Spring 2021

Statistical Methods for Gene-Environment Interactions

Yaqing Xu
Yale University Graduate School of Arts and Sciences, xuyaqing0919@gmail.com

Abstract

Statistical Methods for Gene-Environment Interactions

Yaqing Xu

2021

Despite significant main effects of genetic and environmental risk factors have been found, the interactions between them can play critical roles and demonstrate important implications in medical genetics and epidemiology. Although many important gene-environment (G-E) interactions have been identified, the existing findings are still insufficient and there exists a strong need to develop statistical methods for analyzing G-E interactions. In this dissertation, we propose four statistical methodologies and computational algorithms for detecting G-E interactions and one application to imaging data. Extensive simulation studies are conducted in comparison with multiple advanced alternatives. In the analyses of The Cancer Genome Atlas datasets on multiple cancers, biologically meaningful findings are obtained.

First, we develop two robust interaction analysis methods for prognostic outcomes. Compared to continuous and categorical outcomes, prognosis has been less investigated, with additional challenges brought by the unique characteristics of survival times. Most of the existing G-E interaction approaches for prognosis data share the limitation that they cannot accommodate long-tailed or contaminated outcomes. In the first method, we adopt the censored quantile regression and partial correlation for survival outcomes. Under a marginal modeling framework, this proposed approach is robust to long-tailed prognosis and is computationally straightforward to apply. Furthermore, outliers and contaminations among predictors are observed in real data. In the second method, we propose a joint model using the penalized trimmed regression that is robust to leverage points and vertical outliers. The proposed method respects the hierarchical structure of main effects and interactions and has an effective computational algorithm based on coordinate descent optimization and stability selection.

Second, we propose a penalized approach to incorporate additional information for identifying important hierarchical interactions. Due to the high dimensionality and low signal

levels, it is challenging to analyze interactions so that incorporating additional information is desired. We adopt the minimax concave penalty for regularized estimation and the Laplacian quadratic penalty for additional information. Under a unified formulation, multiple types of additional information and genetic measurements can be effectively utilized and improved identification accuracy can be achieved.

Third, we develop a three-step procedure using multidimensional molecular data to identify G-E interactions. Recent studies have shown that collectively analyzing multiple types of molecular changes is not only biologically sensible but also leads to improved estimation and prediction. In this proposed method, we first estimate the relationship between gene expressions and their regulators by a multivariate penalized regression, and then identify regulatory modules via sparse biclustering. Next, we establish integrative covariates by principal components extracted from the identified regulatory modules. Last but not least, we construct a joint model for disease outcomes and employ Lasso-based penalization to select important main effects and hierarchical interactions. The proposed method expands the scope of interaction analysis to multidimensional molecular data.

Last, we present an application using both marginal and joint models to analyze histopathological imaging-environment interactions. In cancer diagnosis, histopathological imaging has been routinely conducted and can be processed to generate high-dimensional features. To explore potential interactions, we conduct marginal and joint analyses, which have been extensively examined in the context of G-E interactions. This application extends the practical applicability of interaction analysis to imaging data and provides an alternative venue that combines histopathological imaging and environmental data in cancer modeling.

Motivated by the important implications of G-E interactions and to overcome the limitations of the existing methods, the goal of this dissertation is to advance in methodological development for G-E interaction analysis and to provide practically useful tools for identifying important interactions. The proposed methods emerge from practical issues observed in real data and have solid statistical properties. With a balance between theory, computation, and data analysis, this dissertation provide four novel approaches for analyzing interactions to achieve more robust and accurate identification of biologically meaningful interactions.

Statistical Methods for Gene-Environment Interactions

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Yaqing Xu

Dissertation Director: Dr. Shuangge Ma

June 2021

Contents

List of Figures

List of Tables

Acknowledgements

Seven years have passed since I first landed in New Haven. I feel blessed to be a part of the Yale community as a graduate student.

I would like to express my gratitude to my advisor and mentor, Professor Shuangge Ma, who has been offering me insightful criticisms at my rises and generous support at my falls. His dedication to research inspired me. I am deeply grateful for his vision and trust that guided me in making important decisions. I would like to thank Professor Hongyu Zhao and Professor Yawei Zhang for serving on my dissertation committee. I am also grateful to Professor Mengyun Wu for her valuable support to my research and informative discussions about life and career.

Special thanks to my friends at Yale – Dr. Xinyue Li, for the warmest memories and delicious homemade treats; Ana Rosen-Vollmar, for those reassuring conversations; and Hao Mei, for many hearty dinners and cheerful grocery rides. I also would like to thank Wenlan Zang and Cat Doudou for memorable cooking Saturdays.

I dedicate this dissertation to my parents, Xingyi Xu and Jingmin Chen, for their unconditional support and endless love despite the long distance.

Chapter 1

Introduction

1.1 Importance of G-E interactions

Gene-environment (G-E) interactions can contribute to the development of complex diseases, together with the significant main effects of genetic and environmental risk factors (Hunter, 2005). Identifying G-E interactions has important implications for understanding etiology and for describing prognosis and response to treatment (Thomas, 2010). One extensively studied G-E interaction is between smoking and gene NAT2 for bladder cancer. In the Spanish Bladder Cancer Study of 1150 cases and 1149 controls, García-Closas et al. (2005) showed an increased risk of bladder cancer among smokers with NAT2 slow acetylation genotype than that for never smokers, compared to those with NAT2 rapid/intermediate acetylators. Other environmental exposures of the chemical arylamines, which are widely used in hair dyes and other consumer products, were also found to be interacting with NAT2 in multiple studies of cancer risk (Skipper et al., 2003). Substantial evidence of the existence of NAT2-arylamine exposure interaction associated with bladder cancer risk has been extensively investigated and supported by the fact that this gene encodes an enzyme that functions to both activate and deactivate arylamine and hydrazine drugs. These interactions were also confirmed to be biologically reasonable because aromatic amines can be detoxified by NAT2 and are one of the most important bladder carcinogens in tobacco smoke (Green et al., 2000; Hein, 2002).

Beyond the better understanding of complex diseases, G-E interactions can also be infor-

mative for predicting disease risk before diagnosis and for providing personalized preventive advice based on the genetic profiles of patients. In this sense, we consider the interaction to be both the effect of genotypes on disease modified by the environmental exposures, and the environmental exposures on disease risk interacting with different genotypes. For example, red meat consumption is associated with the risk of colorectal cancer, and studies have shown the effect of red meat intake was modified by NAT2 polymorphisms (Chen et al., 1998). Specifically, among carriers of the rapid NAT2 alleles, the association between red meat intake and the risk of colorectal cancer was stronger (Nöthlings et al., 2009). This G-E interaction shows that the polymorphisms in gene NAT2 convey differential susceptibilities to the effect of red meat intake on colorectal cancer risk, providing valuable information for individualized prevention and risk prediction.

Identifying G-E interactions can help to discover important genes that are associated with the disease through interacting effects with no significant marginal effects (McAllister et al., 2017). Similarly, searching for G-E interactions can also reveal the environmental risk factors that influence the etiology of disease among genetically susceptible populations. In addition, when we consider a drug as the environmental exposure of interest, pharmacogenetics is a special case of G-E interactions. It has demonstrated significant applications and potential impact on public health and clinical care (Dempfle et al., 2008). For instance, warfarin is commonly used in anticoagulation therapy, but Higashi et al. (2002) demonstrated patients possessing CYP2C9 polymorphisms have an increased risk of over anticoagulation and of bleeding complications so that a lower dose of warfarin is required. Hence, the existence of such interactions can be applied to personalized treatment in clinical practice, by tailoring the therapy for patients who are at risk of adverse side effects or treatment failure.

From simple dichotomous genotypes and environmental exposures in the examples, both genetic and environmental risk factors can take other forms. For genetic measurements, SNPs, genotypes as categorical, and gene expression levels as continuous are available. Similar to environmental exposures, a variety of measurements are of interest given the category of potential risk factors, which include the chemical environment as in aforementioned NAT2 interactions, physical environment such as sun exposure and air pollution, and clinical risk factors described as physiological attributes related to certain diseases.

For instance, clinically relevant factors such as weight and height are usually measured as continuous ones. The response associated with G-E interactions can be disease status such as diagnosis, continuous disease outcomes such as surrogate biomarker measurements, and survival time with censorship.

1.2 Current methods

Though the importance of G-E interactions has been recognized, existing studies barely scratch the surface of the massive data that have been collected and are readily available for analysis and research. Current findings remain insufficient considering the sophisticated mechanisms of complex diseases (Khoury and Wacholder, 2009). Many statistical approaches have been proposed for detecting G-E interactions, especially for categorical responses such as disease status. Consider the disease outcome or phenotype as Y, and p single nucleotide polymorphisms (SNPs) as $G = [G_1, G_2, \ldots, G_p]$ with a single binary environmental risk factor as E.

1.2.1 Hypothesis testing-based approaches

The simplest approach employs a 3×2 contingency table to test if the relative risk for each SNP is significantly different comparing the exposed to the unexposed subjects. Chi-square tests can be conducted as well as Fisher's exact tests. The importance of potential G-E interactions is evaluated by p-values with multiple testing correction. For example, Travis et al. (2010) studied the effects of 12 polymorphisms among 7610 women with breast cancer and 10196 controls. The per-allele relative risk was calculated using logistic regressions to describe the main effect of each of the 12 SNPs, and then compared across two levels of each of the ten environmental risk factors, including age at menarche, height, and others. ANOVA was applied for comparing the means of continuous variables and conventional chi-square tests were used for the proportions of categorical variables. Given a total of 120 tests, the threshold of p-values for statistical significance was corrected to be 0.0004 and no significant evidence of any G-E interaction was concluded in this analysis. Other examples of hypothesis testing-based analysis include Higashi et al. (2002), Lake and Laird (2004),

and many others. Among the existing hypothesis testing analysis methods, the majority have been designed for case-control data. We omit further discussions and refer García-Closas and Lubin (1999), Albert et al. (2001), and Gauderman (2002) for comprehensive discussions about sample size and power calculation in specific study designs for detecting G-E interactions.

One obvious drawback is that the hypothesis testing-based approaches evaluate the relative risks or odds ratios across different levels of genetic and environmental factors (Chatterjee and Wacholder, 2009). When it comes to continuous genetic measurements such as gene expression data, categorizing expression levels would cause a considerable loss of information. Dependent upon the strength of the association and the magnitude of the interaction, statistical methods with increased power are desired. With low-dimensional covariates, robust methods have demonstrated to be powerful and efficient (Wilcox, 2011). For example, Ritchie et al. (2001) developed the multifactor dimensionality reduction (MDR) method based on hypothesis testing and Ritchie et al. (2003) showed the MDR method retains high power in the presence of genotyping error. Yet, these methods usually have limited applicability and demand certain study designs.

1.2.2 Marginal modeling framework

As a result, the most commonly used marginal model when testing for the existence of G-E interactions between certain genes and environmental risk factors is defined as

$$Y \sim \phi(\beta_k G_k + \gamma E + \theta_k G_k \times E), \text{ for } k = 1, 2, \ldots, p,$$

where β_k and γ are the main effects of G_k and E respectively, θ_k is the interaction effect between G_k and E, and $\phi(\cdot)$ is a known link function. $G_k \times E$ represents a two-way product interaction where $\times$ is element-wise multiplication. Note that we omit other covariates to avoid unnecessary notation but they may be added to the above model. Standard estimation, especially likelihood-based techniques, is conducted. Under such a marginal modeling framework, this model fitting is cycled through all genes and environmental factors. Important interactions are selected based on p-values. For instance, we can use the logit link for

a binary Y such as disease diagnosis. The example of gene NAT2 and smoking interaction discussed in García-Closas et al. (2005) was examined by logistic regression with adjustment for relevant covariates, and the odds ratios were assessed for the effect on bladder cancer risk.

Many methods have been proposed to enhance the power and can be broadly summarized into two categories. One is adding a preliminary screening process to reduce the number of tests. For example, Murcray et al. (2009) proposed to screen the associations between SNPs and the environmental exposures using likelihood ratio tests based on the logistic model. Alternatively, Kooperberg and LeBlanc (2008) suggested screening on marginal genetic effects. In both methods, only for SNPs that were selected by a pre-specified significance level, their corresponding G-E interactions were tested for association of disease status in the second step. Combining these two screening approaches, Murcray et al. (2011) developed a hybrid method and Hsu et al. (2012) introduced a cocktail method. The other category for improving the power aims to combine a group of genetic variants and then to perform a set-based test to reduce the multiple-testing burden. Tzeng et al. (2011) proposed a marker-set approach to detect G-E interactions where the genetic similarity and interaction were regressed on the trait similarity between individuals, and the genetic similarity was used to integrate information from multiple polymorphic sites so that the power was increased by reducing the total number of tests.

In fact, marginal models for identifying G-E interactions discussed above are also based on hypothesis testing. The candidate interactions are described by two-way products and the estimates of their coefficients are obtained. With the null hypotheses that the effect of each interaction equals zero, p-values are produced. Meanwhile, hypothesis testing-based approaches summarized in Section 1.2.1 are built under the marginal analysis category, where one interaction is considered at a time and marginal effects on the outcomes are investigated. We separate hypothesis testing-based methods from Section 1.2.2 for those that do not explicitly denote G-E interactions by element-wise multiplication of genetic and environmental risk factors.

1.2.3 Joint modeling framework

Besides marginal effect modeling for identifying G-E interactions, other approaches assume a joint model of the main and interaction effects as

$$Y \sim \phi(\sum_{k=1}^{p} \beta_k G_k + \gamma E + \sum_{k=1}^{p} \theta_k G_k \times E).$$

Joint modeling framework for analyzing G-E interaction is more challenging mainly for two reasons. First, high dimensionality is problematic due to the number of genetic factors and interactions. For instance, given 1000 genes and 5 environmental risk factors, the number of potential interactions is 5000 and the total number of the covariates in joint analysis adds to 6005. To handle such high-dimensional data, regularized estimation is often adopted and increased computational cost is required (Wu and Ma, 2015). For instance, Lasso (Tibshirani, 1996) and the minimax concave penalty (Zhang et al., 2010), two popular penalization techniques, both demand developed computational algorithms for model fitting. More details about coordinate descent algorithms are discussed in Friedman et al. (2010) and Breheny and Huang (2011). Second, the hierarchical structure of main effects and interactions needs to be accommodated under joint modeling. That is, when an interaction is identified, the corresponding main effect of genetic factor should be simultaneously included in the model. The need to respect the "main effects, interactions" hierarchy has been widely recognized to deliver biologically meaningful findings in the literature (Bien et al., 2013; Hao et al., 2018). Yet, directly imposing regular penalization in the joint analysis that does not guarantee the hierarchical structure, important interactions may be identified without their corresponding main effects. Such identification of G-E interactions can lead to false discovery and difficult interpretation.

Several published studies address these challenges by statistical approaches. For example, Liu et al. (2013) proposed a joint model and adopted the group MCP for penalized estimation and hierarchical structure. Simulation study showed that it outperforms alternatives by identifying more true positives and fewer false positives. Zhu et al. (2014) developed a stagewise strategy and employed ℓ_1 penalization to identify G-E interactions. A coordinate descent method was utilized in computation to produce regularized estimation. Given

the massive amount of data collected and the sophisticated mechanisms of interactions, the existing findings remain insufficient and there is a lack of methodology development for G-E interaction analysis.

We note that other analysis frameworks exist (Cordell, 2009; McKinney et al., 2006). One example category arises from a Bayesian standpoint for selecting G-E interactions. For instance, Mukherjee and Chatterjee (2008) proposed an empirical Bayes-type estimator for case-control data, Mukherjee et al. (2010) introduced a proper full Bayesian approach with sample size determination criteria for both estimation and hypothesis testing for G-E interactions, and Yu et al. (2012) developed a resampling-based test derived from a Bayesian model. We refer Simonds et al. (2016) and Wu and Ma (2019) for further reviews.

1.3 Application

We conduct data analysis on publicly available data collected by The Cancer Genome Atlas (TCGA), which provides comprehensive profiling on more than 30 cancer types with high quality for cancer studies. It serves as benchmark data for conducting and comparing different statistical approaches and is ideal for demonstrating the practical applicability of our proposed methods. More information about TCGA can be found online at `http://cancergenome.nih.gov/`. Here, we use lung adenocarcinoma (LUAD) data as an example to introduce the characteristics of the TCGA data and several performance assessments for comparing different methods.

