Published in final edited form as:

Stat Med. 2019 April 30; 38(9): 1620–1633. doi:10.1002/sim.8064.

Identifying gene-environment interactions incorporating prior information

Xiaoyan Wang^{1,3}, Yonghong Xu², and Shuangge Ma³

¹College of Finance and Statistics, Hunan University, Changsha, China

²School of Economics, Xiamen University, Xiamen, China

³Department of Biostatistics, Yale University, New Haven, Connecticut

Abstract

For many complex diseases, gene-environment (G-E) interactions have independent contributions beyond the main G and E effects. Despite extensive effort, it still remains challenging to identify G-E interactions. With the long accumulation of experiments and data, for many biomedical problems of common interest, there are existing studies that can be relevant and informative for the identification of G-E interactions and/or main effects. In this study, our goal is to identify G-E interactions (as well as their corresponding main G effects) under a joint statistical modeling framework. Significantly advancing from the existing studies, a quasi-likelihood-based approach is developed to incorporate information mined from the existing literature. A penalization approach is adopted for identification and selection and respects the “main effects, interactions” hierarchical structure. Simulation shows that, when the existing information is of high quality, significant improvement can be observed. On the other hand, when the existing information is less informative, the proposed method still performs reasonably (and hence demonstrates a certain degree of “robustness”). The analysis of The Cancer Genome Atlas (TCGA) data on cutaneous melanoma and glioblastoma multiforme demonstrates the practical applicability of the proposed approach and also leads to sensible findings.

Keywords

G-E interaction; penalized joint analysis; prior information; quasi-likelihood

1 INTRODUCTION

For many complex diseases, gene-environment (G-E) interactions have independent contributions beyond the main G and E effects. A myriad of examples is available in the literature. With the high data dimensionality, often low signal levels, and additional complexity brought by interactions beyond main effects, the identification of G-E

Correspondence Yonghong Xu, School of Economics, Xiamen University, Xiamen 361005, China. xustatistics@gmail.com.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

interactions is challenging, and accordingly, extensive statistical developments have been conducted.

Generically, most of the existing approaches can be classified as marginal analysis or joint analysis. In marginal analysis,¹ one or a small number of G measurements are analyzed at a time, and this analysis is usually casted as a hypothesis testing problem. Marginal analysis is easy to implement and has led to many interesting findings, however, may contradict the fact that outcomes and phenotypes of complex diseases are associated with the joint contributions of multiple main G and E effects and their interactions. In joint analysis, the second family of analysis, a large number of G measurements are jointly analyzed. Although multiple approaches have been developed, the most popular, which is adopted in this study, is the statistical modeling approach, under which interactions are represented as product terms (of main effects) in a statistical model. Joint analysis is challenging with the high dimensionality: the model with interactions may have a much higher dimensionality than that with main effects only. In addition, the need for respecting the “main effects, interactions” hierarchy brings significant additional complexity. Recently published joint analysis studies include those by Bien et al,² Liu et al,³ Wu et al,⁴ and others.

It has been well acknowledged that the analysis of genetic data can be challenged by “a lack of information” caused by the high dimensionality, low signal levels, need for joint analysis, and other factors. Such a problem gets even more serious in interaction analysis. Literature review suggests that, for many common biomedical problems such as those analyzed in this article, there exist multiple relevant studies, which may provide valuable “prior information.” Incorporating prior information in analysis is by no means new. When the existing studies and the one to be analyzed have comparable designs, meta-analysis (of summary statistics) or integrative analysis (of raw data) can be conducted. For relevant discussions, we refer to the works of Zeggini et al,⁵ Ma et al,⁶ and Zhao et al.⁷ However, such analysis demands a high level of comparability across studies/datasets, may need to exclude studies that are only partially relevant, and cause a waste of information. Consider, for example, the analysis of melanoma prognosis as presented in Section 5. The study by Lobach et al⁸ also conducts the G-E interaction analysis of melanoma prognosis, and its results can be potentially incorporated via meta-analysis or integrative analysis. But, the study by Merlino and Noonan⁹ is on the effect of solar ultraviolet radiation on melanoma development and hence does not fit the meta-analysis or integrative analysis framework. However, it suggests the association between gene BRAF and melanoma prognosis and hence can be informative for the identification of main G effects. With still limited information, it is desirable to include as many relevant studies as possible.

In this article, our goal is to more effectively conduct G-E interaction analysis by incorporating existing prior knowledge. The proposed method conducts joint analysis under the statistical modeling framework. The most prominent advancement is the incorporation of prior information, which enables boosting the performance of analysis by borrowing information from many existing studies. Significantly different from the existing meta-analysis and integrative analysis, a strict requirement condition on comparability (of studies/datasets) is not required, allowing the incorporation of a large amount of existing information. To the best of our knowledge, the only available approach that shares a similar

strategy is the prior Lasso,¹⁰ which is limited to the analysis of main effects and linear regression model. Also, different from many of the existing approaches, the quasi-likelihood technique is adopted, which can be more robust than the commonly adopted likelihood-based and, more importantly, more naturally “balance” the observed data and prior information for the present analysis. As to be shown in numerical studies, the proposed method has competitive performance. With the importance of interaction analysis, limitations of existing results and methods, and fast accumulation of genetic studies and data, this study can provide a practically useful new venue for G-E interaction analysis.

2 MINING PRIOR INFORMATION

As suggested in the literature, there are multiple ways of defining existing prior information/knowledge. In this study, the main interest is on the identification of important G-E interactions and corresponding main effects. As such, we extract from the existing literature information on the suggested associations between the response variable of interest and G-E interactions and main effects. Here, we propose using qualitative as opposed to quantitative information, that is, only the suggested associations but not their estimated effect sizes. This has been motivated by the following considerations. When multiple studies are independently conducted under different protocols, the direct comparability is questionable. As suggested in recent integrative studies,⁷ qualitative results can be more comparable (and more informative in this case) across studies than quantitative results. In addition, different studies may adopt different analysis methods/models, making the effect sizes not directly comparable. Some studies may only suggest associations without reporting their effect sizes. Last but not least, it is much easier to extract qualitative than quantitative information.

Consider a study with p genes (with a slight abuse of terminology, we use “genes” to generically represent G measurements) and q E variables. The following prior information is considered in our analysis. (a) The first type pertains to G-E interactions and is denoted as S_{G-E} , where a pair $(j, k) \in S_{G-E}$ indicates that the interaction between the j th gene and k th E variable has been suggested as associated with the response variable. (b) The second type of information is on the main G effects. First, consider S_{G_0} , where $j \in S_{G_0}$ indicates that the main effect of the j th gene has been suggested as associated with the response variable. This type of information is “direct.” In addition, the “main effects, interactions” hierarchy has been stressed in recent joint G-E interaction analysis. That is, if an interaction is identified, the corresponding main effect should be automatically identified. Motivated by this hierarchy, we consider S_H , where $j \in S_H$ if $(j, k) \in S_{G-E}$. This type of information is hierarchy-derived and indirect. Note that S_{G_0} and S_H may have overlap. Denote

$S_G = S_{G_0} \cup S_H$, which contains information on main G effects.

