[32]. The regression curve and first order derivative with wild bootstrap-based 95% confidence intervals (shaded area) are presented in Figure S4. The confidence intervals suggest that the first order derivative significantly deviates from constant, and a nonlinear effect is clearly observed. A total of 20,189 gene expression measurements are available on 386 samples. Among them, 381 samples have completely observed survival time, E, and G factors. In this analysis, we simply remove subjects with missing measurements as the missing rates are very low. This analysis can test the approach proposed in Section 2.2. As the number of cancer-related genes is not expected to be large, to improve stability, we conduct a simple prescreening via marginal AFT models. The top 2,000 genes with the smallest p-values are selected for downstream analysis.

As shown in Table 3, the proposed approach identifies 3 main E effects (age, PM, and gender), 45 main G effects, and 23 G-E interactions. The main effects of PM and gender have negative coefficients, and this finding is consistent with that in the literature. The 19 nonlinear age effects are shown in Figure S5. It is observed that most of the estimated effects may not have simple linear approximations, again suggesting the need of nonlinear modeling. Literature search suggests that the identified genes and interactions may have important implications. For example, gene CHRDL2 bounds to bone morphogenetic proteins which may inhibit the proliferation of both normal and malignant gastric epithelial cells[33], suggesting that CHRDL2 can serve as a biomarker of poor prognosis in gastric cancer. Gene LDHB has been found to be down-regulated in gastric cancer samples, resulting in the dysregulation of pyruvic acid efflux in the development of gastric cancer [34]. The copy-number loss of gene LYRM7 has been observed in at least approximately 20% of stomach adenocarcinomas, indicating its important role in stomach adenocarcinomas[35]. The retention of the hyper-phosphorylated state of gene MARCKS has been shown to be responsible for certain mechanisms of protein kinase C (PKC)[36]. Gastric carcinoma and adenocarcinoma cells often show dysregulated PKC-dependent cell signal transduction compared to normal gastric cells, supporting MARCKS as a potential biomarker in gastric cancer[37]. It has been reported that gene PARN is up-regulated in gastric tumor tissues, and PARN-depletion significantly inhibits the proliferation of gastric cancer cell lines MKN28 and AGS and promotes cell death[38]. TOMM20 expression has been detected to be specifically localized to gastric cancer cells and strongly associated with reduced survival, and it has been suggested as a promising biomarker for predicting the prognosis of patients with gastric cancer [39]. Published analysis has also found that gene VAPA is over-expressed in gastric cancer tissues compared to adjacent normal tissues [40].

Beyond the proposed approach, we also analyze data using the alternatives. The summary comparison results are provided in Table S20, including the numbers of overlapping in identified main effects and interactions and corresponding RV-coefficients[41]. The RV-coefficient measures the common information of two data matrices, with a larger value indicating a higher degree of similarity, and provides a more objective measure of overlapping information. Detailed identification results using the alternatives are available

from the authors. It is observed that different approaches identify significantly different sets of interactions and main effects, and the level of overlapping information as measured by the RV-coefficients is moderate. Approach A2, which assumes linear E effects, and approach A4, which does not reinforce the hierarchical structure, identify a small number of interactions. Approach A5, which does not reinforce the hierarchical structure, identifies a few interactions but a very small number of main effects. Both A1 and A3 identify a moderate number of main effects and interactions.

In practical data analysis, it is hard to objectively evaluate the accuracy of identification. To provide partial support to the identification analysis, we evaluate prediction performance using a resampling-based approach. As the response is prognosis, the C-statistic is adopted as the evaluation statistic. With 100 resamplings (5/6 training samples and 1/6 testing samples), we compute the mean C-statistics as 0.65 (proposed), 0.60 (A2), 0.62 (A3), 0.60 (A4), and 0.55 (A5), respectively. The proposed approach has a moderately improved prediction. We also examine stability and compute the observed occurrence index (OOI)[42]. With the same resampling approach as above, the OOI quantifies the probability of a specific effect (interaction or main) identified in random samples, with a larger value indicating higher stability. The mean OOI values across the interactions and main effects identified by the proposed approach is 0.50, compared to 0.43 (A2), 0.46 (A3), 0.11 (A4) and 0.13 (A5). The improved prediction and stability provide support to the validity of the proposed analysis.