Lung cancer is the leading cause of cancer death globally, and adenocarcinoma of the lung is its most common histological type. From molecular profiling, many genetic mutations have been identified as the driver in certain tumors, for example, ALK (Kwak et al., 2010) and EGFR (Paez et al., 2004). However, the additional unexplained mechanisms of pathway activation, suggesting potential G-E interactions may exist (Network et al., 2014). For data analysis, we download the level 3 data from TCGA Provisional using the R package *cgdsr* (Jacobsen, 2017). There is a total of 544 tumor samples, 230 of which have mRNA, copy number variation (CNV) and sequencing data. For example, mRNA gene expression data is collected using the IlluminaHiseq RNAseq V2 platform, containing a total of 20189

measurements. CNV data is obtained using the Genome-Wide Human SNP Array 6.0 platform with 18342 measurements. DNA methylation is obtained using the Illumina Infinium Human DNA Methylation 450 platform with 21231 measurements. For clinical information, 67 measurements about the participants are available and can be regarded as environmental risk factors. For instance, smoking status is recorded for 353 subjects, which is known as a major cause of lung cancer. Age of 516 participants is included, ranging from 33 to 88 with a mean of 65. In this dissertation, we select environmental factors based on the existing findings in biomedical literature and include them in the model for data analysis. For example, age, gender, tumor pathological stage, and smoking status have been found to be associated with lung cancer prognosis (Westcott et al., 2015), and can be used as environmental risk factors for data analysis. In addition, survival times and censoring indicators, as the prognosis outcomes for the proposed methods, are also reported as clinical information with 262 complete cases. It ranges from 0.13 to 238.11 months with 93 deaths during the follow-up period.

Using the proposed methods, identified interactions are confirmed and validated by searching the current literature for biological implications. We expect that some, if not all, of the selected G-E interactions have already been detected and explained in the existing studies, which can be found by their main genetic effects, as well as interactive effects with other environmental risk factors. We also apply the alternative approaches to analyze the TCGA data. Though analytical methods are different, the actual effects contained in the data should produce similar discoveries, which means the identified G-E interactions may be distinct but the information of those identifications can largely overlap. In this sense, we assess how much the identifications from different methods are overlapped by the modified matrix correlation coefficient (RV-coefficient), which describes the common information of two high-dimensional matrices (Smilde et al., 2009). Moreover, we evaluate and compare the stability performance by the observed occurrence index (OOI), which examines the probability of an interaction being identified in random samples and with a larger value indicating higher stability (Huang and Ma, 2010).

1.4 Summary

Motivated by the importance of G-E interactions and the limitations of the existing interaction analysis methods, in this dissertation, we propose two marginal and two joint modeling approaches for analyzing G-E interactions and extend the applicability to histopathological imaging data. For each proposed method, we investigate numerical results using simulated data under various settings in comparison with multiple alternative approaches. We also conduct analysis on publicly available data collected by The Cancer Genome Atlas (TCGA).

The proposed methods are intertwined in serval aspects. All of the proposed methods, except Chapter 2.2, assumes a linear relationship to model the association between the interactions and outcomes. This modeling strategy of using linear regressions is in accordance with the current statistical theories and applications for analyzing main genetic and interaction effects on disease outcomes. Several benefits of using linear regressions that describe the additive effects on disease outcomes are carried out and magnified in the proposed methods. First, the regression coefficients of main genetic and interactions directly represent their effects on the outcomes. We hence enforce the hierarchical structure of main effects and interactions in this dissertation by decomposing the coefficients. In this way, an interaction can be included in the model only if the corresponding main genetic effect is also included, leading to reasonable interpretability and accurate identification. Consequently, the unified modeling scheme also brings advantages to the estimation and selection process. To fit the regression models, the objective functions proposed in this dissertation consist of the loss function and penalty terms. Since the coefficients of interactions are decomposed as multiplications, it is intuitive to apply separate penalty terms for the main effects and interaction respectively. Without further complication in calculating the regularized solution, we adopt the coordinate descent algorithms that are computationally efficient with well-established convergence properties.

The rest of the dissertation is organized as follows. In Chapter 2, we develop two G-E interaction analysis methods with robustness properties to accommodate outlying observations. In Chapter 3, we propose to incorporate additional information into G-E interaction analysis using penalization. In Chapter 4, we extend to multidimensional molecular data

and develop a three-step strategy for analyzing molecular changes-environment interactions. In Chapter 5, we present an application of interaction analysis using histopathological imaging features for cancer modeling. In the following studies, we demonstrate that the methodological advancement of the proposed methods can effectively overcome the limitations of the existing methods and expand the current scope of interaction analysis. Multiple recently collected data are analyzed using the proposed methods and compared with benchmark alternative approaches, which can provide biologically meaningful identifications and potentially reveal important G-E interactions missed by existing studies. In Chapter 6, we summarize the achievements and limitations of this dissertation and discuss potential future work for analyzing G-E interactions.

Chapter 2

Robust Interaction Analysis Methods for Prognostic Outcomes

2.1 Overview

In practical biomedical studies, the presence of irregular noise caused by various sources is commonly observed. For example, in cancer research, observed survival data is often heterogeneous with many possible reasons (Aalen, 1988). The natural course of a disease can be distinct from person to person, which could be affected by the clinical treatment and the influence of risk factors. Also, individual frailty varies and patients who are more frail will die sooner. Even with strict patient selection as in clinical trials, the natural heterogeneity in study populations requires robust methods to deliver accurate identifications. Additionally, complex diseases like cancer and diabetes may have various subtypes and mechanisms, which can result in heterogeneity in survival times as well. The Cancer Genome Atlas Research Network (2014) stated The Cancer Genome Atlas data on lung adenocarcinoma demonstrated diverse patterns of survival outcomes under different molecular subtypes. Finally, potential data contamination such as human error may contribute to the presence of noise. Relevant discussions can be found in Osborne and Overbay (2004) and Shieh and Hung (2009). For example, medical records are vulnerable to entry errors during data collection, and even diagnostic errors may happen: 10-30% of breast cancers are missed on

mammography and 1-2% of cancers are misread on biopsy samples (Graber, 2013). Together with biological variation among the population over time, the heterogeneity and possible contamination in the observed data is not simply a nuisance, but an important characteristic of the data itself, demonstrating the necessity of developing robust methods for identifying G-E interactions.

Given the biological variation and potential contamination in real data, the popular procedures using non-robust methods for identifying G-E interactions have the following issues. (1) The specified model may not be consistent for all subjects due to the heterogeneity among patients. It is possible that the subgroups of patients demonstrate distinct associations so that robust methods are necessary to retain accurate estimates across dissimilar patterns. Considering the large number of genes, chances are that the estimated significance level will be invalid because the strict model assumptions may not be met for every marginal model. For example, some genes exhibit distinctive expression signatures, and the corresponding residuals do not satisfy the strict assumption of the error distribution. Consequently, the regression coefficients can be misleading, resulting in false positive detections of interactions. (2) Different sets of covariates may correlate with different disease subtypes. In this case, since the traditional methods assume that the same set of covariates in the model, the estimated coefficients may not describe the association properly. Including irrelevant variables or omitting significant ones can result in misleading identifications of interactions. Thus, robustness is required given potentially ambiguous disease subtypes. (3) With insufficient prior knowledge, the regression model can even be misspecified, especially when the biological findings are too limited to validate the specified model across subtypes. In contrast, robust models are less sensitive to model misspecification, and accommodate the complexity of the disease by using weaker model assumptions. (4) Additionally, extreme values could easily disturb the non-robust model, leading to biased estimates and misleading inference. A small proportion of data contamination can skew regression estimates dramatically even with the prescreening step combined with non-robust methods, whereas the robust models remain stable and can still provide accurate identifications.

Considering the common presence of noise introduced by the nature of complex diseases, heterogeneous populations, and even human errors, we propose two robust methods for iden-

tifying G-E interactions to discover new interaction effects and to advance in methodological development. In Chapter 2.2, we propose a robust censored quantile partial correlation approach to identify important interactions while properly controlling for the main genetic and environmental effects under a marginal modeling framework. In Chapter 2.3, we develop a robust penalization approach using the trimmed regression technique under joint modeling. Both of the proposed robust methods can accommodate prognostic response.

2.2 Censored Quantile Partial Correlation for Cancer Prognosis

2.2.1 Introduction

Recent studies have shown that G-E interactions play a critical role for the prognosis of many diseases. For instance, it has been suggested that the interaction between gene TP53 and age affects the prognosis of glioblastoma (Batchelor et al., 2004). Literature review suggests that there is less research on G-E interactions for prognosis, which may be caused by the challenging characteristics of prognosis data (non-negative distributions, censoring, etc.). Recent methodological developments for identifying G-E interactions for prognosis include Shi et al. (2014), Sharafeldin et al. (2015), and a few others.

In practical genetic studies, the long-tailed distributions and contaminations in prognostic response are not uncommon. These studies usually cannot afford conducting strict subject selection, and as such, the subjects are less homogeneous than in for example clinical trials. Sometimes there are some extremely good or bad survivals, which has been observed in quite a few studies. In addition, human errors (for example, mistakes in death records) can also cause long-tailed distributions and contaminations. As the demonstrative examples, consider the LUAD and SKCM (cutaneous melanoma) data collected by TCGA. More information on these data can be found in the data analysis section of this article as well as the TCGA website. For the 262 LUAD subjects analyzed in this section, one has survival time 238.11 months, while the rest 261 have survival times ranging from 0.13 to 129.43 months. In addition, for the 225 SKCM subjects, three have survival times 241.20,

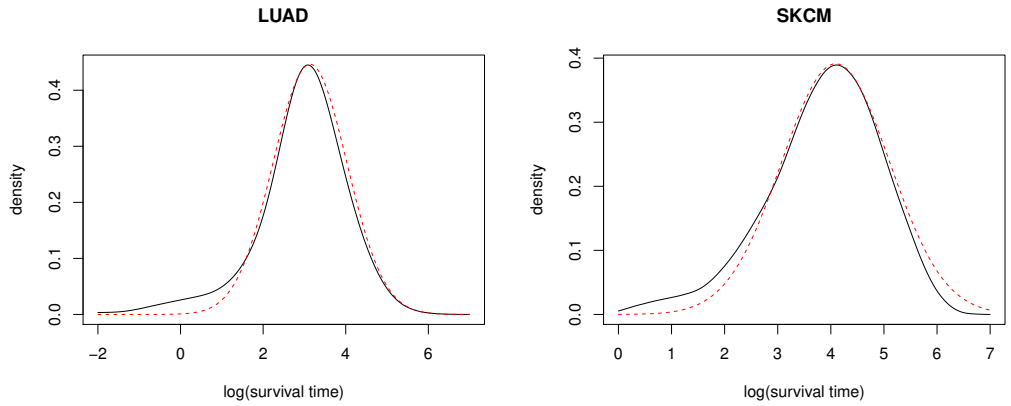


Figure 2.1: Analysis of the LUAD and SKCM data: the empirical distribution of log(survival time) (solid line) and best-fitted normal distribution (dashed line).

268.53, and 339.88 months, while the rest 222 have survival times ranging from 2.04 to 228.42 months. In Figure 2.1, we present the empirical density function of the log survival time as well as the best-fitted Normal density for both datasets. Compared to Normal, we observe the longer left tails (p-values for LUAD and SKCM from the Kolmogorov-Smirnov test are 0.001 and 0.002, suggesting a significant difference from Normal). In "classic" statistical analysis, it has been noted that data with long-tails/contamination cannot be appropriately accommodated by non-robust estimations: even a single extreme value can lead to biased estimation and misleading inference.

For low-dimensional biomedical studies, robust methods have been extensively developed and implemented. For example, Wang and Wang (2009) have proposed the robust censored quantile regression (CQR) approach which is a recursive weighting approach and generalizes the Kaplan-Merier (KM) estimator introduced in published studies. Huang et al. (2007) have developed the robust least absolute deviation estimation based on the AFT model and KM weights (KMW-LAD). Other examples include the rank-based regression (Wang and Zhu, 2006), S-estimation (Tharmaratnam et al., 2010), and others. However, development and implementation in G-E interaction analysis with prognosis data are still much limited.

In this Section, we conduct G-E interaction analysis for data with prognosis responses. To accommodate long-tailed distributions/contamination in the response, we develop a robust censored quantile partial correlation (CQPCorr) approach, which can be applied to

analyze both continuous and categorical variables. This study advances from the existing literature in the following aspects. First, we specifically consider the scenario with long-tailed distributions/contamination in the prognosis response, which is not uncommon but has been little investigated. Second, the proposed approach is built on the quantile regression technique and may have a more solid statistical basis than some alternatives. Quantile regression has been first developed for low-dimensional data (Koenker and Bassett Jr, 1978) and its joint asymptotic distribution, robustness, and statistical inference have been well established (Koenker and Machado, 1999). Compared to least squares regression, quantile regression has been demonstrated to have comparable efficiency for Normal error distribution and perform much better for a wide class of non-Normal error distributions. It has been more recently developed for high-dimensional main effect analysis, and also shown to have good properties, including the consistency and asymptotic normality (Lee et al., 2018; Wang et al., 2012a).Although quantile regression has been a popular tool in statistical analysis, its applications to genetic interaction analysis are still limited. Different from the standard quantile regression technique, the proposed approach adopts data-dependent weights to accommodate censoring. In addition, tailored to interaction analysis, the partial correlation technique is adopted. Third, compared to some alternative robust techniques, the quantile-based is computationally more feasible, making the proposed approach suitable for high-dimensional analysis. It is noted that although components of the proposed approach have roots in existing techniques, development and implementation in the present context are new and innovative. In addition, our extensive numerical study shows that the proposed approach can outperform multiple direct competitors. Overall, this study provides a useful new venue for identifying G-E interactions with prognosis responses.

2.2.2 Methods

Consider a dataset with n independent subjects. For subject i, let T_i be the transformed (e.g., log) survival time of interest, and $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{iq})'$ and $\boldsymbol{Z}_i = (Z_{i1}, \cdots, Z_{ip})'$ be the q- and p-vectors of E and G variables, respectively. To study the interaction between

the kth E factor and jth gene, consider the model

$$T_i = a_{kj} + \alpha_{kj} X_{ik} + \beta_{kj} Z_{ij} + \theta_{kj} X_{ik} Z_{ij} + \epsilon_i, \tag{2.1}$$

where a_{kj} is the intercept, α_{kj}, β_{kj}, and θ_{kj} are unknown coefficients, and ϵ_i is the random error with $P(\epsilon_i < 0 | X_{ik}, Z_{ij}) = \tau$. Note that here a very weak assumption is made on the error distribution, whereas with non-robust estimations, usually very stringent assumptions (for example, normal distribution) are needed. In the above model, one E factor and one G factor are considered. This strategy has been commonly adopted in the literature. See for example Frost et al. (2016) and Zhang et al. (2016). The proposed approach can straightforwardly accommodate multiple E factors and one G factor in a single model. In practice, right censoring is usually present. For subject i, denote C_i as the censoring time which is transferred as the survival time, then we observe $Y_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$.

The CQPCorr approach

Denote X_k, Z_j and T as the random variables corresponding to the kth E factor, jth gene and transformed survival time. In most of the existing studies, the importance of interaction $X_k Z_j$ on T is quantified by the magnitude or p-value of θ_{kj} (Shi et al., 2014). Significantly different from the existing studies, we propose quantifying the importance of interaction $X_k Z_j$ using the quantile partial correlation defined as

$$\text{qpcorr}_\tau(k,j) = \frac{\text{cov}\{\psi_\tau(T - \eta_0^0 - \eta_1^0 X_k - \eta_2^0 Z_j), X_k Z_j - \gamma_0^0 - \gamma_1^0 X_k - \gamma_2^0 Z_j\}}{\sqrt{\text{var}\{\psi_\tau(T - \eta_0^0 - \eta_1^0 X_k - \eta_2^0 Z_j)\}\text{var}(X_k Z_j - \gamma_0^0 - \gamma_1^0 X_k - \gamma_2^0 Z_j)}}. \tag{2.2}$$

Here for a quantile $0 < \tau < 1$, $\psi_\tau(u) = \tau - \mathbf{1}(u < 0)$ and $\rho_\tau(u) = u\psi_\tau(u)$. $(\eta_0^0, \eta_1^0, \eta_2^0) = \text{argmin}\ \mathbb{E}[\rho_\tau(T - \eta_0 - \eta_1 X_k - \eta_2 Z_j)]$ and $(\gamma_0^0, \gamma_1^0, \gamma_2^0) = \text{argmin}\ \mathbb{E}[(X_k Z_j - \gamma_0 - \gamma_1 X_k - \gamma_2 Z_j)^2]$. $\mathbb{E}$ is the expectation function with respect to the random variables X_k, Z_j and T. Note that $\eta_0, \eta_1, \eta_2, \gamma_0, \gamma_1$ and γ_2 take possibly different values for different k and j. We omit the dependence on (k, j) to simplify notations.