A common approach for extracting information from published studies is to mine large publication databases, especially PubMed. Multiple software tools are available for such a purpose, including VxInsight, MedMiner, UALCAN, PubMatrix, and others. In our data analysis, we adopt PubMatrix (<http://pumatrix.grc.nia.nih.gov/>), which is web-based literature mining tool and has been used in the studies of Hansen et al,¹¹ Minafra et al,¹²

Kumar et al.¹³ and many others. More specifically, PubMatrix can conduct a two-dimensional search using any two lists of keywords. Each time, it can search for 100×10 terms and present the results in a form of frequency matrix of term co-occurrence. That is, it can indicate not only whether an association exists but also its amount of evidence. The main steps in using PubMatrix to extract prior information can be summarized as follows (see the Web-based Supplementary Materials for details): (1) enter search queries (gene names and disease name) for generating S_{G_0} ; (2) download the search results that report the pairwise frequency counts found in PubMed; (3) input search queries (gene names, disease name, and E variable names) for generating S_{G-E} ; (4) output the reported results. For more detailed description about PubMatrix, we also refer to the work of Becker et al.¹⁴

To fix ideas, we consider the cutaneous melanoma example analyzed in Section 5.1. Genes are processed in batches of size 100. We search for their interactions with age, stage, gender, and clark level on melanoma, which corresponds to the interaction analysis part. In addition, we also search for their associations with melanoma directly, which corresponds to the main effect analysis part. The PubMatrix search results are presented as matrices in an html form. Partial results are shown in Figure 1. A large number of studies have suggested gene BRAF as one of the most important, if not the most important, for melanoma prognosis. In our search, its association is suggested in 4353 published studies. Its interactions with age, stage, gender, and clark level are suggested in 194, 33, 73, and 11 studies. We are able to manually examine some of the search results and find that they are sensible. For example, one search indicates that the interaction between gene IRF4 and age has been suggested in the work of Kvaskoff et al.¹⁵ A closer examination of this article reveals that it is on nevus-associated loci and melanoma and suggests that gene IRF4 is more likely to be associated with melanoma with a younger age. Although the study design of Kvaskoff et al.¹⁵ differs considerably from that of the TCGA data to be analyzed in Section 5.1 and hence meta-analysis or integrative analysis may not be suitable, this piece of information can be potentially helpful for our analysis.

3 INTERACTION IDENTIFICATION

3.1 Data

Consider a dataset with n independent and identically distributed (iid) subjects. For subject i , let Y_i be the response variable, $X_i = (X_{i1}, \dots, X_{iq})^T$ and $Z_i = (Z_{i1}, \dots, Z_{ip})^T$ be the q - and p -dimensional E and G variables, respectively. Denote $E(Y_i | X_i, Z_i) = \mu_i$. Consider the joint model

$$g(\mu_i) = \eta_i = \alpha_0 + \sum_{k=1}^q X_{ik} \alpha_k + \sum_{j=1}^p \left(Z_{ij} \beta_j + \sum_{k=1}^q Z_{ij} X_{ik} \gamma_{jk} \right) = \alpha_0 + \sum_{k=1}^q X_{ik} \alpha_k \quad (1)$$

$$+ \sum_{j=1}^p \mathbf{b}_j^T \mathbf{w}_{ij},$$

where α_0 is the intercept, $\alpha_k (k = 1, \dots, q)$ is the coefficient of the k th E variable, and $\mathbf{b}_j = (\beta_j, \gamma_{j1}, \dots, \gamma_{jq})^T$ and $\mathbf{w}_{ij} = (Z_{ij}, Z_{ij}X_{i1}, \dots, Z_{ij}X_{iq})^T$ represent all coefficients and observed

data corresponding to the j th gene, respectively. $g(\cdot)$ is the known inverse link function. Further, denote $\mathbf{a} = (a_1, \dots, a_q)^T$, $\mathbf{b} = (b_1^T, \dots, b_p^T)^T$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$, and $\boldsymbol{\gamma} = (\gamma_{11}, \dots, \gamma_{1q}, \dots, \gamma_{p1}, \dots, \gamma_{pq})^T$.

3.2 Penalized identification using quasi-likelihood

For estimation, most of the existing studies adopt likelihood-based approaches. Here, we consider the quasi-likelihood approach, which demands less distributional assumptions and hence can be more robust. We refer to the work of McCullagh and Nelder¹⁶ and Feng et al¹⁷ for discussions on the advantages of quasi-likelihood approaches under low- and high-dimensional settings. In addition, as to be shown below, quasi-likelihood can be a more suitable choice for incorporating prior information than likelihood. For the response variable Y , the quasi-likelihood is defined as $Q(Y, \mu) = \int_Y^\mu \frac{Y - \mu}{\sigma^2 V(\mu)} d\mu$, where μ is the expectation and $V(\cdot)$ is the variance function. Covariates (in this case, G-E interactions and main G and E effects) affect Y through μ . With n iid observations, denote $L(\mathbf{a}_0, \mathbf{a}, \mathbf{b}; \mathbf{Y})$ as the negative log quasi-likelihood function. With low-dimensional data, the quasi-likelihood estimate can usually be obtained using the iterated weighted least squares (IWLS) approach.

When it is desirable to accommodate prior information, a “natural” choice seems to be the Bayesian technique. However, Bayesian techniques may not be easily feasible when it is needed to simultaneously accommodate high dimensionality and conduct the identification of interactions and main effects (while respecting the “main effects, interactions” hierarchy). In this study, we adopt the penalization technique, which has been a popular choice in high-dimensional interaction analysis. For recent penalized interaction analysis, we refer to the works of Liu et al,³ Wu et al,⁴ and Shi et al,¹⁸ for examples. Advancing from such studies, we adopt quasi-likelihood and incorporate prior information.

First, ignore prior information. Consider the penalized quasi-likelihood estimation approach with penalty $P(\mathbf{b}; \lambda)$. In this article, we adopt the sparse group MCP (sgMCP) with

$$P(\mathbf{b}; \xi, \lambda) = \sum_{j=1}^p \left(\rho(\|\mathbf{b}_j\|; \sqrt{(q+1)}\lambda_1, \xi) + \sum_{k=2}^{q+1} \rho(|b_{jk}|; \lambda_2, \xi) \right). \quad (2)$$

λ_1 and λ_2 are data-dependent tuning parameters, ξ is the regularization parameter, and b_{jk} is the k th element of $\mathbf{b}_j$. $\rho(t; \lambda, \xi) = \lambda \int_0^t \left(1 - \frac{x}{\lambda \xi}\right) + dx$ is the minimax concave penalty (MCP).¹⁹

The MCP-based penalty is adopted with its superior statistical and numerical performance as established in multiple published studies. The sgMCP automatically ensures the “main effects, interactions” hierarchy. For related discussions, we refer to the work of Liu et al.³

3.3 Incorporating prior information

Advancing from the previous section and other existing methods, we now consider incorporating prior information. First, consider the scenario where prior information is fully credible. Consider the penalized estimate with objective function

$$L(\alpha_0, \mathbf{a}, \mathbf{b}; \mathbf{Y}) + P_p(\mathbf{b}; \xi, \kappa), \quad (3)$$

where

$$P_p(\mathbf{b}; \xi, \kappa) = \sum_{j \notin S_G} \rho(\|\mathbf{b}_j\|; \sqrt{(q+1)}\kappa_1, \xi) + \sum_{(j,k) \notin S_{G-E}} \rho(\|b_{jk}\|; \kappa_2, \xi). \quad (4)$$

This is a modification of sgMCP. It ensures that the effects identified in the literature (prior information) are automatically identified and only conducts selection with those not previously identified. When the sample size is small to moderate and a large number of effects are included in prior information, to stabilize estimation, a small ridge penalty can be added.