4.2 | Cutaneous melanoma (SKCM) data

The response of interest is the (log-transformed) Breslow's depth, which has a continuous distribution and has been suggested as a prognostic marker in melanoma, with deeper tumors correlated with shorter survival. The raw values of the Breslow's depth are nonnegative, and the direct application of the proposed approach may result in unreasonable negative predicted values. Thus, the log-transformation is conducted. Nine E variables are analyzed, including weight, height, clark level, age, PM, PN, PT, gender, and sample type (type). Among them, weight, height, and age are continuous, and the others are discrete. The regression curves and first order derivatives of weight, height, and age are also studied in Figure S4, together with the 95% confidence intervals. Significant nonlinear effects are observed. The continuous E variables are rescaled to the range of 0 to 1. Re-coding of the discrete variables is described in Supporting Information. mRNA gene expressions are analyzed. From the 20,189 measurements, we conduct a prescreening and select 2,000 for downstream analysis. Data are available on 340 subjects, among which 56% have missing measurements in weight, height, and/or clark level.

We assume model (4) and apply the approach developed in Section 2.3. The analysis results using the proposed approach are presented in Table 4. It identifies 7 main E effects (weight, height, clark level, age, PN, PT, and sample type), 35 main G effects, and 43 G-E interactions. The four identified discrete E variables have positive coefficients, indicating positive correlations with the response, which is consistent with published literature. The nonlinear effects of weight, height, and age are shown in Figures S5-S8. Again, we observe considerable curvatures with no simple linear approximations. For the identified genes, we

search the literature for independent evidences of their associations with cutaneous melanoma and find strong support. For example, gene MCAM is a cell surface adhesion that has been detected to be strongly expressed in metastatic melanoma and involves in tumorigenicity and metastasis. Gene ACSL5 has been observed to be critical to the expression of tumor-related factor MCAM, indicating its potential effect on melanoma[43]. Gene EZH2 has been shown to contribute to the transcriptional silencing of tumor suppressor and differentiation genes and be involved in melanoma progression and metastasis [44]. The expression of gene ATP5A1 has been detected prevalently in the cytoplasm of melanoma cells[45]. MSH6 expression has been observed to be absent or extremely low in benign nevi and increased in a subset of primary melanoma samples and also suggested as a valuable marker to improve prognosis assessment in primary melanoma[46]. Gene NLRC4 is an important regulator of key inflammatory signaling pathways in macrophages, and published studies have demonstrated that it plays a critical role in suppressing tumor growth in cutaneous melanoma[47]. Gene TPSO has been found to be involved in the appearance of skin melanoma due to modified functional activities caused by mutagenic activation of MAPK signaling cascade, and has different expression levels between skin melanoma cells and normal melanocytes[48].

Analysis is also conducted using the alternatives. The summary comparison results are shown in Table S21, and detailed estimation results using the alternatives are available from the authors. It is observed that different approaches identify similar main E effects (except for SBS-A4) but different main G effects and interactions. Measured using the RV-coefficients, different sets of identified main E effects have high similarity. ML-A2 and SBP-A2 identify a very small number of interactions. Different from the STAD analysis, SBS-A4 identifies some interactions but no main G effect. Prediction and stability are evaluated as described above. With a continuous outcome, we compute the prediction MSEs, which are 0.15 (proposed), 0.27 (CC-A1), 0.18 (ML-A2), 0.17 (SBP-A2), 0.18 (MRF-A1), 0.20 (SBS-A3), and 0.22 (SBS-A4), respectively. In addition, for the proposed approach, the average OOI is 0.56, compared to 0.14 (CC-A1), 0.50 (ML-A2), 0.49 (SBP-A2), 0.52 (MRF-A1), 0.44 (SBS-A3), and 0.28 (SBS-A4). The proposed approach is again observed to have better prediction performance and stability.