The adopted quantile partial correlation measure has multiple desirable properties. The

same as the classic Pearson correlation coefficient, it lies between -1 and 1, and is scale-free and easy to compare across variables. Unlike the simple correlation coefficient, it is defined based on quantile and hence is robust to long-tailed distributions/contamination. In the definition, the main effects of G and E variables are first removed from T and $X_k Z_j$, and then the correlation is computed. Thus, the main effects are removed in a more explicit manner. In the literature, the quantile partial correlation has been used for screening predictors under high-dimensional settings (Ma et al., 2017). However, there is a lack of application in the context of G-E interaction analysis. In our analysis, there is one additional significant complication: T is subject to right censoring and not always observable. To tackle this problem, we propose the censored quantile partial correlation (CQPCorr) technique, which advances from the quantile partial correlation by adopting weights to accommodate censoring. Overall, the proposed approach consists of the following steps.

Step I Conduct the censored quantile regression of the prognosis response on the main effects, which corresponds to the first term in the numerator of (2.2). Specifically, $(\eta_0^0, \eta_1^0, \eta_2^0)$ is estimated as

$$(\hat{\eta}_0, \hat{\eta}_1, \hat{\eta}_2) = \text{argmin} \sum_{i=1}^{n} w_i \rho_\tau (Y_i - \eta_0 - \eta_1 X_{ik} - \eta_2 Z_{ij}) + (1 - w_i) \rho_\tau (Y^{+\infty} - \eta_0 - \eta_1 X_{ik} - \eta_2 Z_{ij}). \tag{2.3}$$

$Y^{+\infty}$ is a fixed value that is large enough.

Here we adopt the weights w_i's to accommodate censoring. The basic strategy is to redistribute the mass of a censored observation to the non-censored observations to the right. This is achieved by creating pseudo-observations with weights w_i's for censored observations and complementary weights $1 - w_i$'s at a point large enough. Motivated by the literature (Wang and Wang, 2009), w_i is defined for a censored observations as

$$w_i = \frac{\tau - F(C_i | X_{ik}, Z_{ij})}{1 - F(C_i | X_{ik}, Z_{ij})} \tag{2.4}$$

if $F(C_i | X_{ik}, Z_{ij}) < \tau$, where $F(\cdot | X_{ik}, Z_{ij})$ is the conditional cumulative distribution function of the survival time given the covariates. For better computational feasibility, we ap-

proximate $F(t|X_{ik}, Z_{ij})$ using the Kaplan-Meier (KM) estimator and calculate the weight function at the τth quantile as

$$w_i = \begin{cases} \frac{\tau - \hat{F}(C_i)}{1 - \hat{F}(C_i)}, & \text{if } \delta_i = 0 \text{ and } \hat{F}(C_i) < \tau, \\ 1, & \text{otherwise,} \end{cases}$$

for $i = 1, \ldots, n$. Here $\hat{F}(t) = 1 - \prod_{i:t_{(i)} \le t} [1 - (n-i+1)^{-1}]^{\delta_{(i)}}$, where the subscript "$(i)$" refers to the ith subject in the sorted data (according to the observed times, from the smallest to the largest).

Step II Remove the main G and E effects from the interaction, and obtain the "net" G-E interaction effect. Specifically, estimate $(\gamma_0^0, \gamma_1^0, \gamma_2^0)$ using the simple least squared approach, where

$$(\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2) = \text{argmin} \sum_{i=1}^{n} (X_{ik} Z_{ij} - \gamma_0 - \gamma_1 X_{ik} - \gamma_2 Z_{ij})^2.$$

Step III Results from the above two steps are combined to assess whether the interaction has an effect on prognosis after accounting for the main effects. Specifically, for interaction $X_k Z_j$, the censored quantile partial correlation is defined as

$$\text{cqpcorr}_\tau(k, j) = \frac{n^{-1} \sum_{i=1}^{n} \left[\tau - w_i \mathbf{1}\left(r_i^{(1)}(k, j) < 0\right)\right] r_i^{(2)}(k, j)}{\sqrt{(\overline{w^2}\tau - \overline{w}^2\tau^2)} \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(r_i^{(2)}(k, j)\right)^2}}, \tag{2.5}$$

where

$$r_i^{(1)}(k, j) = Y_i - \hat{\eta}_0 - \hat{\eta}_1 X_{ik} - \hat{\eta}_2 Z_{ij}, r_i^{(2)}(k, j) = X_{ik} Z_{ij} - \hat{\gamma}_0 - \hat{\gamma}_1 X_{ik} - \hat{\gamma}_2 Z_{ij},$$
$$\overline{w} = n^{-1} \sum_i w_i, \quad \text{and} \quad \overline{w^2} = n^{-1} \sum_i w_i^2.$$

As in Step I, the weights are introduced to accommodate censoring.

Remarks Advancing from the existing quantile partial correlation studies, the proposed approach introduces weights to accommodate censoring. In survival analysis, there are many ways to estimate $F(t|X_{ik}, Z_{ij})$ in (2.4) to accommodate censoring. Popular examples include the semi-parametric Cox model, accelerated failure time model and transformation model,

nonparametric KM estimator, and others. We adopt KM estimator as it is computationally simpler and has been a common choice in the literature. It also has the advantage of making no assumption on the underlying data distributions and models, leading to more robust results. It is noted that, although may seem "straightforward", coupling the KM weights with quantile partial correlation to achieve robustness with censored data has not been pursued in the literature. Examining the procedures described above suggests that the proposed approach can be directly applied to analysis with multiple E factors. Setting all weights equal to one, the proposed approach can directly accommodate continuous responses without censoring.

2.2.3 Simulation

Simulation is conducted to gauge performance of the proposed method and compare with direct competitors. For all simulated data, we set $n = 200$, $p = 1000$, and $q = 5$. There are thus a total of 5,000 candidate interactions and 1,005 candidate main effects. Other settings are as follows. (a) The G factors are generated from a multivariate Normal distribution with marginal mean 0 and variance 1. The continuous distribution mimics gene expression data analyzed below. The Normal distribution, although somewhat simpler than practically encountered, has been extensively adopted in published studies. Following published literature, we consider the AR (auto-regressive) structure with different parameters, where the jth and lth G variables have correlation coefficient $\rho^{|j-l|}$. We consider two levels of correlation with $\rho = 0.5$ and 0.3. (b) There are five continuous E factors (E1) that are generated from a multivariate Normal distribution with marginal mean 0, marginal variance 1, and AR correlation ($\rho = 0.5$). (c) The log event time Y is computed from the following accelerated failure time (AFT) model,

$$Y = \sum_{k=1}^{q} \alpha_k X_k + \sum_{j=1}^{p} \beta_j Z_j + \sum_{k=1}^{q} \sum_{j=1}^{p} \theta_{kj} X_k Z_j + \varepsilon, \tag{2.6}$$

where ε is the random error. Note that this is a joint model, under which prognosis is determined by the joint effects of multiple main effects and interactions. We choose this model as it may better describe "biological reality". Thus, it is sensible to conduct marginal analysis

and compare results to the data generating mechanisms described above. Additionally, the log censoring times are generated from uniform distributions and conditionally independent of the event times (conditional on covariates). The parameters are adjusted so that the censoring rates are around 20%. (d) Consider three error distributions: $N(0,1)$ (Error 1), $90\%N(0,1)+10\%N(\pm 50,1)$ (Error 2) and $80\%N(0,1)+20\%N(0,50)$ (Error 3). The last two scenarios represent different types/levels of long-tailed distributions/contamination. (e) There are 16 G-E interactions together with two main E effects and five main G effects. Although the proposed method focuses on interaction identification, the main effects are assumed to make the simulated dataset closer to practical data. Five different coefficient settings are considered.

C1 has $\theta_{kj}=2$ for $k=1,2$ and $j=1,\cdots,5$, and $\alpha_1=\alpha_2=\beta_1=\cdots=\beta_5=1$. Under this setting, the main effects are weaker than the corresponding interactions. In addition, $\theta_{kj}=1$ for $k=3,4,5$ and $j=6,7$. All other coefficients are 0.

C2 has $\theta_{kj}=1.5$ for $k=1,2$ and $j=1,\cdots,5$, and $\alpha_1=\alpha_2=\beta_1=\cdots=\beta_5=1.5$. Under this setting, the main effects and interactions have the same level. In addition, $\theta_{kj}=1$ for $k=3,4,5$ and $j=6,7$. All other coefficients are 0.

C3 is the same as C1 except that the magnitudes of the main effects are larger, that is $\alpha_1=\alpha_2=\beta_1=\cdots=\beta_5=3$.

C4 is the same as C1 except that the magnitudes of the interactions are smaller, that is $\theta_{kj}=0.5$ for $k=1,2$ and $j=1,\cdots,5$, and $\theta_{kj}=0.5$ for $k=3,4,5$ and $j=6,7$.

C5 is the same as C1 except that the interactions with main effects have negative coefficients, that is $\theta_{kj}=-2$ for $k=1,2$ and $j=1,\cdots,5$.

Under all five settings, there are two types of interactions. The first one includes ten interactions (θ_{kj}, $k=1,2$ and $j=1,\cdots,5$) with main effects and the second one includes six interactions (θ_{kj}, $k=3,4,5$ and $j=6,7$) without main effects. Thus, the hierarchy of the second type is violated. There are a total of 21 simulation scenarios, covering a wide spectrum of settings.

Comparison with the alternative methods

Besides the proposed approach, we also consider four alternatives with the same covariate effects as in (2.1), including the AFT model, Cox model, censored quantile regression approach (CQR), the least absolute deviation estimation with KM weights (KMW-LAD). As introduced in Section 1, AFT and Cox models are perhaps the most popular methods for analyzing prognosis data, but without the property of accommodating long-tailed distributions and contamination. Note that under our simulation settings which are based on AFT model, the Cox model is mis-specified. However, due to its popularity, it has been also adopted as the alternative method in many published studies without sufficient model diagnostics (Liang et al., 2016; Song et al., 2014) and is a suitable benchmark for comparison. The CQR and KMW-LAD approaches are also robust. Different from the proposed approach, they consider one interaction and its corresponding main effects in one regression model. For all the proposed method and four alternatives, p-values are computed and used to rank/identify interactions. We note that there are other G-E interaction analysis methods that are potentially applicable to the simulated data. The above four methods are chosen because their analysis frameworks are the closest to the proposed and also because of their popularity and competitive performance demonstrated in published studies. With the proposed approach and CQR, we set the quantile $\tau = 0.5$. Choosing this specific quantile makes the proposed approach more comparable to KMW-LAD (which is a special case of quantile regression with $\tau = 0.5$).

The main goal of our analysis is to accurately identify important interactions. Identification accuracy is evaluated using multiple measures, including: (a) TP20, which is the number of true positives when 20 interactions are selected; (b) TP40, which is defined in a similar way as TP20; (c) pAUC, which is the standardized partial area under the ROC curve when the number of false positives are restricted to 150 (Robin et al., 2011); (d) TP.FDR, which is the number of true positives when the number of important interactions is selected using the FDR (false discovery rate) approach with target FDR 0.1; (e) FP.FDR, which is the corresponding number of false positives; and (f) E.FDR, which is the corresponding estimated FDR. All five measures have been adopted in multiple publications.

Under each setting, we simulate 200 replicates. Summary results for scenarios C1 and C2 are presented in Tables 2.1 and 2.2, respectively. It is observed that the proposed approach has similar or better performance than the alternatives. When there is no contamination (Error 1), the proposed approach may be slightly inferior to the non-robust alternatives. This is reasonable as the non-robust alternatives can be more efficient for data with no contamination. Although the true model is not Cox, the Cox-model-based approach is observed to have satisfactory performance. Both the Cox and AFT models are transformation models. The "robustness" of the Cox model (to model mis-specification) has also been observed in the literature. The proposed approach can more accurately identify important interactions than the robust alternatives. For example in Table 2.2 with $\rho = 0.3$ and Error 1, the proposed approach selects on average 10.3 true nonzero interactions when the model size is 40, while CQR and KMW-LAD select 4.9 and 8.8 on average, respectively. When there is a stronger correlation which is common in practice, the advantage of the proposed approach over the alternatives gets more prominent, even over AFT and Cox for data without contamination. For example in Table 2.1 with $\rho = 0.5$ and Error 1, the proposed approach has pAUC=0.94, compared to 0.84 (AFT), 0.90 (Cox), 0.74 (CQR), and 0.90 (KMW-LAD). When data have contamination, the proposed approach has significant advantages. For example in Table 2.1 with $\rho = 0.3$ and Error 3, the proposed approach has pAUC=0.77, compared to 0.65 (AFT), 0.72 (Cox), 0.62 (CQR), and 0.71 (KMW-LAD). Across all settings, the proposed approach performs moderately or slightly better than the KMW-LAD approach. It is reasonable that the improvement over the KMW-LAD is not dramatic: this approach has a strategy similar to the proposed, with the loss function being a special case of quantile regression, and using the KM weights to accommodate censoring. However, we can still observe improvement which supports the proposed three-steps strategy. We also examine an example of the partial ROC curves in Figure A.1 (Appendix) under setting C1 with $\rho = 0.3$ and Error 3. It is shown that the solid line representing the proposed approach is superior to the others.

Table 2.1: Simulation results for setting C1 with the AR correlation structure. In each cell, mean (sd) based on 200 replicates.

	Error	Method	TP20	TP40	pAUC	TP.FDR	FP.FDR	E.FDR
$\rho = 0.3$	1	AFT	9.8(1.2)	10.7(0.9)	0.79(0.05)	10.9(1.2)	69.0(57.0)	0.78(0.17)
		Cox	9.6(1.6)	10.5(1.7)	0.83(0.05)	9.0(2.0)	16.2(27.8)	0.49(0.22)
		CQR	3.1(1.4)	4.9(1.8)	0.67(0.04)	7.9(2.0)	116.1(31.3)	0.93(0.02)
		KMW-LAD	7.1(2.3)	8.8(2.4)	0.81(0.06)	3.2(2.3)	0.8(1.3)	0.12(0.16)
		CQPCorr	8.6(1.8)	10.2(2.0)	0.84(0.06)	4.8(2.2)	0.8(0.9)	0.11(0.11)
	2	AFT	4.8(1.8)	5.8(1.7)	0.71(0.05)	3.2(2.3)	4.9(6.0)	0.38(0.34)
		Cox	6.9(2.0)	8.3(2.1)	0.78(0.06)	4.1(2.4)	2.2(2.6)	0.32(0.24)
		CQR	2.9(1.3)	4.1(1.8)	0.65(0.05)	6.3(2.5)	94.4(36.4)	0.94(0.02)
		KMW-LAD	6.4(1.7)	8.3(1.7)	0.79(0.05)	1.2(1.1)	0.3(0.6)	0.08(0.17)
		CQPCorr	7.7(1.9)	8.8(2.1)	0.81(0.05)	3.3(1.7)	0.4(0.6)	0.07(0.11)
	3	AFT	3.2(2.4)	4.3(2.8)	0.65(0.08)	1.8(2.3)	5.7(9.0)	0.43(0.41)
		Cox	5.0(2.9)	6.4(2.9)	0.72(0.09)	2.3(2.5)	1.7(1.8)	0.30(0.30)
		CQR	1.8(1.4)	3.0(1.7)	0.62(0.06)	5.8(2.5)	105.0(39.2)	0.94(0.02)
		KMW-LAD	4.0(1.3)	5.4(1.6)	0.71(0.05)	0.9(1.0)	0.2(0.4)	0.11(0.21)
		CQPCorr	6.0(2.4)	7.7(2.6)	0.77(0.07)	2.4(1.9)	0.5(0.8)	0.09(0.15)
$\rho = 0.5$	1	AFT	11.2(1.4)	12.5(1.7)	0.84(0.06)	14.1(1.3)	142.9(165.7)	0.84(0.11)
		Cox	11.6(1.2)	13.2(1.2)	0.90(0.04)	12.9(1.7)	29.8(29.6)	0.60(0.17)
		CQR	4.7(1.6)	6.9(1.8)	0.74(0.06)	11.5(2.0)	133.0(33.1)	0.92(0.02)
		KMW-LAD	10.6(1.8)	12.3(1.7)	0.90(0.05)	7.9(2.3)	2.3(1.6)	0.21(0.13)
		CQPCorr	12.2(1.6)	13.8(1.5)	0.94(0.03)	10.9(2.0)	3.3(2.5)	0.21(0.12)
	2	AFT	9.3(1.7)	10.2(1.5)	0.81(0.04)	9.4(2.7)	22.1(29.6)	0.50(0.29)
		Cox	10.4(1.3)	11.4(1.7)	0.86(0.04)	9.9(1.9)	6.3(4.1)	0.35(0.13)
		CQR	5.2(1.4)	7.1(1.5)	0.73(0.04)	9.6(2.1)	108.4(23.8)	0.91(0.02)
		KMW-LAD	9.0(2.0)	9.9(1.8)	0.84(0.05)	5.6(2.5)	1.2(1.3)	0.17(0.17)
		CQPCorr	10.4(1.7)	12.0(2.0)	0.89(0.05)	8.0(2.4)	1.6(1.2)	0.16(0.09)
	3	AFT	7.0(2.1)	8.1(2.1)	0.77(0.06)	5.9(2.9)	17.1(20.7)	0.56(0.28)
		Cox	9.3(1.5)	10.2(1.6)	0.84(0.05)	8.4(2.1)	8.0(13.2)	0.35(0.22)
		CQR	4.5(1.6)	6.3(1.7)	0.70(0.05)	9.0(1.9)	105.8(44.7)	0.92(0.03)
		KMW-LAD	8.7(1.9)	10.7(1.9)	0.86(0.06)	4.0(2.0)	0.8(1.0)	0.13(0.16)
		CQPCorr	10.7(1.7)	12.2(1.9)	0.90(0.06)	7.5(2.4)	1.3(1.4)	0.14(0.11)

Table 2.2: Simulation results for setting C2 with the AR correlation structure. In each cell, mean (sd) based on 200 replicates.