As has been noted in the literature, false positives are not uncommon. In addition, previous studies may not be fully comparable to the present one, and thus, differences in findings may be expected. As such, we adopt the following strategy that balances between the prior information and present data.

Denote the estimate of (3) as

$$\left(\hat{\alpha}_0^{\text{prior}}, \hat{\mathbf{a}}^{\text{priorT}}, \hat{\mathbf{b}}^{\text{priorT}} \right)^T, \text{ where } \hat{\mathbf{a}}^{\text{prior}} = \left(\hat{\alpha}_1^{\text{prior}}, \dots, \hat{\alpha}_q^{\text{prior}} \right)^T \text{ and } \hat{\mathbf{b}}^{\text{prior}} = \left(\hat{\mathbf{b}}_1^{\text{priorT}}, \dots, \hat{\mathbf{b}}_p^{\text{priorT}} \right)^T$$

for subject i , denote the prior-predicted response as

$$\hat{Y}_i^{\text{prior}} = g^{-1} \left(\hat{\alpha}_0^{\text{prior}} + \sum_{k=1}^q X_{ik} \hat{\alpha}_k^{\text{prior}} + \sum_{j=1}^p \hat{\mathbf{b}}_j^{\text{priorT}} \mathbf{w}_{ij} \right).$$

Further, denote $\hat{\mathbf{Y}}^{\text{prior}} = \left(\hat{Y}_1^{\text{prior}}, \dots, \hat{Y}_n^{\text{prior}} \right)^T$.

We propose the overall penalized objective function

$$(1 - \tau)L(\alpha_0, \mathbf{a}, \mathbf{b}; \mathbf{Y}) + \tau L(\alpha_0, \mathbf{a}, \mathbf{b}; \hat{\mathbf{Y}}^{\text{prior}}) + P(\mathbf{b}; \xi, \lambda). \quad (5)$$

The final estimate is defined as the optimizer of this objective function, and a nonzero estimated coefficient corresponds to an identified interaction or main effect.

Rationale. In the first two terms, we have additionally inserted $\mathbf{Y}$ and $\hat{\mathbf{Y}}^{\text{prior}}$ to indicate that they are constructed using the observed response and that predicted based on prior information. Loosely speaking, they represent lack-of-fit measures from the present data and prior information, respectively. The tuning parameter $0 \leq \tau \leq 1$ describes the balance between the two lack-of-fit measures and will be chosen data dependently. Intuitively, $\tau \rightarrow$

1 corresponds to more reliable prior information, and $\tau \rightarrow 0$ if otherwise. Under the proposed strategy, quasi-likelihood can be more convenient than likelihood. Consider, for example, a binary response. The prior-predicted value would be a probability between 0 and 1, as opposed to 0 or 1. A likelihood function cannot be directly constructed. In contrast, the quasi-likelihood approach only involves an estimating equation where the numerator has the form $Y - \mu$. It still functions well when Y does not take value of 0 or 1. The penalty P is imposed for regularized estimation and selection and has similar interpretations as previously discussed.

The proposed strategy shares a similar spirit with that in the work of Jiang et al.¹⁰ Additional flexibility is introduced with the quasi-likelihood, which makes the proposed approach more broadly applicable. Additional complexity is also brought by interactions beyond main effects. It is noted that the proposed analysis only uses the qualitative information of whether a main G effect or interaction has been previously suggested. Using more quantitative information (for example, the counts of previous studies) may lead to the analysis dominated by a few “simple” signals, as can be partly seen from Figure 1.

3.4 Computation

In analysis that involves quasi-likelihood, the most common computational approach is the IWLS, which transforms the optimization problem into that of a weighted least squares. As such, consider the penalized weighted least squares problem

$$\frac{1}{2n} \sum_{i=1}^n \omega_i \left(Y_i - \alpha_0 - \sum_{k=1}^q X_{ik} \alpha_k - \sum_{j=1}^p \mathbf{b}_j^T \mathbf{W}_{ij} \right)^2 + P(\mathbf{b}; \xi, \lambda). \quad (6)$$

Here, ω_i 's are the weights that need to be updated in each iteration. With simple manipulations ω_i 's can be multiplied inside, and α_0 can be eliminated. With a slight abuse of notations, consider the simplified objective function

$$\frac{1}{2n} \left\| \mathbf{Y} - \mathbf{X}\mathbf{a} - \mathbf{W}\mathbf{b} \right\|_2^2 + P(\mathbf{b}; \xi, \lambda), \quad (7)$$

where $\mathbf{X}$ and $\mathbf{W}$ are the matrices composed of X_{ik} and $\mathbf{W}_{ij}$, respectively.

We adopt an iterative algorithm. With the estimate of $\mathbf{b}$ fixed at $\hat{\mathbf{b}}$, we estimate $\mathbf{a}$ as $\hat{\mathbf{a}} = (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T (\mathbf{Y} - \mathbf{W}\hat{\mathbf{b}}))$. Denote $\mathbf{W}_j$ as the submatrix of $\mathbf{W}$ corresponding to $\mathbf{b}_j$. Then, with the estimate of $\mathbf{a}$ fixed, we first orthogonalize $\mathbf{W}_j$ such that $\frac{1}{n} \mathbf{W}_j^T \mathbf{W}_j = \mathbf{I}_{q+1}$. This can be achieved with the Cholesky decomposition technique. Then, we adopt a group coordinate descent (GCD) technique. The GCD optimizes the objective function with respect to one group of coefficients, which corresponds to the main and interaction effects of one gene, at a time, and iteratively cycles through all genes. Details of the GCD are presented in the Supplementary Materials. With fixed tunings, the main steps of the overall computational

algorithm are as follows. (a) Initialize $\hat{a} = 0$ and $\hat{b}$. (b) With the current estimates of $\mathbf{a}$ and $\mathbf{b}$, compute ω_i 's. (c) Update the estimate of $\mathbf{a}$ as described above. (d) Update the estimate of $\mathbf{b}$ as described in the Supplementary Materials. (e) Repeat Steps (b)-(d) until convergence. In numerical study, we conclude convergence when the relative change between two consecutive estimates is smaller than 10^{-4} . Convergence is achieved in all of our numerical studies, usually within 40 iterations. It is noted that the loss function of the proposed psgMCP is slightly different from that of sgMCP. We write $\mathbf{Y}^* = (1 - \tau) \mathbf{Y} + \tau \hat{\mathbf{Y}}^{\text{prior}}$. Then, (5) can be rewritten as $L(\mathbf{a}_0, \mathbf{a}, \mathbf{b}; \mathbf{Y}^*) + P(\mathbf{b}; \xi, \lambda)$. This also shows the advantage of the quasi-likelihood approach. With data $(\mathbf{X}, \mathbf{W}, \mathbf{Y}^*)$, the computation for psgMCP can be achieved using the above GCD algorithm.