5 | DISCUSSION

G-E interaction analysis has important implications. In this study, we have conducted the challenging joint analysis that respects the “main effects, interactions” hierarchy. We have developed a novel analysis approach, which advances from the existing literature in multiple aspects. To achieve the much desired robustness property, we adopt the Huber’s approach and make important modifications to reduce computational cost. To describe the effects of E variables in a more flexible manner, we adopt semiparametric modeling. In this study, E factors are modeled separately. There are also studies that model multiple E factors together, such as the partial linear varying multi-index coefficient model[49]. These two strategies have been developed with different considerations. Specifically, the proposed one assumes that each E factor has a nonlinear effect, and the latter one assumes that the linear combination of all E factors has a nonlinear effect. Neither of them can be viewed as a special case of the other and is consistently better than the other under all scenarios. In data analysis, it is

difficult to objectively determine the underlying model, and how to construct the most suitable semiparametric model is still an open question. We adopt the former one, as it has satisfactory performance and a simple optimization procedure, and is also a popular choice in recent publications. For estimation and selection, we adopt sparse boosting, which has competitive performance in statistical learning but has not been well adopted in G-E interaction analysis. Significantly advancing from the existing sparse boosting studies, a robust loss and robust criteria are adopted. More importantly, the boosting algorithm is modified to respect the "main effects, interactions" hierarchy. This hierarchy has been extensively studied in the published G-E interaction analysis, especially under the penalization framework. The hierarchical Lasso is perhaps the most representative one[7]. Despite considerable successes, the hierarchical Lasso has a complex optimization problem and high computational cost, and is less desirable for high dimensional interaction analysis. Although the proposed hierarchy strategy may be not as straightforward as under penalization, it is still warranted by having a much simpler optimization and satisfactory numerical performance. Another important problem addressed in this study is the missingness in E variables, which is commonly encountered but has been largely neglected in published studies. For this problem, we have developed a novel imputation approach based on a flexible semiparametric regression model, which explicitly describes the covariate relationships and has lucid interpretations. We adopt sparse boosting for estimation to be "consistent" with the G-E interaction analysis. More importantly, the proposed approach can effectively remove irrelevant variables, conduct imputation based on only the relevant ones, and hence is more effective. Extensive simulations show that the proposed approach significantly outperforms multiple state-of-the-art direct competitors. In the analysis of two TCGA datasets, interactions and main effects different from those using the alternatives are identified. The identified genes have important implications. Satisfactory prediction and stability provide partial support to the validity of the proposed analysis.

This study can be potentially extended in multiple directions. For G-E interaction analysis, robust approaches are still limited. It can be of interest to extend the Huber's approach to other data/model settings and develop other robust measures. Semiparametric modeling, which has been shown to be very powerful in low-dimensional biomedical studies, also has limited applications in genetic interaction analysis. Sparse boosting is a generically applicable learning technique and can be potentially coupled with other loss functions and data/model settings. Multiple robust methods have been developed in both low-dimensional and high-dimensional main effect analysis, among which the least absolute deviation (LAD) is one of the most representative. It has been demonstrated that LAD is suitable for heavy-tail distributions, especially double-exponential distributions, and Huber's loss has good performance for contaminated normal distributions[50,12]. However, no method can perform universally better than the other[13]. The boosting technique with Huber's loss and LAD for low-dimensional main effect analysis has been studied in the literature[19]. Huber's loss is observed to perform well with both normal and slash errors, whereas LAD has limitations with normal errors. Sparse boosting with LAD for interaction analysis has not been well examined and will be deferred to future study. Missing data has been studied in genetic studies, but most of the existing attention has been on the G variables. This study is one of the few with a special emphasis on missingness in E variables. For examining the nonlinear

effects of continuous E factors, confidence intervals are constructed under marginal analysis. Studies on the inference of regression curves under high dimensional interaction models are still limited and expected to be very challenging. In this study, we have focused on estimation and will postpone inference investigation to future research. In data analysis, findings different from the alternatives are made and have important biological implications. The prediction and stability evaluation provide partial support. More definitive confirmation will need to come from functional validations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank the editor and reviewers for their careful review and insightful comments, which have led to a significant improvement of this article. This work was supported by the National Institutes of Health [CA216017, CA121974, CA204120]; National Natural Science Foundation of China [91546202, 71331006]; Bureau of Statistics of China [2018LD02]; “Chenguang Program” supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission [18CG42]; Program for Innovative Research Team of Shanghai University of Finance and Economics.