	Error	Method	TP20	TP40	pAUC	TP.FDR	FP.FDR	E.FDR
$\rho = 0.3$	1	AFT	9.5(1.6)	11.1(2.0)	0.82(0.06)	11.3(3.1)	67.6(64.0)	0.73(0.21)
		Cox	8.8(1.6)	10.5(2.0)	0.85(0.05)	7.2(2.6)	5.1(7.6)	0.29(0.20)
		CQR	3.0(1.6)	4.4(1.9)	0.67(0.05)	8.4(2.3)	122.4(44.7)	0.93(0.02)
		KMW-LAD	6.2(1.8)	8.2(2.0)	0.80(0.06)	2.0(1.5)	1.0(1.2)	0.26(0.30)
		CQPCorr	8.1(2.0)	9.8(2.0)	0.84(0.06)	3.9(2.3)	1.2(1.3)	0.17(0.19)
	2	AFT	3.2(1.8)	4.8(2.4)	0.66(0.07)	1.4(1.9)	4.3(6.6)	0.45(0.43)
		Cox	5.4(2.2)	6.9(2.5)	0.73(0.07)	2.3(2.3)	2.8(4.8)	0.36(0.38)
		CQR	2.3(1.3)	3.5(1.6)	0.63(0.04)	6.1(2.1)	111.3(32.9)	0.94(0.02)
		KMW-LAD	5.7(2.0)	7.4(2.4)	0.77(0.06)	1.5(2.3)	0.2(0.5)	0.04(0.12)
		CQPCorr	7.2(2.4)	9.0(2.2)	0.81(0.06)	2.4(1.9)	0.1(0.4)	0.03(0.07)
	3	AFT	1.5(1.4)	2.2(1.4)	0.58(0.06)	0.4(1.0)	2.7(6.2)	0.47(0.48)
		Cox	3.9(2.3)	4.9(2.8)	0.69(0.09)	1.0(1.6)	1.5(2.6)	0.26(0.39)
		CQR	2.2(1.2)	3.2(1.5)	0.62(0.04)	5.4(2.1)	105.2(38.0)	0.95(0.02)
		KMW-LAD	4.2(1.4)	5.7(1.8)	0.73(0.06)	0.3(0.5)	0.1(0.2)	0.03(0.12)
		CQPCorr	5.4(2.2)	7.2(2.3)	0.76(0.07)	1.2(1.4)	0.1(0.2)	0.01(0.06)
$\rho = 0.5$	1	AFT	11.9(1.4)	13.4(1.4)	0.88(0.05)	14.1(1.5)	86.6(100.4)	0.75(0.14)
		Cox	12.4(1.5)	13.6(1.8)	0.92(0.04)	12.6(2.0)	11.6(13.6)	0.38(0.20)
		CQR	5.0(2.0)	7.0(2.1)	0.73(0.06)	11.0(2.0)	138.2(40.5)	0.92(0.03)
		KMW-LAD	10.9(1.8)	12.5(1.8)	0.91(0.05)	7.4(2.4)	1.9(1.8)	0.18(0.15)
		CQPCorr	12.0(1.5)	13.8(1.5)	0.95(0.03)	10.1(2.5)	2.5(2.0)	0.17(0.12)
	2	AFT	7.7(1.9)	9.1(2.0)	0.79(0.06)	7.3(3.4)	27.8(49.0)	0.53(0.27)
		Cox	10.1(2.1)	11.3(2.1)	0.86(0.06)	8.1(3.4)	3.2(4.8)	0.20(0.20)
		CQR	4.8(1.5)	6.4(2.0)	0.72(0.05)	9.9(2.0)	112.1(39.1)	0.91(0.03)
		KMW-LAD	9.5(2.1)	11.6(2.1)	0.88(0.06)	5.8(2.3)	1.1(1.2)	0.14(0.15)
		CQPCorr	11.2(2.0)	12.8(2.0)	0.91(0.05)	8.2(2.5)	1.4(1.3)	0.13(0.10)
	3	AFT	5.0(3.1)	6.3(3.3)	0.72(0.10)	4.3(4.4)	12.1(17.4)	0.56(0.38)
		Cox	6.9(2.7)	8.7(3.0)	0.80(0.08)	5.0(4.0)	5.0(11.3)	0.26(0.28)
		CQR	3.9(1.5)	5.5(1.5)	0.70(0.04)	8.8(2.0)	110.5(36.4)	0.92(0.02)
		KMW-LAD	8.5(2.0)	10.7(1.9)	0.85(0.05)	4.0(2.5)	1.0(1.3)	0.15(0.18)
		CQPCorr	9.6(1.8)	11.2(2.3)	0.87(0.06)	5.8(3.0)	1.4(1.5)	0.15(0.13)

With target FDR 0.1, it can be seen that the proposed method performs better in achieving the nominal FDR control and has the smallest estimated FDR under most settings. Except KMW-LAD, the alternatives cannot control the FDR. For example, in Table 2.1 with $\rho = 0.3$ and Error 1, the proposed method has the estimated FDR 0.11, compared to 0.78 (AFT), 0.49 (Cox), 0.93 (CQR), and 0.12 (KMW-LAD). Under the settings with a weak correlation, the values of TP.FDR with the proposed method are relatively small due to the limited sample size. We further examine the results for scenario C1 with $\rho = 0.3$

and various values of sample size in Tables A.1-A.3 (Appendix). With a large enough sample size, the proposed method is able to identify majority of the true positives with the estimated FDR approximately being 0.1. The improvement of TP.FDR is also observed when there is a stronger correlation ($\rho = 0.5$) even with a small sample size.

In addition, we conduct analysis on the simulated datasets under coefficient scenarios C3-C5 with $\rho = 0.5$. Summary results are provided in A.4-A.6 (Appendix). It it can be seen that all methods perform slightly worse under these three scenarios compared to scenario C1. This may due to that the relative magnitudes of the interactions to main effects under scenarios C3 and C4 are smaller, and the interaction and its corresponding main effects have different directions under scenario C5. Similar to under the previous simulation scenarios, the proposed method performs better than or comparable to the alternatives. For example in Table A5 with Error 2 (Scenarios C4), the proposed method has TP20=7.6, compared to 1.2 (AFT), 4.2 (Cox), 3.4 (CQR), and 7.2 (KMW-LAD). For settings C1 and C2, we also examine other scenarios with G factors with banded correlation structure, E factors with binary measurements, and higher censoring rate (35%). Detailed results are provided in Appendix. Similar conclusions can be drawn for the G factors with banded correlation structure. The performance of all methods decay when the datasets are with binary E factors or a higher censoring rate, which is as expected. However, the proposed CQPCorr still has superior or comparable performance. An advantage of quantile-based approaches is that multiple quantiles can be potentially examined to generate a more comprehensive picture. We analyze the simulated datasets under coefficient scenarios C1 with $\rho = 0.5$ using the proposed method and CQR with various values of τ, and present the summary results in Table A.15 (Appendix). Slight differences across the results are observed. The proposed method can achieve favorable performance with multiple quantiles.

Computational cost Simulation suggests that the proposed analysis is computationally feasible. The analysis of 5000 interactions (along with the corresponding main effects) can be accomplished within ten seconds using a laptop with standard configurations. Although a large number of permutations may need to be computed, as they can be analyzed in a highly parallel manner, the overall computational cost is still much affordable. For example, for the 10,000 permutations, the analysis can be accomplished within 10 minutes using 100

parallel jobs on our cluster. More parallel jobs can be conducted if less computational time is desirable.

2.2.4 Data analysis

In the following, we analyze the TCGA data on lung adenocarcinoma (LUAD) and cutaneous melanoma (SKCM). With a high quality, TCGA provides an ideal testbed for new analysis approaches. Although TCGA data have been analyzed in multiple published studies, as described in the first section, it is worthwhile re-examining data using the new robust approach. We refer to the TCGA website for more information on the study design. Data analyzed are downloaded from TCGA Provisional using the R package *cgdsr*.

Analysis of LUAD data

Lung cancer is the leading cause of cancer death globally, and adenocarcinoma of lung is its most common histological type. In analysis, we focus on primary tumor samples of the Whites. The prognosis response of interest is overall survival. Data are available for 262 subjects, among whom 93 died during followup. The survival times range from 0.13 to 238.11 months with median 20.65 months. The E factors analyzed include smoking pack years (smoking), age, American Joint Committee on Cancer (AJCC) tumor pathologic stage (stage), and gender, all of which have been suggested as potentially associated with lung cancer prognosis (Westcott et al., 2015). Following the literature, here we take a loose definition of E factors to also include clinical variables. For G factors, we analyze mRNA gene expressions, which have been collected using the IlluminaHiseq RNAseq V2 platform. A total of 20,189 measurements are available. As the number of relevant genes is not expected to be large, we conduct a simple prescreening and select the top 2,000 genes with the largest variances across all the samples for downstream analyses.

When applying the proposed approach, we compute p-values based on 10,000 permutations and use the FDR (false discovery rate) approach to identify important interactions. With a target FDR of 0.1, 48 G-E interactions are identified, and the CQPCorr values are shown in Table 2.3. Literature search suggests that the identified genes and interactions may have important biological implications. For example, a negative correlation between

survival and the AP3D1-Gender interaction is observed. Gene AP3D1 has been reported as involved in fusions in lung cancer and overexpressed in lung adenocarcinoma in women compared with men. Gene BPIFB1 (LPLUNC1) is a secretory protein that is predominantly present in lung tissues and has been shown to be potentially relevant to lung carcinogenesis. Gene CHEK2 is a cell cycle-control gene encoding a pluripotent kinase that can cause arrest or apoptosis in response to DNA damage, and its mutations have been shown to be associated with an increased risk of lung cancer. CPSF4 has been found to play an important role in regulating lung cancer cell proliferation and survival, and has been suggested as a potential prognostic biomarker and therapeutic target for lung adenocarcinoma. Gene DKK1 has been observed to increase the migratory activity of mammalian cells and suggested as a novel serologic and histochemical biomarker for lung adenocarcinoma. Published analysis has also suggested that inhibition of gene PCSK9 induces apoptosis and inhibits proliferation of lung adenocarcinoma cells via endoplasmic reticulum stress and mitochondrial signaling pathways. WFS1 protein is expressed in various tissues but at higher levels in the lung and has been found to probably contribute to the relationship of cigarette smoking and lung cancer.

Data are also analyzed using the alternatives. The summary of comparison is presented in the upper sub-table of Table A.16 (Appendix). When evaluating the differences in findings, we use both the simple numbers of findings as well as the RV-coefficients (Smilde et al., 2009), which measure the common information of two matrices of interactions, with a larger value indicating a higher degree of similarity. The RV-coefficient can effectively account for correlations of different genes and is a more objective and rigorous measure of overlap. More detailed identification results of the alternative approaches are available from the authors. Table A16 suggests that although there are overlapping identifications, the proposed approach identifies a different set of interactions. As the numbers of identifications identified by different approaches are quite different, we also consider the top 40 interactions and evaluate overlap. Note that because of ties, the numbers can be slightly off. The results are shown in the lower sub-table of Table A.16 (Appendix). Again it is observed that although there are overlaps, the proposed approach makes different findings. With practical data, it is difficult to objectively evaluate identification accuracy. Here we evaluate the stability of

findings, which may provide some insight into the analysis. Specifically, we compute the observed occurrence index (OOI) (Huang and Ma, 2010), which lies between 0 and 1 and can be roughly interpreted as the probability of an interaction being identified in random samples and with a larger value indicating higher stability. For the interactions identified using the FDR controlling procedure, we compute the OOI values. The proposed approach has mean OOI (across the identified interactions) 0.41, compared to 0.26 (AFT), 0.34 (Cox), 0.18 (CQR), and 0.14 (KMW-LAD). The OOI values with proposed and alternative methods are all moderate, which has been also observed in the literature. This may due to the more complex correlation structures, lower signal-to-noise ratios, higher censoring rates, small sample size, and other factors in real datasets. However, the proposed method still has slightly better stability, which provide support to a large extent to the superiority of the proposed approach. More results and discussions on stability with simulated datasets are provided in Appendix.

Analysis of SKCM data

The occurrence of skin cancer is rapidly increasing over the last decade, and cutaneous melanoma is responsible for approximately 75% of all deaths from skin cancer. In analysis, we focus on metastatic samples of the Whites. Data are available for 225 subjects. The prognosis response of interest is overall survival. Among the subjects, 93 died during followup, with survival times ranging from 2.04 to 339.88 months (median 56.31 months). For E variables, we consider Breslow thickness at diagnosis, Clark level, age, AJCC tumor pathologic stage, and gender, all of which have been suggested in the literature. For G variables, we consider gene expressions, for which 20,189 measurements are available. With the same processing as above, 2,000 gene expressions are selected for downstream analysis.

The proposed approach identifies 80 G-E interactions with the FDR control. Details are presented in Table 2.4. Most of the identified interactions are with Breslow thickness and Clark level, which are the most important prognostic parameters in evaluating the primary tumors (Dickson and Gershenwald, 2011). Published studies suggest potentially important

Table 2.3: Analysis of the LUAD data using CQPCorr: identified G-E interactions.

	Smoking	Age	Stage	Gender
ABI2	-0.178			
ABR				-0.200
AKR1D1	0.186			
AP3D1				-0.197
BPIFB1			0.133	
BRE.AS1	0.206			
C19ORF57			0.200	
C1ORF229			0.188	
C1RL	0.193			
C3ORF38	-0.185			
C6ORF163	0.187			
CAPN7	-0.175			
CHEK2			0.185	
CST5		0.188		
CSTF2				-0.197
DAGLA				-0.210
DKK1				-0.184
EIF2B5				0.197
ETV5			-0.188	
FAF2			-0.214	
FAM114A2			-0.260	
HABP4	-0.204			
HIST2H2AC		-0.222		
LINC01547			0.209	
LINGO1	0.183			
MFAP3			-0.187	
MMP25				-0.222
MRFAP1L1				0.214
MTF2				0.197
MZF1.AS1		0.212		
NCAPD2		-0.182		
PAXIP1.AS1			0.172	
PCDHA11			-0.207	
PCSK9	0.181			
PIGR				0.176
RAET1L		0.192		
RCOR2			0.196	
RNF14			-0.207	
SNX4				0.236
SP2				-0.197
TAPT1				0.191
TTTY14	0.197			
UBE2S		-0.191		
UBLCP1			-0.212	
UGT1A3	0.185			
WFS1	0.185			
ZNF174			-0.199	
ZNF721				0.211

implications of the findings. For example, gene GSN has been shown to be crucial for migration and invasion of melanoma cell lines, indicating its potential effects on cutaneous melanoma. Gene NFKBIE has been suggested as a candidate oncogene in melanomas, of which recurrent mutations have been found at several nearby hotspots in melanomas. The expression levels of gene PEBP1 (RKIP) in melanoma cancer cell lines have been found to be low relative to primary melanocytes, indicating its important role in melanoma turmorgenesis. Gene PLD1 has been observed to be strongly expressed in primary and metastatic melanomas, enhancing the activity of basal phospholipase D enzyme in a protein phosphorylation-independent manner in melanoma cells. Gene RNF144A has been found to be specifically upregulated in melanocytes, which function to avoid uncontrolled proliferation and to be a part of embryonic development, acting as cancer development modulators. Gene SSR2 exerts a prosurvival functionality in human melanoma cells, and high expression levels of SSR2 have been observed to be associated with an unfavorable disease outcome in primary melanoma patients. Gene TRPM2 is capable of inducing melanoma apoptosis and necrosis and has been suggested as an important diagnostic and prognostic marker for primary cutaneous melanoma.