The proposed method involves tuning parameters $(\tau, \lambda_1, \lambda_2)$. In addition, the MCP involves the regularization parameter ξ . In our numerical study, we fix $\xi = 6$ as in published studies. For the determination of τ, λ_1 , and λ_2 , we conduct a three-dimensional grid search in both simulation and real data analysis. A similar search procedure has been adopted in the literature, for example, in the work of Tan et al.,²⁰ where three tuning parameters are selected using a Bayesian information criterion (BIC)-type criterion. To select the optimal tuning combination, we adopt the extended BIC (EBIC),²¹ which has demonstrated satisfactory performance in high-dimensional studies. For the alternative approaches, EBIC is also used. All computations are implemented using R. The code is publicly available at <https://github.com/Xu-Yonghong>. As the computational algorithm described above only involves simple calculations, the overall computational cost is affordable. For example, for a simulated dataset with four E variables and 1000 genes, analysis can be accomplished within ten minutes using a regular laptop.

4 SIMULATION

Simulation is conducted to gauge practical performance of the proposed approach. The most significant component of the proposed approach is the incorporation of prior information, for which we consider three settings. Under S_1 , about 80% of the prior information is true, that is, 80% of the effects suggested by prior information are actually associated with the response in the present data. The rest 20% of the prior information is false. Under S_2 , about 50% of the prior information is true. Under S_3 , only about 30% of the prior information is true. For a large number of diseases, these three settings may well describe the different levels of available information. In addition, to test the “robustness” of the proposed approach, for some simulation scenarios, we also consider 20% (S_4) and 10% (S_5) of the true information.

For the G variables, we generate from a multivariate normal distribution with marginal mean zero and covariance matrix $\Sigma = (\rho_{ij})_{p \times p}$ which has been commonly adopted in the literature and may mimic gene expression data analyzed in the next section. Four E variables are simulated. To be comprehensive, the following four scenarios are considered. Under Scenario 1, the G variables have an autoregressive (AR) correlation structure with $\rho_{ij} = \rho^{|i-j|}$ and $\rho = 0.3$. The four E variables are generated from a multivariate normal distribution with an AR correlation structure and $\rho = 0.5$. Under Scenario 2, the E variables are the same as under Scenario 1. The G variables have a banded correlation (BC) structure, where $\rho_{ij} = I_{(i=j)}$

$+ 0.33I_{(j=1)}$. Under Scenario 3, two categorical E variables are generated by dichotomizing the continuous ones generated above, and the G variables are the same as under Scenario 1. Under Scenario 4, there are also two categorical E variables, and the G variables have a BC structure.

To evaluate the broad applicability of the proposed method, both continuous and count responses are considered, each with three simulation settings. Specifically, under Simulations 1 to 3, continuous responses are generated from linear regression models; and under Simulations 4 to 6, count responses are generated from Poisson models. In the quasi-likelihood estimation, the commonly adopted variance functions for continuous and count data are used. The details are as follows.

Simulation 1. $n = 200$ and $p = 1000$. There are four main E effects, eight main G effects, and six G-E interaction effects. The nonzero coefficients are randomly generated from $\text{Uniform}(0.2, 0.5)$.

Simulation 2. There are 16 main G effects and 10 G-E interactions. This setting can test if the proposed approach can “scale up.”

Simulation 3. Same as Simulation 1, except that $p = 1500$. This can test performance of the proposed approach when there are more noises.

Simulation 4. $n = 400$ and $p = 1000$. There are eight main G effects and seven G-E interactions. The nonzero coefficients of E effects are generated from $\text{Uniform}(0.5, 0.8)$. The nonzero coefficients of G effects and G-E interactions are also generated from $\text{Uniform}(0.5, 0.8)$, and their signs are positive (negative) with a probability of 0.5.

Simulation 5. We increase the numbers of main G effects and G-E interactions to 16 and 12, respectively. The nonzero coefficients are generated in a similar way as Simulation 4 but from $\text{Uniform}(0.3, 0.5)$.

Simulation 6. Same as Simulation 5, except with $p = 1500$.

For comparison, we consider two closely related alternatives. The first is sgMCP, which analyzes the present simulated data without considering prior information. The second is Prior, which fully trusts prior information and is defined in the first part of Section 3.3. We note that there are multiple other approaches that can analyze the simulated data. The above two have analysis frameworks closest to the proposed and can directly test the benefit of balancing between prior information and present data. The three approaches are compared using the following criteria, and with the differences between main G effects and interactions, separate evaluations are conducted. (1) TP and FP, which are the numbers of true and false positives identified; (2) Bias, which is defined as $\frac{\|\hat{\beta} - \beta\|_1}{\|\beta\|_1}$ and $\frac{\|\hat{\gamma} - \gamma\|_1}{\|\gamma\|_1}$ for the main G effects and interactions, respectively; (3) PMSE, which is the prediction mean squared error (MSE) evaluated on an independently generated testing dataset that has the same data generating mechanism as the training data, and defined as $E[(Z\beta - Z\hat{\beta})^T(Z\beta - Z\hat{\beta})]$

and $E\left[\sum_{j=1}^p \sum_{k=1}^q \left(Z_j X_k (\gamma_{jk} - \hat{\gamma}_{jk})\right)^2\right]$ for the main G and interaction effects, respectively.

Summary statistics are computed based on 200 replicates.

Results for Simulation 1 are shown in Table 1. The rest of the simulation results are in the Supplementary Materials. Across the whole spectrum of simulation, the proposed approach is observed to have competitive performance. Consider, for example, Scenario 3 in Table 1. For the identification of interactions, when the prior information is largely accurate (S_1), the three approaches have the same performance in terms of TP (5.90). In terms of FP, Prior has the best performance (3.37), and the proposed approach (5.67) significantly outperforms sgMCP (16.40). For Bias and PMSE, the proposed approach and Prior have comparable performance and are better than sgMCP. When the prior information is largely inaccurate (S_3), the proposed approach (5.70) is inferior to sgMCP but outperforms Prior (5.50). It is noted that it significantly outperforms sgMCP with a much smaller FP. In terms of Bias and PMSE, the proposed approach outperforms both alternatives. Similar patterns are observed for the main G effects. For example, sgMCP has TP 7.47; under S_1 , the proposed approach has TP 7.60, and Prior has TP 7.87. Under S_3 , the proposed approach has TP 7.07, compared to 6.27 of Prior.