References

1. Thomas D Gene-environment-wide association studies: emerging approaches. Nat Rev Genet. 2010;11:259–272. [PubMed: 20212493]
2. Simonds NI, Ghazarian AA, Pimentel CB, Schully SD, Ellison GL, Gillanders EM, Mechanic LE. Review of the gene-environment interaction literature in cancer: what do we know? Genet Epidemiol. 2016;40(5):356–365. [PubMed: 27061572]
3. Zhang P, Lewinger JP, Conti D, Morrison JL, Gauderman WJ. Detecting gene-environment interactions for a quantitative trait in a genome-wide association study. Genet Epidemiol. 2016;40(5):394–403. [PubMed: 27230133]
4. Xu Y, Wu M, Zhang Q, Ma S. Robust identification of gene-environment interactions for prognosis using a quantile partial correlation approach. Genomics. 2018. doi:10.1016/j.ygeno.2018.07.006
5. Hung H, Lin Y T, Chen P, Wang CC, Huang SY, Tzeng JY. Detection of gene-gene interactions using multistage sparse and low-rank regression. Biometrics. 2016; 72(1):85–94. [PubMed: 26288029]
6. Wu C, Jiang Y, Ren J, Cui Y, Ma S. Dissecting gene-environment interactions: a penalized robust approach accounting for hierarchical structures. Stat Med. 2018;37:437–456. [PubMed: 29034484]
7. Bien J, Taylor J, Tibshirani R. A Lasso for hierarchical interactions. Ann Stat. 2013;41(3):1111–1141. [PubMed: 26257447]
8. Hao N, Zhang H. A note on high dimensional linear regression with interactions. Am Stat. 2017;71:291–297.
9. Osborne JW, Overbay A. The power of outliers (and why researchers should always check for them). Pract Assess Res Eval. 2010;9(6):1–12.
10. Fan J, Fan Y, Barut E. Adaptive robust variable selection. Ann Stat. 2014;42(1):324–351. [PubMed: 25580039]
11. Zhong W, Zhu L, Li R, Cui H. Regularized quantile regression and robust feature screening for single index models. Stat Sinica. 2016;26(1):69–95.
12. Wu C, Ma S. A selective review of robust variable selection with applications in bioinformatics. Brief Bioinform. 2015;16(5):873–883. [PubMed: 25479793]
13. Wu M, Ma S. Robust genetic interaction analysis. Brief Bioinform. 2019;20(2):624–637. [PubMed: 29897421]