Data are also analyzed using the alternatives. The summary comparison results are shown in Table A.16 (Appendix). Both the FDR control results and (roughly) top forty lists suggests that the proposed approach identifies interactions different from the alternatives. Stability is also evaluated. For the proposed approach, the average OOI is 0.37, compared to 0.26 (AFT), 0.28 (Cox), 0.19 (CQR), and 0.22 (KMW-LAD).

2.2.5 Discussion

The identification of G-E interactions is an important task in genetic epidemiology studies. In this article, we focus on prognosis data. Prognosis is an essential endpoint in the study of cancer, cardiovascular diseases, and many others. Different from most of the existing studies, we have developed a novel approach which can accommodate long-tailed distributions/contamination in the prognosis response. The proposed approach has an intuitive

Table 2.4: Analysis of the SKCM data using CQPCorr: identified G-E interactions.

	Breslow thickness	Clark level	Age	Stage	Gender
ABCA8	-0.198				
ADGRD1	-0.197				
AGPAT2	-0.211				
ANAPC2	-0.194				
ATAD3A	-0.211				
ATP5G2	-0.223				
ATP5SL			0.217		
AURKAIP1	-0.217				
BOLA2	-0.205				
C15ORF41		0.222			
C19ORF53	-0.251				
C1ORF204				0.220	
C1ORF226			0.231		
C4A		-0.200			
C9ORF85					-0.220
CASP7	0.205				
CD164	0.240				
CECR1		-0.198			
CEP57L1	0.211				
CHMP1A			0.197		
CHRD				-0.215	
COX6A1	-0.219				
CTXN2			0.222		
DDT	-0.210				
DERL3		-0.204			
DPPA3				-0.211	
DUSP26		-0.222			
E2F6		0.212			
ECSIT	-0.212				
EIF3G	-0.226				
FATE1		-0.221			
FGFR1OP	0.208			0.241	
GADD45GIP1	-0.223				
GSN	-0.208				
KCNE3		-0.213			
KCNK17	-0.195				
KIAA2013	-0.194				
KLK4	-0.203				
LHB	-0.192				-0.200
LRSAM1					-0.209
LYRM5				0.221	
MAF1	-0.191				
MAGOHB				0.218	
MAPK4	-0.207				
MZB1		-0.202			
NCKAP1		0.214			
NDUFA11	-0.206				
NDUFB7	-0.221				
NFKBIE		-0.222			
NKX2.4					-0.197
NOS1AP			0.221		
NTMT1	-0.222				
NUDT19		0.235			
PARVB	-0.191				
PDSS1		0.219			
PEBP1	-0.199				
PLD1	0.197				
PRSS37					0.232
RNF144A		0.236			
SMYD4				0.235	
SRR				0.233	
SSR2	-0.216				
SURF2	-0.215				
TBC1D10A		-0.218			
TCTA		-0.215			
TCTE1	-0.220				
THEM6	-0.203				
TMEM159		0.217			
TPRN					-0.208
TRPM2		-0.206			
UQCRQ	-0.216				
VAMP4	0.207				
VCAN	-0.199				
VSTM5	0.231				
WDR4	-0.209				
ZFP41	-0.213				
ZNF671		-0.239			
ZUFSP	0.198				

formulation and solid statistical basis and can more explicitly remove main G and E effects so as to facilitate the analysis of interactions. By examining a wide spectrum of simulation settings, we have shown that the proposed approach can outperform direct competitors. It is interesting to note that it has more accurate identification than two robust approaches. In the analysis of TCGA lung and skin cancer data, interactions different from using the alternatives are identified. Literature search shows that the identified genes and interactions have sound biological interpretations. In addition, the proposed approach has more stable identifications.

The proposed approach conducts marginal analysis, which is more popular than joint analysis in the current literature. The proposed approach can be potentially extended to joint analysis. The formulations in the three steps may directly hold. However, with the high dimensionality of joint analysis, the estimation demands regularization. This extension is expected to be highly nontrivial and warrants a separate investigation. The proposed method may not respect the "main effects, interactions" hierarchical constraint, which is often explored in recent G-E interaction analysis (Liu et al., 2013; Wu et al., 2018). Under this constraint, an interaction can be selected only if the corresponding main effects are also selected. In (2.5), when the main E and G factors are not associated with the response, the estimated η_0, η_1 and η_2 in $r_i^{(1)}$ can be approximately zero. Then no information is removed from the response and the proposed CQPCorr can still work. Thus the identified interactions are not necessary to have corresponding main effects. As our main interest is to identify interactions, no specific attention is paid to the selection of main effects. More studies on the identification of main effects and "main effects, interactions" hierarchy are deferred to future investigation. In Step II, we adopt least square as it is computational simpler. Data analysis demonstrates that the proposed method identifies biological sensible interactions with better stability. If needed, robust regression, such as quantile-based method, can be adopted as in Step I. Besides KM estimator, it can be of interest to estimate the conditional cumulative distribution function $F(t|X_{ik}, Z_{ij})$ using other approaches, for example the Cox model or AFT model. The weights so estimated may generate different results. The details will be studied in the future. The proposed method can be also extended to accommodate non-linear or nonparametric gene-environment interactions. In

Steps I and II, the nonparametric models, such as the varying coefficients model, can be adopted. In Step III, the censored quantile partial correlation can be developed based on some correlations measuring nonlinear dependence, for example distance correlation. In the study, we have focused on methodological development and numerical examination. Theoretical study for robust methods under high-dimensional settings is still much limited and will be postponed to future research.n numerical study, we set $\tau = 0.5$ which is one of the most popular choices in the literature, and generate satisfactory results. More extensive numerical analysis with multiple τ may be of interest. For example, we can compare the identified interactions across different τ to explore some interesting findings, such as that some interactions are important across all τ, while some variables may be important only for certain τ. In data analysis, significant differences across approaches are observed. High-dimensional interaction identification can be more challenging than the identification of main effects. Even in simulation (which has simpler settings), a few false positives are observed. The significant differences observed in Table A.16 (Appendix) are at least partly attributable to the potential false positives. In the literature, G-E interaction analysis for lung and skin cancers is still limited. The sound biological implications of the identified genes provides at least partial support to the validity of our analysis. This is further supported by the improved stability measured using OOI. More functional studies are needed to confirm the findings.

2.3 Penalized Trimmed Estimation and Selection for Joint Interaction Analysis

2.3.1 Introduction

Among successful approaches developed for detecting important G-E interactions associated with the etiology, diagnosis and prognosis of many complex diseases, joint analysis has attracted increasing interest. It can accommodate all genetic and environmental risk factors, and their interactions in a single model, given that the biological processes are usually dominated by the joint effects of multiple genetic changes. To facilitate the estimation and

interpretation, the "main effects, interactions" hierarchical constraint is often imposed (Bien et al., 2013; Wu et al., 2018), where an interaction can be identified only if its corresponding main effects are also identified. Compared to marginal analysis, there are more challenges in joint analysis due to the high dimensionality of genomic measurements and hierarchical constraint (Chai et al., 2017).

Despite many advantages, most of the existing interaction analysis approaches have the limitation of nonrobustness. They usually assume that data have no outliers/contaminations. However, in practice, outliers/data contaminations are not uncommon in both predictor and response spaces (Osborne and Overbay, 2004), which are known as leverage points and vertical outliers. More specifically, for some types of G factors, such as gene expression, outliers/contaminations may occur because of technical problems in profiling, human errors and genetic abnormalities (Li and Wong, 2001). For the disease-related clinical response (for example, Breslow's depth for skin cutaneous melanoma), outliers/contaminations can be caused by errors in data collection and recording and inadvertently incorrect sampling. In addition, sometimes there are extremely long or short survivals in prognosis studies due to the mistakes in death records as well as misclassification in the cause of death. In Figure 2.2, we show the distributions of some G factors and Breslow's depth for the SKCM (skin cutaneous melanoma) data collected by TCGA (The Cancer Genome Atlas), where both leverage points and vertical outliers are clearly observed. More information on this data is available in the data analysis section of this article. For nonrobust approaches, it has been shown that these outliers can lead to biased estimation and false marker identification. Recently, a few approaches have been developed for robust G-E interaction analysis, including those based on quantile regression (Wang et al., 2017) or correlation (Xu et al., 2019), least absolute deviation (LAD) loss (Wu et al., 2018), rank-based loss function (Wu et al., 2015), and others. However, these approaches are only robust to outliers in response but cannot accommodate leverage points in predictor space. The interaction studies on both vertical outliers and leverage points are still much limited (Wu and Ma, 2019).

In this Section, we develop a joint model respecting the "main effects, interactions" hierarchical structure for G-E interaction analysis. The unique characteristic of this study is accommodating outliers/contaminations in both predictor and response spaces. The

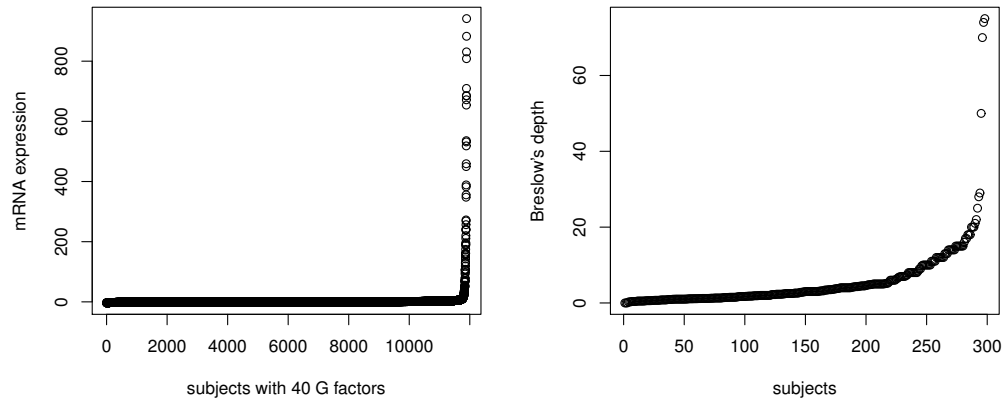


Figure 2.2: Analysis of SKCM data: the distributions of some G factors and the Breslow's depth.

proposed approach is built on the robust trimmed regression technique, which can accommodate many types of data, such as continuous biomarkers and censored survival times. It significantly differs from least absolute deviation regression and other robust approaches which only have robustness property towards vertical outliers. Our study extends the traditional trimmed regression to interaction analysis and develops the "coefficient decomposition+penalization" framework for hierarchical selection, which may have independent methodological value. Advanced from the existing trimmed regression approaches which are usually built with the predefined size of trimmed set, we propose a more flexible data-driven process to determine the number of outliers, leading to satisfactory efficiency and robustness. In addition, a stability selection strategy is adopted to more accurately select the trimmed subject set. Overall, this study provides a practically useful new venue for G-E interaction analysis.

2.3.2 Methods

For a subject, let y be the response of interest, which can be a continuous marker, categorical disease status, or survival time. Let $\boldsymbol{z} = (z_1, \cdots, z_q)$ be the q environmental/clinical variables and $\boldsymbol{x} = (x_1, \cdots, x_p)$ be the p genetic variables. We consider the joint regression

model with all G and E effects and their interactions,

$$\mathbb{E}(y;\boldsymbol{z},\boldsymbol{x})=\phi\left(\alpha_0+\boldsymbol{z}\boldsymbol{\alpha}+\boldsymbol{x}\boldsymbol{\beta}+\sum_{k=1}^{q}\boldsymbol{w}^{(k)}\boldsymbol{\eta}_k\right), \tag{2.7}$$

where ϕ is the known link function, $\mathbb{E}(\cdot)$ denotes expectation, α_0 is the intercept, $\boldsymbol{\alpha}=(\alpha_1,\cdots,\alpha_q)'$, $\boldsymbol{\beta}=(\beta_1,\cdots,\beta_p)'$ and $\boldsymbol{\eta}_k=(\eta_{k1},\cdots,\eta_{kp})',k=1,\cdots,q$ are the regression coefficients for main E factors, main G factors and their interactions, respectively, and $\boldsymbol{w}^{(k)}=(z_kx_1,\cdots,z_kx_p)$.

We assume n independent subjects and use the subscript "i" to denote the ith subject. Denote the design matrices of E and G variables as $\boldsymbol{Z}_{n\times q}$ and $\boldsymbol{X}_{n\times p}$, and the response vector as $\boldsymbol{y}_{n\times 1}$. Under model (2.7), the unknown parameters $\boldsymbol{\theta}=\left(\alpha_0,\boldsymbol{\alpha}',\boldsymbol{\beta}',\boldsymbol{\eta}_1',\cdots,\boldsymbol{\eta}_q'\right)'$ can be estimated by minimizing the negative log-likelihood function,

$$L\left(\boldsymbol{\theta};\boldsymbol{Z},\boldsymbol{X},\boldsymbol{y}\right)=\frac{1}{n}\sum_{i=1}^{n}l_i\left(\boldsymbol{\theta}\right),$$

with the deviance $l_i\left(\boldsymbol{\theta}\right)$, which are usually not robust to vertical outliers or leverage points.

Robust trimmed estimation and selection

Instead of using the negative log-likelihood function directly, we propose the following robust objective function based on trimming technique,

$$L(\boldsymbol{\theta};\boldsymbol{Z},\boldsymbol{X},\boldsymbol{y},\mathcal{S})=\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}l_i\left(\boldsymbol{\theta}\right), \tag{2.8}$$

where $\mathcal{S}$ is an outlier-free subset of $\{1,2,\cdots,n\}$ and $|\mathcal{S}|$ denotes the cardinality of set $\mathcal{S}$. We first consider the most popular linear regression model,

$$y_i=\alpha_0+\boldsymbol{z}_i\boldsymbol{\alpha}+\boldsymbol{x}_i\boldsymbol{\beta}+\sum_{k=1}^{q}\boldsymbol{w}_i^{(k)}\boldsymbol{\eta}_k+\varepsilon_i, \tag{2.9}$$

with

$$l_i\left(\boldsymbol{\theta}\right)=\left(y_i-\alpha_0-\boldsymbol{z}_i\boldsymbol{\alpha}-\boldsymbol{x}_i\boldsymbol{\beta}-\sum_{k=1}^{q}\boldsymbol{w}_i^{(k)}\boldsymbol{\eta}_k\right)^2\triangleq r_i^2,$$

where ε_i is the random error.

Let $\boldsymbol{r} = (r_1, \cdots, r_n)'$, then $\mathcal{S}$ is defined as

$$\mathcal{S} = \{1 \leq i \leq n : |r_i - \text{median}(\boldsymbol{r})| < \mu \text{ MAD}(\boldsymbol{r})\}, \tag{2.10}$$

where median($\boldsymbol{r}$) and MAD($\boldsymbol{r}$) are the median and median absolute deviation of vector $\boldsymbol{r}$ adjusted by a factor 1.4826, and $\mu > 0$ is a tuning parameter.

The penalization is adopted for regularized estimation and variable selection, which has been a popular choice in several recent studies. For respecting "main effects, interactions" hierarchy, the coefficient for the interaction term $\boldsymbol{\eta}_k$ is decomposed as $\boldsymbol{\eta}_k = \boldsymbol{\beta} \odot \boldsymbol{\gamma}_k$, where $\odot$ represents the component-wise multiplication. Then, the following robust penalized objective function is proposed,

$$\begin{aligned} L_p(\boldsymbol{\theta}; \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{y}, \mathcal{S}) &= \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \left(y_i - \alpha_0 - \boldsymbol{z}_i \boldsymbol{\alpha} - \boldsymbol{x}_i \boldsymbol{\beta} - \sum_{k=1}^{q} \boldsymbol{w}_i^{(k)} (\boldsymbol{\beta} \odot \boldsymbol{\gamma}_k) \right)^2 \\ &+ \sum_{j=1}^{p} \rho(|\beta_j|; \lambda_1, \xi) + \sum_{k=1}^{q} \sum_{j=1}^{p} \rho(|\gamma_{kj}|; \lambda_2, \xi), \end{aligned} \tag{2.11}$$

where $\rho(|\nu|; \lambda_1, \xi) = \lambda_1 \int_0^{|\nu|} \left(1 - \frac{x}{\lambda_1 \xi}\right)_+ dx$ is the minimax concave penalty (MCP) (Zhang et al., 2010), λ_1 and λ_2 are data-dependent tuning parameters, and ξ is the regularization parameter. The proposed estimate $\hat{\boldsymbol{\theta}}$ is defined as the minimizer of (2.11) with the optimal subset $\hat{\mathcal{S}}$. The nonzero components of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}} \odot \hat{\boldsymbol{\gamma}}_k (k = 1, \cdots q)$ are regarded as the important main G effects and interactions that are associated with the response.