Under Simulation 2 (Table 1 of the Supplementary Materials), which has more main effects and interactions, all methods identify more TPs and FPs (that is, can scale up). psgMCP still outperforms sgMCP when the quality of prior information is high (S_1). When the prior information is of low quality (S_3), the proposed approach may adaptively accommodate false information, whereas Prior identifies more false effects. Under Simulation 3 with more noisy variables (with results shown in Table 2 of the Supplementary Materials), all methods have increased FP, Bias, and PMSE values. However, the relative superiority of the proposed approach remains. For count responses, patterns observed in Tables 3 to 5 of the Supplementary Materials are comparable, although there are numerical differences.

We also take a closer look at the values of the tuning parameter τ . Consider, for example, Simulation 3 under Scenario 1. The median values of τ are 0.55 (S_1), 0.22 (S_2), and 0.11 (S_3), respectively. That is, when the quality of prior information gets worse, the proposed approach is able to adaptively detect that and puts a less weight on the prior information. This demonstrates the adaptiveness of the proposed approach, which is highly desirable, as in practice, the quality of prior information is unknown.

5 ANALYSIS OF TCGA DATA

For a large number of cancer types, G-E interactions have independent contributions beyond the main G and E effects. In this section, we analyze two cancer datasets from TCGA (<http://tcga-data.nci.nih.gov/tcga/>), which is a recent effort organized by NCI and has recently published high quality data on multiple cancers. The first dataset is on skin cutaneous melanoma (referred to as “SKCM”). The second dataset is on glioblastoma multiforme (referred to as “GBM”). The practical importance of the two cancers has been well established in the literature and will not be reiterated here. Processed level III data are downloaded using the R package TCGA2STAT. For G measurements, we choose gene

expressions, which have been analyzed in many recent studies. For E measurements, we take a loose definition and also include clinical variables, also as in recent studies.

5.1 Skin cutaneous melanoma data

For each sample, expression measurements are available for 18 351 genes. To reduce computational cost and improve the stability of analysis, we conduct a supervised marginal screening and select 1350 gene expressions for downstream analysis. Four E variables are included in analysis, namely, age, AJCC TUMOR PATHOLOGIC PT (referred to as “stage”), gender, and CLARK LEVEL AT DIAGNOSIS (referred to as “clark level”). Samples with missing E measurements are removed from analysis, leading to an effective sample size of 294. The response variable is the overall survival (analyzed with a logarithm transformation), which ranges from 0.01 to 29 years, with a median of 2.85. It is treated as a continuous variable, and in the quasi-likelihood estimation, the default variance function is adopted. Additional considerations are needed to accommodate censoring. We adopt the Kaplan-Meier weighted approach,²² which adds a nonzero weight to each event and a zero weight to each censored subject. This approach can be easily coupled with the quasi-likelihood estimation and causes minimum changes to computation.

The results for mining prior information are shown in Figures 1 and 2. For the main G effects, a total of 388 genes (out of 1350) have been previously suggested as associated with cutaneous melanoma. Among them, 29 have been mentioned in more than 50 publications. The highest counts (most prior information) correspond to genes BRAF, HR, NRAS, and CASP3, all of which have been extensively examined and confirmed as melanoma markers. This may provide a partial support to the validity of the prior mining procedure. For the purpose of reliability, we use these top 29 genes to construct S_{G_0} . Next, we mine G-E

interactions. Comparatively, research on G-E interactions is much limited. In addition, 148 G-E interactions have been suggested in the literature. Among them, 38 have been suggested more than 5 times, which are used to construct S_{G-E} . Figure 2 suggests that most of the G-E interactions identified in the literature are with age, which also has sound biological basis. Among the genes, HR has the most interactions.

With the above prior information, the proposed approach identifies 33 main G effects and 53 G-E interactions. The detailed estimation results are shown in Table 2. All four E variables are identified as having interactions with genes, with the most interactions for age and stage. Among the 33 identified genes, 20 are included in prior information, suggesting a reasonable degree of alignment with the existing literature. On the other hand, new findings are also made. A quick literature search suggests that the findings may be biologically sensible. For example, ASL plays an important role for the catalyzation of the arginine, which is essential for melanoma growth. A decreased expression of BAX is associated with a poorer survival rate and tumor progression in cutaneous melanoma. GSTP 1 is a major GST isoenzyme expressed in the melanocytes of the normal skin basal layers, as well as in cutaneous melanoma. IRF4 is a gene related to melanocytic nevus count (MNC), and a higher MNC may indicate a higher risk of cutaneous melanoma. IRF4 has also been suggested as having a strong genotype-by-age interaction effect on melanoma. KY is involved in the function of neuromuscular junction, and studies have suggested that neurofibromatosis is associated

with melanocytic malignancy. NRAS, providing proliferation signaling from surface receptors to nucleus, is correlated with melanocytic nevus syndrome and exists in most melanomas of skin primary. SF3B1 mutations at codon 625 occur in cutaneous melanoma although with a low frequency. SOX10 is important for Nestin activation in melanoma cell lines, and a high expression of SOX10 has been observed in melanoma.

Data are also analyzed using the two alternatives. Detailed information on identification and estimation using the alternatives is available in the Supplementary Materials. Summary comparison results are provided in Table 4. It is observed that, in terms of main G effects and G-E interactions identified, different approaches have moderate overlaps. For example, sgMCP and Prior identify 51 and 44 interactions, respectively, with 35 and 22 overlaps with the proposed approach. It has been recognized that different genes can have similar functions/expression values. As such, we compute the modified random variable (RV) coefficients (MRVC),²³ which can evaluate the overlap of information of two gene sets. The MRVC analysis suggests a higher level of overlap. However, differences among the three approaches are still prominent. To further compare the three approaches, a cross-validation-based approach is applied, and the median absolute prediction error (MAPE) is computed. The proposed approach has a small improvement in prediction. We further evaluate the similarity in the estimation results of the three methods. Specifically, we compute the MSEs and correlation coefficients (COR) for the overlapping effects. Table 7 (Supplementary Materials) suggests that the proposed approach and sgMCP generate similar estimates, which differ from those of Prior. The stability of finding is also evaluated. Specifically, the observed occurrence index (OOI), which evaluates the probability of a finding being made, is computed. It is noted that the OOI evaluation is not conducted for Prior, which forces prior findings in the model and may have superficially higher OOIs. For the proposed approach and sgMCP, the mean OOI values are 0.62 and 0.50, respectively. The prediction and stability evaluation provide some support to the proposed approach.