14. Stark A, Stahl MS, Kirchner HL, Krum S, Prichard J, Evans J. Body mass index at the time of diagnosis and the risk of advanced stages and poorly differentiated cancers of the breast: findings from a case-series study. Int J Obes. 2010;34(9):1381–1386.
15. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res. 2013;22(3):278–295. [PubMed: 21220355]
16. Sindelar R, Babuska R. Input selection for nonlinear regression models. IEEE Trans Fuzzy Syst. 2004;12(5):688–696.
17. Wu C, Shi X, Cui Y, Ma S. A penalized robust semiparametric approach for gene-environment interactions. Stat Med. 2015;34:4016–4030. [PubMed: 26239060]
18. Huber PJ. Robust estimation of a location parameter. Ann Math Stat. 1964;35:73–101.
19. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat. 2001;29(5): 1189–1232.
20. Lutz RW, Kalisch M, Buhlmann P. Robustified L2 boosting. Comput Stat Data Anal. 2008;52(7): 331–3341.
21. Huang Y, Liu J, Yi H, Shia BC, Ma S. Promoting similarity of model sparsity structures in integrative analysis of cancer genetic data. Stat Med. 2017;36(3):509–559. [PubMed: 27667129]
22. Bühlmann P, Yu B. Sparse boosting. J Mach Learn Res. 2006;7:1001–1024.
23. Zhu R, Zhao H, Ma S. Identifying gene-environment and gene-gene interactions using a progressive penalization approach. Genet Epidemiol. 2014;38(4):353–368. [PubMed: 24723356]
24. Stute W Distributional convergence under random censorship when covariables are present. Scand J Stat. 1996;23:461–471.
25. Noh M, Lee Y. Robust modeling for inference from generalized linear model classes. J Am Stat Assoc. 2007;102(479):1059–1072.
26. Morris TP, White IR, Carpenter JR, Stanworth SJ, Royston P. Combining fractional polynomial model building with multiple imputation. Stat Med. 2015,34(25):3298–3317. [PubMed: 26095614]
27. Fan J, Guo S, Hao N. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. J Royal Stat Soc B. 2012;74(1):37–65.
28. Meinshausen N, Buhlmann P. Stability selection. J Royal Stat Soc B. 2010;72(4):417–473.
29. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med. 2011;30(4):377–399. [PubMed: 21225900]
30. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. Ann Stat. 2010;38(2):894–942.
31. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med. 2011;30(10):1105–1117. [PubMed: 21484848]
32. Sestelo M, Villanueva NM, Meiramachado L, et al. npregfast: An R package for nonparametric estimation and inference in life sciences. J Stat Softw. 2017; 82(12). doi: 10.18637/jss.v082.i12
33. Zhang J, Ge Y, Sun L, et al. Effect of bone morphogenetic protein-2 on proliferation and apoptosis of gastric cancer cells. Int J Med Sci. 2012;9(2):184–192. [PubMed: 22359486]
34. Cai Z, Zhao J, Li J, et al. A combined proteomics and metabolomics profiling of gastric cardia cancer reveals characteristic dysregulations in glucose metabolism. Mol Cell Proteomics. 2010;9(12):2617–2628. [PubMed: 20699381]
35. Cutcutache I, Wu A, Suzuki Y, et al. Abundant copy-number loss of CYCLOPS and STOP genes in gastric adenocarcinoma. Gastric Cancer. 2016;19(2):453–465. [PubMed: 26205786]
36. Yokoyama Y, Ito T, Hanson V, et al. PMA-induced reduction in invasiveness is associated with hyperphosphorylation of MARCKS and talin in invasive bladder cancer cells. Int J Cancer. 1998;75(5):774–779. [PubMed: 9495248]
37. Fahrmann M Targeting protein kinase C (PKC) in physiology and cancer of the gastric cell system. Curr Med Chem, 2008;15(12):1175–1191. [PubMed: 18473812]
38. Zhang L, Yan Y. Depletion of poly (A)-specific ribonuclease (PARN) inhibits proliferation of human gastric cancer cells by blocking cell cycle progression. Biochim Biophys Acta Mol Cell Res. 2015;1853(2):522–534.