The proposed approach is motivated by the following considerations. As opposed to the nonrobust squared loss, the robust trimmed squared loss is adopted in (2.11) based on a subset $\mathcal{S}$ of subjects. The definition of $\mathcal{S}$ in (2.10) can exclude those subjects with extreme absolute residuals due to the deviated values in the spaces of predictors and/or response. It significantly advances from the existing robust G-E interaction analyses (Wang et al., 2017; Wu et al., 2018, 2015) which can only accommodate outliers in response but not in predictors. Besides, the robust measures of central location (median) and scale (MAD) are adopted in $\mathcal{S}$, leading to more accurate detection of the number of outliers. Different from

the existing studies on the least trimmed squares estimator (Alfons et al., 2013; Kurnaz et al., 2018) where the size of $\mathcal{S}$ is predefined, the proposed approach determines the value of $|\mathcal{S}|$ based on the residuals themselves and data-driven parameter μ. The identification of $\mathcal{S}$ becomes more flexible to achieve sufficiently high efficiency for the dataset without outliers and satisfactory robustness against data contamination. When μ is large enough, the proposed approach is reduced to the squared loss. In addition, motivated by the pairwise interaction analysis with strong hierarchal constraint developed in Choi et al. (2010), we adopt the decomposition $\boldsymbol{\eta}_k = \boldsymbol{\beta} \odot \boldsymbol{\gamma}_k$ so that if an interaction term is selected ($\beta_j \gamma_{kj} \neq 0$), the corresponding main genetic effect must also be selected ($\beta_j \neq 0$). The MCP penalty is then imposed on β_j and γ_{kj} for variable selection given its satisfactory statistical and numerical properties. Here, E factors are not subject to penalized selection and always included in the model as they are usually pre-selected by clinical evidences and with low dimensionality. This decomposition framework for respecting hierarchical G-E interaction structure has the advantage of lucid interpretation and a less complex computational algorithm.

We also modify $l_i(\boldsymbol{\theta})$ to accommodate other types of response variables. For example, for the right-censored survival response with observed logarithm survival time y and censoring indicator δ, we consider the weighted squared loss under the accelerated failure time (AFT) model,

$$l_i(\boldsymbol{\theta}) = w_i \left(y_i - \alpha_0 - \boldsymbol{z}_i \boldsymbol{\alpha} - \boldsymbol{x}_i \boldsymbol{\beta} - \sum_{k=1}^{q} \boldsymbol{w}_i^{(k)} \boldsymbol{\eta}_k \right)^2 \triangleq \left(r_i^{(w)} \right)^2,$$

where the data $\{(\boldsymbol{x}_i, \boldsymbol{z}_i, y_i, \delta_i), i = 1, \ldots, n\}$ have been sorted by y_i in ascending order, and the weight w_i is the Kaplan-Meier (KM) estimator defined as $w_1 = \frac{\delta_1}{n}$, $w_i = \frac{\delta_i}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_j}$, $i = 2, \cdots, n$. This weighted approach has been adopted in many published studies due to its considerably low computational cost and good statistical properties (Huang et al., 2006). Using the subjects with nonzero weights and their corresponding $r_i^{(w)}$, the proposed approach can then proceed in the same manner. For categorical and count data under generalized linear model, a similar weighted squared loss can be conducted based on the Taylor series expansion. In numerical study, we examine both continuous data under the linear regression model and survival data under the AFT model.

Algorithm

A modified C-steps algorithm is developed to obtain the optimal subset $\hat{\mathcal{S}}$ and corresponding estimation $\hat{\boldsymbol{\theta}}$, which is motivated by the stability selection (Meinshausen and Bühlmann, 2010). We present the proposed algorithm in Algorithm 1. In this algorithm, the most challenging step is the optimization of the objective function (2.11) given the outlier-free subset $\mathcal{S}$. In Algorithm 2, we adopt an iterative coordinate descent (CD) algorithm, which optimizes $L_p\left(\boldsymbol{\theta}; \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{y}, \mathcal{S}\right)$ with respect to one parameter at a time and iteratively cycles through all parameters until convergence. Denote $\boldsymbol{y}_{\mathcal{S}}$ as the components of $\boldsymbol{y}$ indexed by $\mathcal{S}$ and $\boldsymbol{X}_{\mathcal{S}}$ as the rows of $\boldsymbol{X}$ indexed by $\mathcal{S}$.

Algorithm 1: Robust trimmed estimation and selection

Step 1: For $t = 1, \cdots, T$,

Step 1.1 Set $m = 0$. Draw $q + 10$ observations from the dataset at random as the elemental subset $\mathcal{S}^{(t,0)}$. Compute

$$\boldsymbol{\theta}^{(t,0)} = \text{argmin}_{\boldsymbol{\theta}} L_p\left(\boldsymbol{\theta}; \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{y}, \mathcal{S}^{(t,0)}\right)$$

Step 1.2 Set $m = m + 1$. Compute

$$\boldsymbol{r}^{(t,m)} = \boldsymbol{y} - \alpha_0^{(t,m-1)} - \boldsymbol{Z}\boldsymbol{\alpha}^{(t,m-1)} - \boldsymbol{X}\boldsymbol{\beta}^{(t,m-1)} - \sum_{k=1}^{q} \boldsymbol{W}^{(k)}\left(\boldsymbol{\beta}^{(t,m-1)} \odot \boldsymbol{\gamma}_k^{(t,m-1)}\right),$$

$$\mathcal{S}^{(t,m)} = \left\{1 \le i \le n : \left| r_i^{(t,m)} - \text{median}\left(\boldsymbol{r}^{(t,m)}\right)\right| < \mu \text{ MAD}\left(\boldsymbol{r}^{(t,m)}\right)\right\},$$

and

$$\boldsymbol{\theta}^{(t,m)} = \text{argmin}_{\boldsymbol{\theta}} L_p\left(\boldsymbol{\theta}; \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{y}, \mathcal{S}^{(t,m)}\right)$$

Step 1.3 Repeat Step 1.2 until convergence, where the convergence criterion is taken as

$$\frac{|L_p\left(\boldsymbol{\theta}^{(t,m)}; \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{y}, \mathcal{S}^{(t,m)}\right) - L_p\left(\boldsymbol{\theta}^{(t,m-1)}; \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{y}, \mathcal{S}^{(t,m-1)}\right)|}{|L_p\left(\boldsymbol{\theta}^{(t,m-1)}; \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{y}, \mathcal{S}^{(t,m-1)}\right)|} < 10^{-4}.$$

Step 1.4 Return the subset $\mathcal{S}^{(t,m_{\text{stop}})}$ of the subjects selected at the stopping iteration m_{stop}.

Step 2: Compute the final set $\hat{\mathcal{S}}$ of the selected subjects,

$$\hat{\mathcal{S}} = \left\{ i : \frac{1}{T} \sum_{t=1}^{T} I\left(i \in \mathcal{S}^{(t,m_{\text{stop}})} \right) > \tau \right\},$$

where $I(\cdot)$ is the indicator function and $\tau \in (0,1)$ is a tuning parameter.

Step 3: Compute the final estimation $\hat{\boldsymbol{\theta}}$ of the unknown parameters,

$$\hat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}} L_p\left(\boldsymbol{\theta}; \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{y}, \hat{\mathcal{S}} \right).$$

Algorithm 2: Iterative coordinate descent (CD) algorithm

Step 1 Initialize $b = 0$, $\left(\alpha_0^{(b)}, \left(\boldsymbol{\alpha}^{(b)} \right)' \right)' = (\tilde{\boldsymbol{Z}}_{\mathcal{S}}' \tilde{\boldsymbol{Z}}_{\mathcal{S}})^{-1} \tilde{\boldsymbol{Z}}_{\mathcal{S}}' \boldsymbol{y}_{\mathcal{S}}$ with $\tilde{\boldsymbol{Z}} = (\boldsymbol{1}_{n\times 1}, \boldsymbol{Z})$, $\boldsymbol{\beta}^{(b)} = \boldsymbol{0}$, and $\boldsymbol{\gamma}_k^{(b)} = \boldsymbol{0}$, where we denote b as the index of iteration.

Step 2 Set $b = b + 1$. With α_0, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}_k$ fixed at $\alpha_0^{(b-1)}$, $\boldsymbol{\alpha}^{(b-1)}$ and $\boldsymbol{\gamma}_k^{(b-1)}$, optimize (2.11) with respect to $\boldsymbol{\beta}$. Let $\tilde{\boldsymbol{y}}^{(b)} = \boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\alpha}^{(b-1)} - \alpha_0^{(b-1)}$ and $\tilde{\boldsymbol{X}}^{(b)} = \boldsymbol{X} + \sum_{k=1}^{q} \boldsymbol{W}^{(k)} \odot \left(\boldsymbol{1}_{n\times 1} \boldsymbol{\gamma}_k^{(t-1)} \right)'$, then

$$\boldsymbol{\beta}^{(b)} = \text{argmin} \frac{1}{|\mathcal{S}|} \left\| \tilde{\boldsymbol{y}}_{\mathcal{S}}^{(b)} - \tilde{\boldsymbol{X}}_{\mathcal{S}}^{(b)} \boldsymbol{\beta} \right\|_2^2 + \sum_{j=1}^{p} \rho(|\beta_j|; \lambda_1, \xi). \tag{2.12}$$

For $j = 1, \cdots, p$, conduct the following steps sequentially,

Step 2.1 Compute

$$\boldsymbol{r}_{(-j)}^{(b)} = \tilde{\boldsymbol{y}}_{\mathcal{S}}^{(b)} - \sum_{l<j} \tilde{\boldsymbol{x}}_{\mathcal{S},l}^{(b)} \beta_l^{(b)} - \sum_{l>j} \tilde{\boldsymbol{x}}_{\mathcal{S},l}^{(b)} \beta_l^{(b-1)},\ \chi_j^{(b)} = \frac{1}{n} \left(\tilde{\boldsymbol{x}}_{\mathcal{S},j}^{(b)} \right)' \tilde{\boldsymbol{x}}_{\mathcal{S},j}^{(b)},\ \varphi_j^{(b)} = \frac{1}{n} (\tilde{\boldsymbol{x}}_{\mathcal{S},j}^{(b)})' \boldsymbol{r}_{(-j)}^{(b)},$$

Step 2.2 Update the estimate of β_j as

$$\beta_j^{(b)} = \begin{cases} \text{ST}\left(\varphi_j^{(b)}, \lambda_1 \right) / \left(\chi_j^{(b)} - \frac{1}{\xi} \right) & \left| \varphi_j^{(b)} / \chi_j^{(b)} \right| \le \lambda_1 \xi, \\ \varphi_j^{(b)} / \chi_j^{(b)} & \text{else}, \end{cases}$$

where $\mathrm{ST}(\nu, \lambda_1) = sgn(\nu)(|\nu| - \lambda_1)_+$ is the soft-thresholding operator.

Step 3 With α_0, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ fixed at $\alpha_0^{(b-1)}$, $\boldsymbol{\alpha}^{(b-1)}$ and $\boldsymbol{\beta}^{(b)}$, optimize (2.11) with respect to $\boldsymbol{\gamma}_k, k = 1, \cdots, q$. Let $\breve{\boldsymbol{y}}^{(b)} = \boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\alpha}^{(b-1)} - \boldsymbol{X}\boldsymbol{\beta}^{(b)} - \alpha_0^{(b-1)}$ and $\left(\tilde{\boldsymbol{W}}^{(k)}\right)^{(b)} = \boldsymbol{W}^{(k)} \odot \left(\mathbf{1}_{n\times 1}\boldsymbol{\beta}^{(b)}\right)'$, then

$$\left((\boldsymbol{\gamma}_1^{(b)})', \cdots, (\boldsymbol{\gamma}_q^{(b)})'\right)' = \mathrm{argmin}\frac{1}{|\mathcal{S}|}\left\|\breve{\boldsymbol{y}}_{\mathcal{S}}^{(b)} - \sum_{k=1}^{q}\left(\tilde{\boldsymbol{W}}_{\mathcal{S}}^{(k)}\right)^{(b)}\boldsymbol{\gamma}_k\right\|_2^2 + \sum_{k=1}^{q}\sum_{j=1}^{p}\rho(|\gamma_{kj}|; \lambda_2, \xi) \quad (2.13)$$

For $k = 1, \cdots, q$ and $j \in \left\{j : \beta_j^{(b)} \neq 0\right\}$, conduct the two steps similar to Step 2.1 and Step 2.2 sequentially.

Step 4 Compute

$$\left(\alpha_0^{(b)}, \left(\boldsymbol{\alpha}^{(b)}\right)'\right)' = (\tilde{\boldsymbol{Z}}_{\mathcal{S}}'\tilde{\boldsymbol{Z}}_{\mathcal{S}})^{-1}\tilde{\boldsymbol{Z}}_{\mathcal{S}}'\left(\boldsymbol{y}_{\mathcal{S}} - \boldsymbol{X}_{\mathcal{S}}\boldsymbol{\beta}^{(b)} - \sum_{k=1}^{q}\boldsymbol{W}_{\mathcal{S}}^{(k)}\left(\boldsymbol{\beta}^{(b)} \odot \boldsymbol{\gamma}_k^{(b)}\right)\right).$$

Step 5 Repeat Steps 2-4 until convergence, where the convergence criterion is taken as

$$\frac{|L_p\left(\boldsymbol{\theta}^{(b)}; \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{y}, \mathcal{S}\right) - L_p\left(\boldsymbol{\theta}^{(b-1)}; \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{y}, \mathcal{S}\right)|}{|L_p\left(\boldsymbol{\theta}^{(b-1)}; \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{y}, \mathcal{S}\right)|} < 10^{-4}.$$

Different from the original C-steps algorithm which conducts a sufficiently large number of initial subsampling (500 adopted in Alfons et al. (2013); Kurnaz et al. (2018)) and returns the results with the smallest objective function, the proposed algorithm identifies the optimal outlier-free subset based on the stability selection. With stability selection, we do not simply select one model which may not be optimal with insufficient initializations. The subset selection depends on the whole process where the outliers have smaller probability to be included, leading to more accurate detection and the lower requirement for a large number of initializations. In our numerical study, we set $T = 50$, which generates satisfactory result. Another advantage of the proposed algorithm is in Step 3 of Algorithm 2. Due to the decomposition $\eta_{kj} = \beta_j\gamma_{kj}$, we only need to update γ_{kj} when $\beta_j \neq 0$, dramatically reducing the searching space and computational cost. Both algorithms are guaranteed to converge as the value of the objective function (2.11) decreases at each step. It is observed that convergence is achieved in a small to moderate number of iterations in both simulation and

case study. For a simulated dataset with $q = 5$, $p = 1000$ and $n = 250$, the analysis with $T = 50$ takes about five minutes using a laptop with standard configurations.

Tuning parameters We set $\mu = 2.5$ in our numerical studies based on the 99.5% quantile of the standard normal distribution, motivated by that 1% of the observations are expected to be outliers for the normal distribution. For simulation scenarios with continuous G factors and AR structure under linear model (see the next section for details), we further examine the outlier detection results (as a function of μ) to better comprehend the effects of μ. In Table A.17, two specific measures are considered, including true positive (TP) and false positive outliers (FP). For the five different error distributions, a larger μ detects fewer false positives but also fewer true positives. On the other hand, a smaller μ produces more true positives as well as more false positives. When $\mu = 2.5$, it is observed to be able to effectively control the false positives and have satisfactory performance on the detection of true positives. As suggested by Meinshausen and Bühlmann (2010), the stability selection results are not sensitive to the threshold value τ in a range of $(0.6, 0.9)$. In our numerical studies, we set $\tau = 0.6$. For the regularization parameter ξ in the MCP penalties, we follow the published studies (Shi et al., 2014) and set $\xi = 6$. A grid search is conducted to choose the values of (λ_1, λ_2) of the MCP penalties using BIC criterion with model size as the degrees of freedom.