5.2 Glioblastoma multiforme data

In this analysis, the response variable is also survival. Samples without response information are removed. In addition, samples with a Karnofsky performance score (KPS) less than 60 are also eliminated since they might have died for reasons other than the disease itself.²⁴ A total of 300 samples are available for downstream analysis. For each sample, measurements on 17 814 gene expressions are available. A supervised prescreening is conducted, leading to 1314 genes for further analysis. The E variables considered include age, gender, KPS, and race.

The results for mining prior information are shown in Figures 5 and 6 of the Supplementary Materials. For the main G effects, a total of 369 genes have been previously suggested as associated with GBM. However, it is observed that the number of studies on GBM is considerably smaller than that on melanoma. Specifically, 323 (out of 369) genes appear in less than 10 studies; and only 20 have been suggested in 30 or more studies. For reliability, those 20 genes are used to construct S_{G_0} . Mining of G-E interactions is also conducted (Figure 6 of the Supplementary Materials). A total of 17 interactions are used to construct S_{G-E} .

For this dataset, survival information is presented in a “discrete” form (in days). To test the applicability of the proposed approach, we treat the response variable as a count in the quasi-likelihood estimation and adopt the default mean and variance functions. Weights are again imposed to accommodate censoring. With the proposed approach, a total of 50 main G effects and 46 G-E interactions are identified. Detailed results are shown in Table 3. Most of the identified interactions are with age and gender. Among the 50 identified genes, 20 have been suggested in the literature. A quick literature search suggests that the findings can be biologically sensible. For example, studies have suggested a significant association between the ANGPT1/ANGPT2 balance and GBM survival. CASP9 has a critical cancer-related function and may be related to the epigenetic deregulation of the mitochondria-independent apoptosis in recurrent GBM. An early study has also suggested a potential link between the CD109 expression in CECs with GBM survival. CUZD1 is mapped at chromosome 10q26.13, and the loss of 10q is common in the development and progression of GBM, which suggests a potential association between CUZD1 and GBM. A decreasing expression of ABCC1 is associated with the mechanism by which FK506 sensitizes GBM cells.

We also analyze data using the two alternatives. Summary comparison results are presented in Table 4. Detailed estimation results using the alternatives are available in the Supplementary Materials. Table 4 suggests that, in terms of the sets of identified main G effects and interactions, the three approaches differ significantly. Much smaller discrepancies are observed if measured using the MRV coefficients. Prediction evaluation is also conducted, and the proposed approach has performance slightly better than sgMCP but much better than Prior. When examining the estimates of the overlapping genes, Table 7 (Supplementary Materials) again suggests a higher degree of similarity between the proposed approach and sgMCP, which are considerably different from Prior. In the evaluation of stability, the proposed approach has a mean OOI of 0.57, better than sgMCP (a mean of 0.45).

6 DISCUSSION

With the goal of improving the identification of G-E interactions (and main effects), in this study, we have developed a new method that can incorporate prior information in identification and estimation. The strategy of creating a balance between the present observed data and prior information in goodness-of-fit is intuitive. The adoption of quasi-likelihood, which has an estimating equation form, makes the proposed approach more robust (by making fewer assumptions) and is more natural with the predicted responses. The adopted penalization can respect the “main effects, interactions” hierarchy and can be feasibly realized. Simulation shows that the proposed approach has competitive performance. It is noted that when the quality and amount of prior information are high, the Prior approach can be favorable; In contrast, with low quality prior information, the analysis of present data can be preferred. However, in practical data analysis, *the quality and amount of prior information is unknown*. Simulation and data analysis suggest that the proposed approach can data-dependently choose τ and find the proper balance between the two goodness-of-fit functions and can outperform the approach when a wrong decision on prior information is made.

This study can be potentially extended in multiple directions. The quasi-likelihood can be potentially replaced with other estimating equations. There are multiple ways of generating prior information. The adopted mining PubMed has been popular in the literature. As discussed above, to avoid analysis being dominated by a small number of easy targets, we have chosen to use qualitative as opposed to quantitative prior information. We conjecture that adaption through weights may be possible to accommodate quantitative prior information—we will leave this to future study. In this article, we have focused on methodological development and numerical studies. It will be of interest to investigate statistical properties in future research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We thank the editor and reviewers for careful review and insightful comments, which have led to a significant improvement of this article. The study of Wang was supported by the National Natural Science Foundation of China (71601076), Humanities and Social Sciences Youth Foundation of Ministry of Education of China (16YJCZH104), and Social Science Foundation of Hunan Province (15YBA085). The study of Xu was supported by the National Social Science Foundation of China (17CTJ007). The study of Ma was supported by the National Institutes of Health (CA204120, CA191383, and CA121974).

Funding information

National Natural Science Foundation of China, Grant/Award Number: 71601076; Humanities and Social Sciences Youth Foundation of Ministry of Education of China, Grant/Award Number: 16YJCZH104; Social Science Foundation of Hunan Province, Grant/Award Number: 15YBA085; National Social Science Foundation of China, Grant/Award Number: 17CTJ007; National Institutes of Health, Grant/Award Number: CA204120, CA191383, and CA121974

REFERENCES

1. Witten DM, Tibshirani R. Survival analysis with high-dimensional covariates. *Stat Methods Med Res.* 2010;19(1):29–51. [PubMed: 19654171]
2. Bien J, Taylor J, Tibshirani R. A lasso for hierarchical interactions. *Ann Stat.* 2013;41(3):1111–1141. [PubMed: 26257447]
3. Liu J, Huang J, Zhang Y, et al. Identification of gene–environment interactions in cancer studies using penalization. *Genomics.* 2013;102(4):189–194. [PubMed: 23994599]
4. Wu M, Zang Y, Zhang S, Huang J, Ma S. Accommodating missingness in environmental measurements in gene–environment interaction analysis. *Genet Epidemiol.* 2017;41(6):523–554. [PubMed: 28657194]
5. Zeggini E, Scott LJ, Saxena R, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet.* 2008;40(5):638–645. [PubMed: 18372903]
6. Ma S, Huang J, Wei F, Xie Y, Fang K. Integrative analysis of multiple cancer prognosis studies with gene expression measurements. *Statist Med.* 2011;30(28):3361–3371.
7. Zhao Q, Shi X, Huang J, Liu J, Li Y, Ma S. Integrative analysis of ‘-omics’ data using penalty functions. *WIREs Comput Stat.* 2015;7(1):99–108.
8. Lobach I, Fan R, Manga P. Genotype-based association models of complex diseases to detect gene–gene and gene–environment interactions. *Stat Its Interface.* 2014;7(1):51–60.
9. Merlino G, Noonan FP. Modeling gene–environment interactions in malignant melanoma. *Trends Mol Med.* 2003;9(3):102–108. [PubMed: 12657431]