39. Zhao Z, Han F, He Y, et al. Stromal-epithelial metabolic coupling in gastric cancer: stromal MCT4 and mitochondrial TOMM20 as poor prognostic factors. Eur J Surg Oncol. 2014;40(10):1361–1368. [PubMed: 24821064]
40. Gao W, Xu J, Wang F, et al. Plasma membrane proteomic analysis of human Gastric Cancer tissues: revealing flotillin 1 as a marker for Gastric Cancer. BMC Cancer. 2015;15(1):367. [PubMed: 25948494]
41. Smilde AK, Kiers HAL, Bijlsma S, Rubingh CM, Van Erk MJ. Matrix correlations for high-dimensional data: the modified RV-coefficient. Bioinformatics. 2009;25(3):401–405. [PubMed: 19073588]
42. Huang J, Ma S. Variable selection in the accelerated failure time model via the bridge method. Lifetime Data Anal. 2010;16(2):176–195. [PubMed: 20013308]
43. Mashima T, Sato S, Sugimoto Y, et al. Promotion of glioma cell survival by acyl-CoA synthetase 5 under extracellular acidosis conditions. Oncogene. 2009;28(1):9–19. [PubMed: 18806831]
44. Mahmoud F, Shields B, Makhoul I, et al. Role of EZH2 histone methyltrasferase in melanoma progression and metastasis. Cancer Biol Ther. 2016;17(6):579–591. [PubMed: 27105109]
45. Ho J, de Moura MB, Lin Y, et al. Importance of glycolysis and oxidative phosphorylation in advanced melanoma. Mol Cancer. 2012;11(1):76. [PubMed: 23043612]
46. Alvino E, Passarelli F, Cannavo E, et al. High expression of the mismatch repair protein MSH6 is associated with poor patient survival in melanoma. Am J Clin Pathol. 2014;142(1):121–132. [PubMed: 24926095]
47. Janowski AM, Colegio OR, Hornick E, et al. NLRC4 suppresses melanoma tumor progression independently of inflammasome activation. J Clin Investig. 2016; 126(10):3917–3928. [PubMed: 27617861]
48. Gyrylova SN, Ruksha TG, Komina AV. TSPO ligand PK11195 and MAPK inhibitor UO126 modulate TSPO expression in melanoma cells. Cell Tissue Biol. 2013;7(3):266–270.
49. Liu X, Cui Y, Li R. Partial linear varying multi-index coefficient model for integrative gene-environment interactions. Stat Sinica. 2016;26:1037–1060.
50. Bradic J, Fan J, Wang W. Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. J Royal Stat Soc B. 2011;73(3):325–349.

TABLE 1

Simulation Scenarios 1–4 without missingness. In each cell, mean (sd) based on 200 replicates.

	TP.L	TP.NL	FP	EMSE	EMISE	PMSE
				Scenario 1		
A1	8.8(1.7)	8.9(1.5)	9.7(3.6)	0.69(0.24)	1.03(0.60)	4.18(1.83)
A2	7.0(1.0)	2.1(0.4)	14.1(3.4)	1.63(0.35)	38.22(7.32)	14.40(4.63)
A3	9.3(1.8)	9.1(1.4)	22.5(5.8)	0.65(0.23)	1.08(0.59)	4.42(2.08)
A4	4.8(1.5)	4.5(1.5)	34.1(3.6)	1.23(0.35)	2.16(0.58)	12.13(4.73)
A5	2.9(1.2)	4.6(1.1)	24.5(8.6)	1.53(0.25)	42.43(7.83)	9.27(3.16)
				Scenario 2		
A1	7.6(1.5)	7.7(2.0)	12.1(4.7)	0.92(0.29)	1.39(0.65)	8.03(3.94)
A2	6.2(1.2)	2.2(0.4)	13.7(3.9)	1.92(0.41)	37.37(7.44)	18.54(5.28)
A3	4.8(2.7)	4.5(2.6)	35.6(6.9)	39.16(204.26)	64.57(269.66)	179.77(713.88)
A4	3.6(1.7)	3.1(1.0)	36.5(4.6)	1.36(0.33)	3.00(0.53)	19.49(6.41)
A5	1.4(1.3)	2.0(1.6)	17.4(12.8)	1.77(0.43)	22.78(16.59)	39.98(60.82)
				Scenario 3		
A1	8.2(1.5)	8.3(1.7)	10.5(4.2)	0.79(0.26)	1.25(0.77)	5.79(2.64)
A2	6.5(1.1)	2.1(0.3)	12.8(3.9)	1.76(0.38)	37.58(6.92)	15.51(4.15)
A3	3.8(2.6)	3.3(2.5)	37.4(6.6)	42.00(131.25)	71.65(281.32)	181.86(395.73)
A4	3.4(2.0)	2.7(0.7)	31.0(9.3)	1.30(0.34)	3.45(1.01)	19.40(9.23)
A5	0.8(1.2)	1.2(1.5)	13.2(13.5)	1.81(0.40)	15.34(15.00)	58.93(80.15)
				Scenario 4		
A1	8.6(1.6)	8.5(1.5)	10.3(4.5)	0.74(0.25)	1.10(0.63)	5.51(2.55)
A2	6.7(1.2)	2.1(0.4)	13.8(3.8)	1.75(0.37)	38.60(6.96)	16.66(5.04)
A3	7.9(2.5)	7.5(2.5)	28.1(7.0)	1.13(0.95)	2.60(2.88)	11.74(13.22)
A4	4.1(1.7)	3.4(1.1)	36.8(2.9)	1.27(0.23)	2.47(0.67)	18.49(11.75)
A5	2.5(1.3)	3.7(1.6)	26.0(12.0)	1.67(0.39)	39.12(14.23)	15.91(11.57)