2.3.3 Simulation

We assess the performance of the proposed analysis with extensive simulations. A total of forty simulation scenarios are considered. Under all scenarios, we set $q = 5$ and $p = 1,000$. There are thus a total of 1,005 main effects and 5,000 interactions. (a) Two types of G factors are considered, mimicking continuous gene expression and categorical SNP data, respectively. The continuous G variables are generated from a multivariate normal distribution with marginal means 0 and marginal variances 1. We consider two correlation structures. The first is an AR (auto-regressive) structure where the correlation between the jth and kth G variables is $0.3^{|j-k|}$. The second is a Band (banded) structure where the correlation between jth and kth G variables is 0.33 if $|j - k| = 1$ and 0 otherwise. For the discrete G variables, we further dichotomize the above continuous variables at the

1st and 3rd quartiles and generate 3-level measurements $(0, 1, 2)$. (b) There are three continuous and two binary E factors, where the three continuous ones are simulated from a multivariate normal distribution with marginal means 0 and the AR structure as mentioned above, and the two binary ones are simulated from a binomial distribution with a success probability of 0.6. (c) All E factors, eight main G factors and fourteen G-E interactions are assumed to have nonzero coefficients randomly generated from Uniform(0.6, 1), where the strong hierarchy is satisfied. The rest coefficients are zero. (d) We consider two types of response variables and models. The first is a continuous response under the linear model (2.9). The second is the censored survival data under the AFT model, where the observed logarithm survival times are generated based on model (2.9), and the censoring times generated from an exponential distribution with the parameter adjusted so that the censoring rate is around 20%. (e) Five types of data contaminations are considered. The first three ones have no outliers in predictors. The first one (D1) has error distribution $N(0, 1)$ which is also without outliers in response. The second (D2) and third (D3) ones have error distribution $90\%N(0, 1) + 10\%Cauchy(0, 5)$ and $90\%N(0, 1) + 10\%N(20, 1)$, where outliers exist in response. The fourth (D4) and fifth (D5) ones are assumed to contain leverage points. Specifically, for dataset with continuous G factors, 2% and 8% of the subjects have G factor measurements added by 20 and $N(0, 2)$, respectively. For dataset with categorical G factors, 10% of the subjects are re-generated from a multinomial distribution with probability $(0.5, 0.3, 0.2)$ for $(0, 1, 2)$. The error distributions for D4 and D5 are $N(0, 1)$ and $90\%N(0, 1) + 10\%Cauchy(0, 5)$. Thus, D4 only has outliers in predictors, while D5 has outliers in both predictor and response spaces. (f) We set the sample size $n = 250$ and $n = 300$ for the continuous and survival responses, respectively.

Besides the proposed approach (referred to as "**LTS-MCP-Hier**"), the following alternatives for joint analysis are also considered. The first four approaches conduct variable selection on all G factors and G-E interactions directly, without considering the hierarchical structure. **LS-MCP** is based on the nonrobust squared loss function and MCP penalty, implemented by the R package *ncvreg*. **LAD-LASSO** consists of the robust least absolute deviations and LASSO penalty which has robustness property towards vertical outliers. It is realized using the R package *quantreg*. **RLARS** is the robust least angle regression with

robust correlation measure for variable selection (Khan et al., 2007) and is realized using the R package *robustHD*. It has been demonstrated to be robust to both vertical outliers and leverage points. **LTS-MCP** is similar to the proposed, except that the hierarchical structure is not reinforced and the original C-steps algorithm is used instead of stability selection. The last one is **LS-MCP-Hier**, which has the same modeling framework as the proposed, except that the nonrobust squared loss function is adopted. The above alternative approaches cover different types of G-E interaction analyses and can comprehensively evaluate the merits of the proposed approach. They are chosen due to their popularity and competitive performance among the existing approaches.

For each approach, we evaluate the identification performance for main effects (M) and interactions (I) separately, by the number of true positives M:TP and I:TP and the number of false positives M:FP and I:FP. In addition, the root of the sum squared error $||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0||_2$ (RSSE) is used to assess the estimation accuracy, where $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^0$ are the estimated and true values of $\boldsymbol{\theta}$. We also examine the prediction performance using an independent testing set with 100 subjects under the same simulation scenarios. We adopt the prediction mean squared error (PMSE) for continuous outcome and C-statistic (Cstat) for survival outcome. The C-statistic quantifies the overall adequacy of risk prediction for censored survival data based on the time-integrated AUC (area under curve), where a larger value indicates better prediction (Uno et al., 2011).

For each scenario, 200 replicates are simulated, and summary statistics (mean and standard deviation) are computed. Summary results for the scenarios with continuous G factors and AR structure under linear and AFT models are shown in Tables 2.5 and 2.6, respectively. The rest of the results are provided in Appendix. The proposed LTS-MCP-Hier is observed to have competitive performance under all simulation scenarios. For the dataset without contamination (D1), the proposed approach can achieve satisfactory efficiency that is comparable to the nonrobust LS-MCP-Hier, and outperforms the robust alternatives and even nonrobust LS-MCP. The majority of true positives are identified by the proposed approach while with a small number of false positives. The advantage of the proposed approach over the alternatives becomes prominent for the datasets with different types of contaminations. For example, for the scenario with outliers in predictors (D4) under lin-

ear model (Table 2.5), the proposed approach has (M:TP, M:FP, I:TP, I:FP)=(7.4, 3.8, 11.1, 2.7), compared to (1.4, 22.6, 3.1, 68.0), (4.1, 4.0, 4.2, 13.4), (7.2, 0.7, 6.9, 11.6), (6.2, 7.9, 10.0, 30.1), and (5.4, 54.5, 3.9, 5.4) for LS-MCP, LAD-LASSO, RLARS, LTS-MCP and LS-MCP-Hier, respectively. The superior identification performance of the proposed approach over LAD-LASSO and RLARS provides a strong support to the proposed trimming strategy for accommodating outliers. In addition, it performs better than LTS-MCP, which suggests that the "coefficient decomposition" and stability selection framework can improve the identification of both main effects and interactions. The proposed approach also behaves better in terms of estimation and prediction. For example, for the scenario with contamination type D2 under AFT model (Table 2.6), the proposed approach has (ESSE, Cstat)=(2.71, 0.89), compared to (46.11, 0.55), (4.11, 0.74), (4.83,0.73), (3.71,0.82), and (59.00,0.58) for LS-MCP, LAD-LASSO, RLARS, LTS-MCP and LS-MCP-Hier, respectively. For the datasets with categorical G variables, the similar pattern is observed that the proposed approach demonstrates superior or comparable performance compared to five alternatives in identification, estimation and prediction accuracy.

In practical genetic interaction analyses, the important interactions may have different magnitude of signals, including those with weak but nonzero effects (Gao et al., 2017). To be thorough, we also examine the scenarios with both moderately large and weak effects. Specifically, we consider data with continuously distributed G factors and AR correlation structure, and with a continuous outcome under the linear regression model. The simulation settings for coefficients are similar to those in (c) as mentioned above. One different is that seven of the fourteen important interactions are with weaker signals equal to 0.2. Results with five types of data contaminations are shown in Table A.24. It can be seen that the performance of all approaches decay compared to those in Table 2.5. However, the proposed approach is again observed to have favorable performance. For example, under the scenario with D4, the values of (I:TP, I:FP) for interactions are (7.7, 1.4) (proposed), (2.3, 69.4) (LS-MCP), (3.2, 14.2) (LAD-LASSO), (5.0, 10.1) (RLARS), (7.2, 27.0) (LTS-MCP), and (3.7, 5.1) (LS-MCP-Hier).

In the interaction analysis literature, it has been suggested that there may exist important interactions in the absence of the corresponding main effects (Thomas, 2010). For

Table 2.5: Summary results under simulation scenarios with continuous G factors and AR structure under linear model. In each cell, mean (sd) based on 200 replicates.

	M:TP	M:FP	I:TP	I:FP	RSSE	PMSE
			D1: $N(0,1)$			
LTS-MCP-Hier	7.8(0.4)	0.6(1.6)	12.7(1.9)	0.7(0.8)	2.15(0.49)	0.99(0.43)
LS-MCP	5.7(0.9)	3.0(3.5)	10.8(0.9)	10.7(10.7)	2.80(0.41)	1.29(0.56)
LAD-Lasso	8.0(0.0)	10.6(5.6)	13.3(1.2)	28.0(11.4)	1.68(0.33)	1.35(0.45)
RLARS	7.5(0.6)	0.5(0.8)	7.3(1.9)	12.5(8.2)	3.27(0.42)	2.51(0.91)
LTS-MCP	6.4(0.9)	6.9(2.8)	11.0(1.1)	26.4(7.4)	2.39(0.53)	1.23(0.28)
LS-MCP-Hier	8.0(0.0)	0.3(1.2)	13.0(1.0)	0.4(0.6)	1.70(0.30)	0.80(0.18)
			D2: $0.9N(0,1)+0.1Cauchy(0,5)$			
LTS-MCP-Hier	7.9(0.3)	0.6(1.8)	12.0(1.5)	0.9(0.9)	2.12(0.38)	1.12(0.34)
LS-MCP	2.2(1.8)	18.0(8.0)	2.7(2.6)	71.0(10.6)	30.42(40.46)	555.38(1853.39)
LAD-Lasso	7.8(0.5)	2.2(1.5)	7.6(2.3)	7.0(3.3)	3.01(0.36)	3.85(1.23)
RLARS	7.2(0.7)	0.7(1.0)	5.7(1.8)	11.0(5.7)	3.68(0.41)	3.55(1.24)
LTS-MCP	6.2(1.1)	7.8(3.3)	10.6(1.3)	30.9(9.7)	2.55(0.55)	1.18(0.32)
LS-MCP-Hier	5.8(1.5)	151.3(125.9)	2.6(3.4)	25.6(59.8)	28.80(42.28)	1351.47(5973.38)
			D3: $0.9N(0,1)+0.1N(20,1)$			
LTS-MCP-Hier	7.9(0.3)	0.6(1.8)	12.0(1.6)	0.9(0.8)	2.01(0.41)	1.03(0.40)
LS-MCP	2.9(1.2)	24.3(4.7)	3.1(1.4)	66.2(5.5)	9.82(0.68)	32.66(6.95)
LAD-Lasso	7.5(0.7)	2.6(1.7)	6.1(2.3)	8.2(3.2)	3.29(0.33)	4.68(1.46)
RLARS	6.3(1.0)	1.4(1.5)	3.8(1.7)	11.7(5.8)	4.25(0.48)	5.23(1.79)
LTS-MCP	6.4(1.0)	7.6(3.0)	10.9(1.1)	28.3(6.2)	2.44(0.53)	1.09(0.27)
LS-MCP-Hier	6.5(0.9)	94.1(5.9)	2.4(1.5)	5.8(5.6)	8.81(0.64)	33.23(7.21)
			D4: $N(0,1)$ and with leverage points			
LTS-MCP-Hier	7.4(1.0)	3.8(8.0)	11.1(3.1)	2.7(2.1)	2.12(0.79)	1.08(2.02)
LS-MCP	1.4(0.9)	22.6(5.1)	3.1(2.0)	68.0(6.4)	7.38(1.03)	19.15(6.64)
LAD-Lasso	4.1(1.3)	4.0(2.4)	4.2(2.2)	13.4(3.6)	3.99(0.35)	9.27(2.47)
RLARS	7.2(0.8)	0.7(1.2)	6.9(2.0)	11.6(7.1)	3.42(0.34)	2.92(0.91)
LTS-MCP	6.2(1.2)	7.9(3.6)	10.0(1.3)	30.1(9.2)	2.47(0.60)	2.43(0.40)
LS-MCP-Hier	5.4(1.5)	54.5(37.3)	3.9(3.2)	5.4(3.1)	5.52(1.39)	14.61(9.16)
			D5: $0.9N(0,1)+0.1Cauchy(0,5)$ and with leverage points			
LTS-MCP-Hier	7.7(0.7)	3.4(8.4)	10.6(2.6)	2.3(2.3)	2.20(0.75)	1.02(1.77)
LS-MCP	0.7(0.7)	18.0(8.9)	1.5(1.4)	69.3(17.5)	25.80(32.38)	271.98(796.53)
LAD-Lasso	3.8(1.4)	4.0(1.9)	4.0(2.0)	12.5(3.5)	4.02(0.36)	9.20(2.59)
RLARS	6.8(0.9)	0.9(1.2)	5.6(2.0)	11.4(6.8)	3.79(0.42)	3.77(1.04)
LTS-MCP	6.3(1.1)	8.6(3.9)	10.8(1.2)	31.9(10.1)	2.47(0.57)	2.05(0.32)
LS-MCP-Hier	4.5(1.5)	152.6(99.4)	1.0(1.6)	24.6(62.0)	27.97(39.71)	1088.91(4898.79)

Table 2.6: Summary results under simulation scenarios with continuous G factors and AR structure under AFT model. In each cell, mean (sd) based on 200 replicates.

	M:TP	M:FP	I:TP	I:FP	RSSE	Cstat
			D1: $N(0,1)$			
LTS-MCP-Hier	7.8(0.5)	5.6(9.6)	11.0(2.7)	0.9(1.1)	2.48(0.52)	0.90(0.03)
LS-MCP	6.3(1.1)	12.3(4.8)	11.3(1.2)	38.3(9.6)	2.39(0.66)	0.92(0.02)
LAD-Lasso	7.5(0.8)	15.5(8.1)	8.4(3.9)	36.5(15.7)	3.06(0.59)	0.85(0.05)
RLARS	7.3(0.7)	10.3(3.8)	3.2(1.6)	21.8(4.3)	4.22(0.35)	0.78(0.04)
LTS-MCP	6.0(1.0)	14.8(4.9)	6.4(1.8)	57.4(10.4)	3.37(0.34)	0.85(0.04)
LS-MCP-Hier	8.0(0.2)	1.0(1.9)	12.1(1.5)	0.6(0.8)	1.94(0.34)	0.92(0.02)
			D2: $0.9N(0,1)+0.1Cauchy(0,5)$			
LTS-MCP-Hier	7.7(0.5)	5.5(4.1)	9.1(2.8)	1.3(1.1)	2.71(0.58)	0.89(0.03)
LS-MCP	1.2(1.4)	13.9(8.6)	1.1(1.4)	59.6(10.7)	46.11(87.78)	0.55(0.07)
LAD-Lasso	5.8(1.7)	4.6(2.1)	1.7(1.3)	12.0(3.4)	4.11(0.40)	0.74(0.07)
RLARS	6.3(1.6)	7.6(4.9)	1.6(1.3)	22.9(6.3)	4.83(0.68)	0.73(0.06)
LTS-MCP	6.0(1.0)	15.4(4.0)	5.5(1.8)	59.7(5.5)	3.71(0.33)	0.82(0.03)
LS-MCP-Hier	5.4(1.5)	196.6(162.2)	2.0(2.4)	70.9(223.4)	59.00(119.77)	0.58(0.07)
			D3: $0.9N(0,1)+0.1N(20,1)$			
LTS-MCP-Hier	8.0(0.2)	2.2(4.8)	11.9(1.6)	0.9(0.9)	2.01(0.38)	0.92(0.01)
LS-MCP	2.5(1.1)	24.6(4.9)	2.4(1.4)	72.2(6.1)	10.72(0.71)	0.64(0.04)
LAD-Lasso	6.6(1.2)	3.9(2.2)	2.7(1.6)	11.1(3.3)	3.79(0.28)	0.78(0.04)
RLARS	6.4(1.0)	4.2(3.2)	1.4(1.1)	12.4(6.3)	4.41(0.42)	0.78(0.04)
LTS-MCP	6.1(1.0)	11.4(4.1)	9.0(1.7)	48.9(10.3)	2.95(0.49)	0.89(0.02)
LS-MCP-Hier	5.8(1.1)	100.5(7.8)	2.5(1.5)	8.3(7.2)	9.75(0.56)	0.66(0.03)
			D4: $N(0,1)$ and with leverage points			
LTS-MCP-Hier	7.1(1.0)	10.9(14.7)	9.0(4.0)	1.2(1.2)	3.18(0.83)	0.84(0.07)
LS-MCP	3.4(1.1)	14.7(4.5)	4.9(2.1)	52.7(6.5)	4.89(0.60)	0.75(0.05)
LAD-Lasso	6.1(1.2)	7.2(5.0)	3.4(2.0)	17.8(11.8)	3.88(0.31)	0.77(0.04)
RLARS	7.0(0.8)	11.9(3.6)	2.6(1.4)	21.5(4.5)	4.37(0.36)	0.77(0.04)
LTS-MCP	5.5(1.3)	17.0(4.0)	5.2(1.8)	61.4(6.0)	3.77(0.42)	0.81(0.04)
LS-MCP-Hier	6.4(1.0)	42.4(24.3)	4.6(2.5)	2.9(2.1)	4.08(0.66)	0.78(0.05)
			D5: $0.9N(0,1)+0.1Cauchy(0,5)$ and with leverage points			
LTS-MCP-Hier	7.1(1.1)	12.9(14.1)	9.3(3.9)	1.5(1.4)	3.08(0.81)	0.85(0.07)
LS-MCP	1.1(1.1)	12.6(7.8)	1.3(1.3)	56.3(9.8)	35.96(69.84)	0.56(0.06)
LAD-Lasso	5.7(1.5)	4.3(2.3)	2.0(1.5)	12.2(3.4)	4.12(0.36)	0.74(0.06)
RLARS	6.5(1.4)	8.8(4.6)	2.2(1.5)	21.6(6.4)	4.79(1.30)	0.74(0.06)
LTS-MCP	5.7(1.1)	16.1(4.2)	5.1(2.0)	60.4(4.6)	3.77(0.37)	0.81(0.04)
LS-MCP-Hier	5.1(1.6)	174.4(158.2)	2.4(2.6)	67.4(229.7)	54.36(131.61)	0.57(0.07)

comprehensive consideration, we conduct another analysis on scenarios where the "main effects, interactions" hierarchy is violated for some interactions. Specifically, data with continuous G factors, AR correlation structure, and a continuous response are generated. Besides the fourteen nonzero G-E interactions as described above, six additional nonzero interactions are considered without the corresponding main G effects. As shown in Table A.25, the proposed approach performs slightly worse than LTS-MCP which is similar to the proposed but does not reinforce the hierarchy. However, it still outperforms other alternatives, including two nonrobust approaches LS-MCP and LS-MCP-Hier, and two robust ones LAD-Lasso and RLARS which do not respect the hierarchy and may be favored here.