10. Jiang Y, He Y, Zhang H. Variable selection with prior information for generalized linear models via the prior lasso method. *J Am Stat Assoc.* 2016;111(513):35–376.
11. Hansen MS, Tencerova M, Frølich J, Kassem M, Frost M. Effects of gastric inhibitory polypeptide, glucagon-like peptide-1 and glucagon-like peptide-1 receptor agonists on bone cell metabolism. *Basic Clin Pharmacol Toxicol.* 2018;122(1):25–37. [PubMed: 28722834]
12. Minafra L, Bravatà V, Cammarata FP, Russo G, Gilardi MC, Forte GI. Radiation gene-expression signatures in primary breast cancer cells. *Anticancer Res.* 2018;38(5):2707–2715. [PubMed: 29715090]
13. Kumar A, Thakur P, Gupta K, Pal A. Text mining approach to analyse the relation between obesity and breast cancer data. *Int Lett Nat Sci.* 2015;44:1–9.
14. Becker KG, Hosack DA, Dennis G, et al. PubMatrix: a tool for multiplex literature mining. *BMC Bioinform.* 2003;4(1):1–6.
15. Kvaskoff M, Whiteman DC, Zhao ZZ, et al. Polymorphisms in nevus-associated genes MTAP, PLA2G6, and IRF4 and the risk of invasive cutaneous melanoma. *Twin Res Hum Genet.* 2011;14(5):422–432. [PubMed: 21962134]
16. McCullagh P, Nelder JA. *Generalized Linear Models.* Boca Raton, FL: CRC Press; 1989.
17. Feng W, Sarkar A, Lim CY, Maiti T. Variable selection for binary spatial regression: penalized quasi-likelihood approach. *Biometrics.* 2016;72(4):1164–1172. [PubMed: 27061299]
18. Shi X, Liu J, Huang J, Zhou Y, Xie Y, Ma S. A penalized robust method for identifying gene–environment interactions. *Genet Epidemiol.* 2014;38(3):220–230. [PubMed: 24616063]
19. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat.* 2010;38(2):894–942.
20. Tan KM, London P, Mohan K, Lee SI, Fazel M, Witten D. Learning graphical models with hubs. *J Mach Learn Res.* 2014;15(1):3297–3331. [PubMed: 25620891]
21. Chen J, Chen Z. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika.* 2008;95(3):759–771.
22. Stute W. Distributional convergence under random censorship when covariables are present. *Scand J Stat.* 1996;23(4):461–471.
23. Smilde AK, Kiers HAL, Bijlsma S, Rubingh CM, van Erk MJ. Matrix correlations for high-dimensional data: the modified RV-coefficient. *Bioinformatics.* 2008;25(3):401–405. [PubMed: 19073588]
24. Srinivasan S, Patric IRP, Somasundaram K. A ten-microRNA expression signature predicts survival in glioblastoma. *PloS One.* 2011;6(3):e17438. [PubMed: 21483847]

FIGURE 1.

Analysis of skin cutaneous melanoma (SKCM) data. Left: numbers of publications that suggest the main effects of genes (for the 29 genes with counts at least 50). Right: boxplot of counts for genes with counts below 50 [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 2.

Analysis of skin cutaneous melanoma (SKCM) data. Numbers of publications that suggest gene-environment interactions. A, Gene-age interaction; B, Gene-stage interaction; C, Gene-gender interaction; D, Gene-Clark-level interaction [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1

Simulation 1: mean (sd) in each cell

sgMCP			S_1		S_2		S_3	
		Prior	psgMCP	Prior	psgMCP	Prior	psgMCP	
Scenario 1	Main Effect							
	TP	7.12(0.74)	7.77(0.43)	7.70(0.53)	6.30(0.84)	7.53(0.68)	5.75(1.75)	7.38(0.74)
	FP	5.00(1.88)	2.13(0.35)	1.43(2.42)	5.30(0.79)	3.50(2.33)	6.12(0.35)	4.25(2.92)
	Bias	0.41(0.09)	0.23(0.05)	0.35(0.09)	0.36(0.16)	0.37(0.21)	0.46(0.13)	0.39(0.09)
	PMSE	0.26(0.08)	0.08(0.03)	0.11(0.04)	0.30(0.30)	0.24(0.22)	0.30(0.23)	0.23(0.07)
	Interaction Effect							
	TP	5.75(0.46)	6.00(1.29)	5.97(0.18)	5.00(1.39)	5.90(0.55)	4.88(0.35)	6.00(0.46)
	FP	13.07(4.61)	3.03(0.18)	4.13(5.91)	2.47(1.17)	6.01(4.23)	4.12(0.35)	5.00(0.32)
	Bias	0.45(0.09)	0.31(0.11)	0.38(0.11)	0.33(0.15)	0.39(0.12)	0.45(0.12)	0.39(0.12)
	PMSE	0.21(0.12)	0.18(0.23)	0.09(0.08)	0.20(0.11)	0.09(0.06)	0.20(0.12)	0.11(0.05)
Scenario 2	Main Effect							
	TP	7.17(0.78)	7.84(0.48)	7.47(0.72)	6.30(0.75)	7.50(0.76)	6.00(1.28)	7.01(0.80)
	FP	6.33(1.05)	2.07(0.25)	2.67(3.68)	5.30(0.61)	3.33(3.93)	6.73(1.03)	3.67(3.20)
	Bias	0.49(0.15)	0.32(0.12)	0.43(0.14)	0.50(0.23)	0.43(0.14)	0.57(0.13)	0.41(0.08)
	PMSE	0.26(0.11)	0.10(0.03)	0.16(0.06)	0.19(0.14)	0.18(0.13)	0.23(0.08)	0.19(0.08)
	Interaction Effect							
	TP	5.83(0.38)	5.93(0.11)	5.87(0.35)	5.00(1.14)	5.83(0.38)	6.00(0.65)	5.83(0.38)
	FP	10.53(6.96)	3.00(0.00)	3.50(9.61)	2.00(0.94)	6.00(5.08)	4.00(1.40)	5.97(5.59)
	Bias	0.52(0.14)	0.54(0.28)	0.47(0.15)	0.41(0.13)	0.40(0.10)	0.57(0.16)	0.49(0.10)
	PMSE	0.17(0.06)	0.06(0.26)	0.13(0.11)	0.11(0.06)	0.09(0.06)	0.12(0.07)	0.10(0.06)
Scenario 3	Main Effect							
	TP	7.47(0.63)	7.87(0.35)	7.60(0.57)	6.33(0.71)	7.37(0.61)	6.27(0.99)	7.07(0.87)
	FP	8.83(3.89)	2.43(0.86)	2.23(3.44)	5.63(0.76)	4.50(4.24)	6.77(0.94)	2.50(4.12)
	Bias	0.41(0.09)	0.23(0.05)	0.35(0.09)	0.48(0.18)	0.40(0.10)	0.58(0.16)	0.43(0.11)
	PMSE	0.44(0.10)	0.26(0.10)	0.35(0.09)	0.21(0.18)	0.17(0.16)	0.28(0.14)	0.23(0.13)
	Interaction Effect							
	TP	5.90(0.40)	5.90(1.04)	5.90(0.40)	4.93(1.28)	5.83(0.59)	5.50(0.63)	5.70(0.75)
	FP	16.40(7.42)	3.37(1.07)	5.67(7.15)	2.00(1.87)	1.12(0.10)	4.50(1.98)	4.50(1.72)
	Bias	0.62(0.13)	0.43(0.19)	0.49(0.12)	0.49(0.17)	0.54(0.12)	0.66(0.20)	0.53(0.10)
	PMSE	0.14(0.10)	0.06(0.14)	0.06(0.14)	0.12(0.16)	0.10(0.16)	0.14(0.06)	0.11(0.06)
Scenario 4	Main Effect							
	TP	6.90(0.84)	8.00(0.00)	7.30(0.71)	6.70(0.47)	6.94(0.84)	6.33(0.71)	7.37(0.61)
	FP	9.13(2.97)	2.33(0.84)	2.07(2.21)	5.73(1.20)	4.00(4.04)	5.63(0.76)	4.50(4.38)
	Bias	0.50(0.09)	0.23(0.08)	0.36(0.10)	0.54(0.17)	0.45(0.14)	0.61(0.16)	0.46(0.10)
	PMSE	0.49(0.17)	0.10(0.07)	0.22(0.12)	0.26(0.11)	0.31(0.16)	0.28(0.13)	0.28(0.13)
	Interaction Effect							
	TP	5.27(0.78)	5.60(0.56)	5.57(0.68)	5.20(0.76)	5.53(0.57)	5.00(1.28)	6.00(0.59)
	FP	17.50(7.01)	3.13(0.43)	6.60(7.91)	2.37(1.13)	7.50(10.93)	3.00(1.87)	10.00(11.27)