TABLE 2

Simulation Scenarios 5–8 without missingness. In each cell, mean (sd) based on 200 replicates.

	TP.L	TP.NL	FP	EMSE	EMISE	PMSE
			Scenario 5			
A1	8.6(1.4)	8.9(1.2)	9.8(3.2)	0.72(0.20)	1.54(0.81)	0.87(0.04)
A2	7.1(1.2)	2.3(0.5)	13.5(3.2)	1.25(0.28)	27.70(4.77)	0.79(0.03)
A3	9.0(1.7)	9.4(0.8)	19.0(5.7)	0.67(0.21)	1.48(0.82)	0.88(0.04)
A4	3.7(1.3)	3.2(1.1)	34.3(2.5)	1.30(0.24)	2.96(1.13)	0.77(0.06)
A5	2.7(0.9)	5.0(0.9)	34.9(12.9)	1.48(0.22)	35.27(5.09)	0.85(0.03)
			Scenario 6			
A1	8.2(1.3)	8.1(1.6)	10.6(4.3)	0.79(0.20)	1.75(0.89)	0.83(0.05)
A2	6.5(1.2)	2.3(0.5)	13.3(3.5)	1.40(0.29)	27.60(5.26)	0.76(0.05)
A3	5.5(2.6)	4.9(2.6)	31.8(6.0)	3.76(9.49)	20.15(90.37)	0.70(0.10)
A4	3.1(1.1)	3.4(0.8)	35.5(3.4)	1.52(0.33)	3.12(0.50)	0.71(0.05)
A5	1.4(1.3)	2.8(2.0)	28.5(19.8)	1.68(0.29)	20.90(14.65)	0.69(0.14)
			Scenario 7			
A1	8.8(1.6)	8.7(1.3)	9.6(4.2)	0.70(0.21)	1.56(0.80)	0.86(0.04)
A2	6.9(1.1)	2.3(0.5)	13.9(3.9)	1.32(0.26)	28.60(5.25)	0.78(0.04)
A3	8.8(1.5)	9.2(1.1)	20.1(5.6)	0.71(0.22)	1.49(0.75)	0.85(0.04)
A4	4.3(1.3)	3.6(1.1)	34.7(2.3)	1.26(0.25)	2.73(0.91)	0.76(0.05)
A5	3.0(1.0)	5.1(0.9)	33.3(9.9)	1.43(0.25)	34.19(5.24)	0.83(0.04)
			Scenario 8			
A1	8.3(1.2)	8.6(1.3)	10.4(3.8)	0.72(0.19)	1.65(0.69)	0.86(0.05)
A2	7.0(1.3)	2.2(0.6)	13.7(3.6)	1.29(0.32)	28.21(6.00)	0.79(0.04)
A3	7.8(2.2)	7.7(2.3)	25.4(8.3)	8.17(55.15)	22.12(146.44)	0.80(0.11)
A4	3.9(1.5)	3.6(0.8)	33.5(3.0)	1.30(0.34)	2.76(0.63)	0.76(0.06)
A5	2.2(1.4)	4.1(1.6)	34.8(15.4)	1.58(0.32)	30.91(9.78)	0.79(0.12)

TABLE 3

Analysis of STAD data using the proposed approach: identified main effects and interactions.

	Main G	Age	PM	Gender
Main E		Nonlinear	−0.086	−0.017
ADCY10P1	0.103			
ARHGEF39	−0.014		−0.091	
C12ORF56	−0.059		−0.140	
C4ORF32	−0.189			
C5ORF58	−0.108			
C9ORF40	−0.069			
CHRDL2	−0.052	Nonlinear	−0.093	
CISH	−0.151			
DCST2	−0.047	Nonlinear		
DCTN5	−0.062	Nonlinear		
DDX59	−0.048			
DRD4	−0.101			
EMG1	−0.037			
FAHD1	−0.070			
GCNT1	−0.104	Nonlinear		
GP9	−0.004	Nonlinear		
GTF3C6	−0.050			
HAAO	−0.058	Nonlinear	0.046	
HPDL	−0.028			
HS6ST2	−0.016	Nonlinear	0.054	
KLHL30	−0.061			
LDHB	0.030			
LECT1	−0.152	Nonlinear		
LETM2	−0.088			
LINC00998	−0.046			
LINC01003	−0.112			
LYRM7	−0.011	Nonlinear	−0.181	
MAK16	0.076			
MAN2A2	0.056			
MARCKS	−0.059			
MB21D2	−0.055	Nonlinear		
NXF3	−0.077	Nonlinear		
OR10A5	−0.045	Nonlinear		
OXCT2	0.170			
PAGE2	−0.041			
PARN	−0.087			
PITX2	−0.091	Nonlinear		
PLA2G2F	−0.170	Nonlinear		

	Main G	Age	PM	Gender
RAB41	−0.060			
SS18L2	0.030			
TAF4B	−0.035	Nonlinear		
TEX36.AS1	−0.662			
TOMM20	−0.070	Nonlinear		
TRMT61B	0.010	Nonlinear		
VAPA	0.015	Nonlinear		

TABLE 4

Analysis of SKCM data using the proposed approach: identified main effects and interactions.

	Main G	Weight	Height	Clark level	Age	PN	PT	Type
Main E		Nonlinear	Nonlinear	0.258	Nonlinear	1.142	0.108	0.575
ACSL5	−0.001				Nonlinear			
AKR1C6P	−0.006	Nonlinear	Nonlinear					
ANKRD26P3	0.101				Nonlinear			
ARMCX2	−0.043	Nonlinear	Nonlinear					
ARSH	0.083	Nonlinear	Nonlinear		Nonlinear			
ATP13A2	0.021							
ATP5A1	0.029		Nonlinear		Nonlinear			
ATRIP	0.009		Nonlinear					
DLL4	0.010							
ELFN1	0.005	Nonlinear	Nonlinear		Nonlinear			
EZH2	0.018							
FOLH1B	−0.008							
GAS2	−0.006					−0.037		
GOLGA7B	0.006							
HEPHL1	0.013							
HEXA	0.014		Nonlinear		Nonlinear			
LCN2	0.012		Nonlinear		Nonlinear			
MCAM	0.097							
MSH6	−0.008		Nonlinear					
NLRC4	0.023							
NPB	−0.047		Nonlinear		Nonlinear			
OR10C1	−0.134					−0.112		
PAPOLG	0.013							
PLK5	0.029		Nonlinear		Nonlinear			
PSMD7	−0.016	Nonlinear	Nonlinear		Nonlinear	−0.052		
SCARNA9	−0.007							
SDF2	0.016							
SDF4	0.028							
SDPR	−0.039	Nonlinear	Nonlinear					−0.469
SLC4A10	−0.003							
SLC9A8	0.031		Nonlinear		Nonlinear			
SNX3	−0.048	Nonlinear	Nonlinear		Nonlinear			−0.035
SPRYD7	0.008	Nonlinear				0.047		
TMEM97	0.032		Nonlinear					
TSPO	−0.028	Nonlinear						