		sgMCP		S_1		S_2		S_3	
				Prior	psgMCP	Prior	psgMCP	Prior	psgMCP
Bias	0.63(0.11)	0.38(0.20)	0.48(0.12)	0.43(0.19)	0.55(0.18)	0.68(0.25)	0.58(0.14)		
PMSE	0.36(0.14)	0.09(0.05)	0.18(0.13)	0.14(0.09)	0.24(0.17)	0.22(0.22)	0.27(0.18)		

Abbreviations: FP, false positive; PMSE, prediction mean squared error; TP, true positive.

TABLE 2

Analysis of skin cutaneous melanoma (SKCM) data using the proposed method: identified main effects and interactions

	MainG	Age	Stage	Gender	Clark Level
Main E		−0.11	−0.64	0.13	−1.22
AMN *	0.04	−0.01			
ASL *	0.18			−0.43	
BAX *	0.10	0.36			
FTLP10 *	0.13				
GSTP1 *	−0.57	0.26	−0.45	−0.19	−0.02
HCAR3 *	−0.02	0.81		0.38	
HR *	0.80	−0.44	1.59	−0.72	0.30
IRF4 *	−0.23	−0.09	0.40		
KY *	0.04	−0.46	0.49	0.50	0.74
MMP10 *	0.35				
NRAS *	−0.40		−0.57		−0.18
PIK3CA *	0.39				
POLK *	0.32	0.22			
PPP6R3 *	0.46				
SON *	0.80	−0.37			
SOX10 *	−0.26	0.39	−0.35	−0.99	
SF3B1 *	−0.08				
HCAR2 *	−0.24	−0.62	1.01	−0.69	
PAK1 *	−0.17				
PML *	−0.25		−0.85	−0.24	
CFAP126	−0.17	0.08	0.21		−0.54
CGB2	0.11				
CNPPD1	0.72	0.01			
COMMD4	−0.76		1.79	−1.51	0.75
DPP3	−0.25	0.43			
HGS	−0.15		−0.80	0.59	0.13
NSUN5	0.10				
PHKB	−0.22				
PITPNA	0.23	−0.35	0.35	0.83	
PMS2P4	0.17	0.17	−1.75	0.95	
SERP2	−0.53		1.07		
SMIM21	−0.41	0.54	−0.97		
TAS2R1	0.16	0.70	−0.15		

* Genes that not only included in the prior information but also identified by the Prior method.

TABLE 3

Analysis of glioblastoma multiforme (GBM) data using the proposed method: identified main effects and interactions

	MainG	Age	Gender	KPS	Race
Main E		−0.22	0.17	0.01	−0.41
ACOT9 *	−0.26	0.05			
BCL2L1 *	−0.06	0.06	−0.05		0.07
C13orf33 *	0.11				
CTDSPL *	−0.10	−0.29	0.06		
ABCA3 *	−0.05	0.11	0.02	0.08	
ABCC1 *	0.13	0.15	0.02		
AIFM2 *	−0.02	0.10	0.02		
ANGPT1 *	−0.01		0.01		
ARF1 *	0.32				
ARPC1B *	−0.25				
ATG12 *	−0.04		0.01		−0.01
CASP9 *	0.06				
CD109 *	0.09				
DLL3 *	0.22		−0.04		0.02
FGF13 *	0.19	−0.10			
HLA-B *	−0.19		−0.08		
LAMA4 *	0.32				
PRNP *	−0.17		−0.11		
SKAP2 *	−0.16	−0.15	0.21		
SLPI *	−0.17	−0.09			
ABCA5	0.26		−0.03		−0.02
ACADS	−0.50	0.20	−0.01		
ACTR1A	0.06				−0.16
AIM1	0.11				
AMAC1	−0.24				
ARMCX3	−0.01				
BEST3	−0.12				
BST1	0.46	−0.57			
C10orf4	−0.05		0.01		
C1GALT1C1	0.29				
C1orf109	0.12	−0.02	0.04		
C1orf75	−0.07				
C5orf30	−0.26	0.17			
C9orf142	−0.03				0.20

	MainG	Age	Gender	KPS	Race
CDH20	−0.03		0.01		
CHCHD7	−0.07	−0.24			
CUZD1	−0.11		0.02		
EID3	0.29				
FAM57B	−0.33	0.03			
FGD6	0.08	0.02	−0.01		
FLRT2	0.02	0.10			
FMR1NB	0.24				
HFM1	0.17		−0.01		
LOC493869	0.13				
PCDHGB3	0.13				
PCSK1	0.04	−0.01			
PTP4A2	0.06	0.14			
RP13–102H20.1	0.12	−0.05			
SCGN	−0.25				
TMEM132D	0.14				

* Genes that not only included in the prior information but also identified by the Prior method. Abbreviation: KPS, Karnofsky performance score.

TABLE 4

Data analysis: comparison of the three methods

	MainG			Interaction			MRVC		MAPE
	sgMCP	Prior	psgMCP	sgMCP	Prior	psgMCP	Prior	psgMCP	
SKCM									
sgMCP	49	15	27	51	17	35	0.72	0.86	5.01
Prior		41	17		44	22		0.82	4.82
psgMCP			33			53			4.73
GBM									
sgMCP	44	4	35	44	3	36	0.85	0.95	28
Prior		45	3		44	3		0.88	38
psgMCP			50			46			24

Abbreviations: GBM, glioblastoma multiforme; MAPE, median absolute prediction error; MRVC, modified random variable coefficients; SKCM, skin cutaneous melanoma.