2.3.4 Data Analysis

The Cancer Genome Altas provides comprehensive profiling data in various cancer types. With high quality and public availability, the TCGA data have contributed to thousands of genetic studies and serve us as an ideal testbed. In this section, we analyze TCGA data on skin cutaneous melanoma (SKCM) and breast invasive carcinoma (BRCA). The processed level 3 data are considered which can be downloaded from TCGA Provisional using the R package *cgdsr*.

Skin Cutaneous Melanoma (SKCM) Data

Cutaneous melanoma, the most dangerous type of skin cancer, has been demonstrated to account for approximately 75% of all deaths from skin cancer. The response of interest is the continuous (log2-transformed) Breslow's depth, which is analyzed using a linear model. It describes the thickness of the tumor, which is considered as one of the most significant factors in predicting progression of melanoma (Breslow, 1970). For E variables, we include age, American Joint Committee on Cancer (AJCC) tumor pathologic stage, gender, and Clark level. For G variables, we consider mRNA gene expressions, which are collected using the IlluminaHiseq RNAseq V2 platform and have been lowess-normalized, log-transformed, and median centered. There are 298 subjects available with 18,355 measurements of gene expressions. We conduct a simple prescreening as the number of cancer-related genes is not expected to be large, which selects the top 2,000 genes with the largest variances across all

the samples for downstream analyses.

The estimated coefficients with the proposed approach are listed in Table 2.7. Compared to age and gender, stage and Clark level are more relevant to the Breslow's depth, which is consistent with the literature. The proposed approach identifies a total of 43 important genes and 26 G-E interactions associated with Breslow's depth. Existing literature shows potentially useful implications of our findings. For instance, gene FGFR3 has been shown to deactivate the malignant transformation as a tumor suppressor in melanoma cancer cells. An increased expression of antigen from gene FMR1NB has been found in melanoma stem cells, which may be a cause of treatment failure. Gene LAMP1 has been observed to express on the surface of metastatic melanoma cells, and its downregulation could reduce lung metastasis. Gene SPRR1A has been found to express dramatically higher levels in thin melanomas. In addition, gene SPRR2G has been characterized as keratinocyte-associated and has been found to have decreased expression in the primary melanoma. Gene S100A7, known as psoriasin, has been observed to significantly over-express in human epithelial skin tumors, as well as in breast and bladder cancer.

We also analyze the data using the alternatives, and the comparison results are summarized in Table A.26. The numbers of overlapping identifications of main effects and interactions are presented, respectively, along with the corresponding RV-coefficients (Smilde et al., 2009). The RV-coefficient evaluates the similarity of two data matrices with a larger value indicating a higher degree of similarity. It is observed that significantly different sets of main effects and interactions are found by different approaches with moderate RV-coefficients. LS-MCP, LAD-LASSO, RLARS and LTS-MCP, which do not reinforce the hierarchical structure, identify much smaller number of main effects compared to that of interactions. Both LTS-MCP-Hier and LS-MCP-Hier identify a moderate number of main effects and interactions.

To provide an indirect support to the identification analysis, we evaluate the prediction accuracy using PMSE based on 200 times resampling (9/10 training subjects and 1/10 testing subjects), which has also been adopted in the literature. The proposed approach is observed to have the best prediction performance with PMSE=0.26, compared to 1.01 (LS-MCP), 0.32 (LAD-LASSO), 0.49 (RLARS), 0.87 (LTS-MCP) and 0.58 (LS-MCP-Hier).

We also examine the selection stability by calculating the observed occurrence index (OOI) (Huang et al., 2006). Using the same resampling strategy, the OOI measures the identified probability for each main effect or interaction, where a larger value indicates better stability in identification among random samples. The mean OOI of the identified main effect and interactions using the proposed approach is 0.85, compared to 0.32 (LS-MCP), 0.81 (LAD-LASSO), 0.50 (RLARS), 0.10 (LTS-MCP) and 0.81 (LS-MCP-Hier), suggesting satisfactory stability of the proposed approach.

Breast Invasive Carcinoma (BRCA) Data

Breast cancer is the second cause of cancer death among female, which can be influenced by a number of environmental and genetic factors (Shipitsin et al., 2007). The response of interest is the censored survival time, which is analyzed based on AFT model. In this section, we focus on the female Whites with primary tumor. Data are available on 353 subjects, with 60 deaths during the follow-up period. For E variables, we include age, AJCC tumor pathologic stage, ER status (positive/negative) and weight. For G variables, there are 16,277 measurements of mRNA expressions and the top 2,000 genes are selected for the downstream analyses using the same prescreening as described in the previous section.

The coefficients estimated from the proposed approach are provided in Table 2.8. The three E variables age, stage and weight have negative coefficients, indicating that higher levels are associated with shorter survival, and the positive coefficient of ER status suggests that the subjects with negative ER status tend to have better prognosis. In addition, there are 32 important main effects along with 43 interactions. These findings are validated by the literature search. For example, gene ASH2L has been shown to be over-expressed in human breast cancer among other candidate oncogenes. Gene ATAD1 has been found to be down-regulated in different subtypes of breast tumors in gene expression profiling, whose interactions with age, tumor stage and ER status are identified using the proposed approach. Abnormal expression of gene FGF4 has been found in human breast cancer cells, and the up-regulation of endogenous FGF4 expression indicates its biological significance in tumorigenesis. Gene KAT6A has been suggested to be a novel oncogene in breast cancer as a chromatin modifier. Gene MED1 has been demonstrated a key role in tamoxifen

Table 2.7: Analysis of SKCM data using the proposed approach: coefficients of identified main effects and interactions

	Main:G	Age	Stage	Gender	Clark level
Main:E		-0.0100	1.2197	-0.0587	0.3307
AADACL3	0.0004				
AMBN	0.0005				
ATP1A2	-0.0011				
BCAR4	0.0004				
BPIFA2	0.0001				
C7ORF69	0.0038		0.0046		
C8ORF34	0.0056		0.0101		
CALCA	0.0029	0.0010	0.0008		
CLNS1A	0.0066		0.0118	0.0020	
CNBD2	0.0011				
CYP1A2	0.0008				
CYP7A1	0.0031		0.0025		
DEFA5	0.0056		0.0100		
DEFB4A	0.0023		0.0016		
DGKB	-0.0029	0.0027			
DGKK	0.0029		0.0018		
DPRX	0.0018		0.0004		
FAM131B	-0.0025		-0.0014		
FAM9B	0.0028	0.0020			
FGF4	0.0006				
FGFR3	0.0026		0.0015		
FMR1NB	0.0038		0.0042		
GLYATL3	-0.0006				
IFNA14	-0.0004				
IL17A	0.0012				
KRT16	0.0065		0.0124		
LAMP1	0.0010				
LCE3C	0.0002				
LPO	0.0001				
MEP1A	0.0029		0.0019		
NPS	-0.0006				
OR2V2	-0.0002				
OR5M8	0.0011				
PHOX2B	-0.0026		-0.0018		
RETNLB	-0.0028	-0.0004			
RIIAD1	0.0079	0.0103	0.0111		
S100A7	-0.0006				
S100A7A	-0.0003				
SEMG2	0.0002				
SPINK9	-0.0049		-0.0046		
SPRR1A	-0.0026			-0.0003	
SPRR2G	0.0011				0.0003
TRIM55	-0.0019			-0.0010	

resistance of human breast cancer cells, suggesting its potential as a therapeutic target in cancer treatment. Over-expression of gene MTBP has been observed to be strongly correlated with reduced breast cancer patient survival. Gene NSD3 has been showed to be amplification in primary breast carcinomas, suggesting a possible involvement in human tumorigenesis. Gene PHB2 has been demonstrated to play a crucial role in modulation of ER status in breast cancer cells.

Data are also analyzed using the alternatives. The summary results of comparison are shown in Table A.27. Small numbers of overlapping main effects and interactions are found across different approaches, whereas moderate common information is contained among different identifications given the values of RV-coefficients. We also compute C-statistics to evaluate the prediction accuracy of survival response using the same resampling process. The proposed approach demonstrates improved prediction ability with a C-statistic value of 0.55, compared to 0.49 (LS-MCP), 0.49 (LAD-LASSO), 0.47 (RLARS), 0.51 (LTS-MCP) and 0.47 (LS-MCP-Hier). In addition, the proposed approach has better stability with the average OOI as 0.49, compared to 0.09 (LS-MCP), 0.43 (LAD-LASSO), 0.27 (RLARS), 0.08 (LTS-MCP) and 0.4 (LS-MCP-Hier). The improved prediction and stability confirm the validity of the proposed analysis.

2.3.5 Discussion

Identifying important G-E interactions associated with complex multifactorial human diseases is an important goal of high-dimensional cancer studies. In this Chapter, we propose a novel effective interaction analysis approach based on the least trimmed regression. The proposed approach can accommodate the vertical outliers as well as the leverage points, which are not uncommon in practice but have not been well studied. It differs significantly from the existing robust interaction analyses that usually focus on model mis-specification or outliers/contaminations in response. A robust criterion based on the (weighted) residuals is developed for choosing the optimal number of outliers, which can accommodate multiple types of responses, such as continuous biomarkers and censored survival time. The coefficient of each interaction is decomposed as the product of the corresponding main effect and interaction-specific coefficient, which has an intuitive formulation to automatically

Table 2.8: Analysis of BRCA data using the proposed approach: coefficients of identified main effects and interactions

	Main:G	Age	Stage	ER status	Weight
Main:E		-0.1594	-0.1089	0.2705	-0.1219
AASDHPPT	0.0885	0.0347		-0.0069	
ASH2L	0.0006				
ATAD1	0.1274	0.0058	0.0016	-0.0078	
AXDND1	-0.1061				-0.0094
BRD1	0.0293				
CCT6A	-0.0701	-0.0076			
CD5L	-0.0776				
FGF4	0.0292				
ITLN2	-0.1221	-0.0113			0.0069
KAT6A	0.0123				
MAEA	0.0453				
MED1	-0.0649	-0.0226	-0.0254	-0.0013	-0.0058
MRPL45	0.0512				-0.0013
MTBP	0.0127				
NARS2	0.0197				
NSD3	0.0112				
NUFIP2	-0.0297				
PHB	0.0984	0.0015		0.0008	0.0005
PHB2	0.0832			-0.0032	0.0025
PMVK	0.1227	0.0064	-0.0016	-0.0216	-0.0564
RAD21	-0.0555	-0.0311			
SEZ6	-0.1450	-0.0320	-0.0017		
SMIM19	0.0950	0.0379	0.0127	0.0022	-0.0136
SUPT4H1	-0.1278		0.0127		0.0027
SUPT5H	-0.0240				
TBC1D21	-0.0571				
TBC1D23	-0.0526				
TRIM11	-0.1352	-0.0314			0.0071
UBE2Z	0.0895	-0.0003	-0.0031		0.0002
UBE4A	-0.0055				
ZNF572	0.0053				
ZNF597	0.0932	0.0065	0.0205		

respect the strong hierarchical structure. The modified stability selection-based C-steps algorithm and iterative coordinate descent algorithm are adopted to optimize the objective function, which leads to the estimation of main effects and interactions as well as the optimal outlier-free subject set. Extensive simulations are conducted, including various scenarios without data contamination, with vertical outliers, and with leverage points. The results demonstrate the competitive performance of the proposed analysis in terms of identification, estimation and prediction. In the data analysis of cutaneous melanoma and breast invasive carcinoma with gene expression measurements, the proposed approach identifies biologically sensible markers with better prediction performance and stability.

In this Section, we have considered a continuous response under the linear model and a censored survival time under the AFT model. For the categorical and count data under generalized linear models, the iterated weighted squared loss can be adopted as an approximation to the negative log-likelihood. Thus, with minor modifications, the proposed approach can be extended to accommodate other types of responses. The proposed approach is built on the trimmed regression which has been demonstrated to have solid statistical properties for the analysis of low-dimensional data and high-dimensional main effects. Thus it may be reasonable to conjecture that the proposed approach also has good theoretical properties. The detailed study is postponed to future research. In simulation, we focus on the leverage points in G factors, more extensive numerical studies with outliers in E factors are deferred to future investigation. In data analysis, more biological and functional analyses are needed to provide more evidence of the identified interactions.

Chapter 3

Incorporating Additional Information Using Marginal Penalized Regression for Interaction Identification

3.1 Introduction

Many statistical methods have been proposed for identifying G-E interactions, among which marginal modeling framework becomes more popular due to less computation and simpler interpretation (Sun et al., 2018b; Xu et al., 2019; Zhang et al., 2019). Though marginal analysis is computationally simpler, the "main effects, interactions" hierarchy is not automatically guaranteed, leading to difficult interpretation. That is, an interaction term may be identified due to a significant p-value, but the corresponding main effects are not. Comparatively, the joint analysis that models a large number of genetic factors and interactions in a single model, the importance of respecting such hierarchical structure for producing statistically and biologically meaningful findings have been demonstrated (Bien et al., 2013; Hao and Zhang, 2017) and a few approaches have been developed such as Shi et al. (2014) and Zhu et al. (2014). In the current literature, the hierarchy structure in marginal analysis

shares equal importance yet has been less studied Bien et al. (2015).

Considering the high dimensionality and low signal levels, it becomes more challenging to identify important G-E interactions beyond the main effects without sufficient information. In recent literature, incorporating additional information for main effect analysis have been adopted to facilitate effective and biologically meaningful discoveries for complex diseases. For example, consider the adjacency structure of SNPs. Due to linkage disequilibrium, SNPs that are physically close can demonstrate similar associations with the disease outcomes (Ardlie et al., 2002). Multiple statistical methods using penalization have been proposed that account for high correlations among closely located markers, including fused lasso (Tibshirani et al., 2005), smooth lasso (Hebiri et al., 2011), spline lasso (Guo et al., 2016), and so on. Extensive research has shown that combining biological knowledge as a priori can lead to more accurate and interpretable estimation and identification. Yet, almost all of the existing G-E interaction analyses omit such biological knowledge. Another example of additional information arises from published studies. Literature review suggests that, for many common problems in the field of biology and biomedicine, multiple relevant investigations were conducted and published, which may provide valuable and comprehensive input for the current study. To incorporate existing studies, meta-analysis or integrative analysis can be conducted. We refer Zeggini et al. (2008), Guerra and Goldstein (2009), and Ma et al. (2011) for further discussion. Despite the great achievement, such analysis procedure requires highly comparable design across available studies and datasets. Excluding partially relevant ones may cause a waste of information, and it is desired to include as many related studies as possible to add to G-E interaction analysis.

In this Chapter, our goal is to incorporate additional information for G-E interaction analysis. Motivated by the lack of information in G-E interaction analysis and the success of utilizing available information in the main effect analysis, we propose a new G-E interaction analysis method under a marginal modeling framework. Using penalized regression, the proposed method respects the 'main effects, interactions' hierarchical structure. That is, when an interaction is identified, the corresponding main effect of the genetic factor is simultaneously included in the model. In addition, significantly advancing from the existing G-E interaction analysis, the proposed method can incorporate additional infor-

mation, especially including the adjacency structure of SNPs and mined data extracted from relevant literature. We propose using penalization to address the lack of information problem, which provides a coherent formulation for multiple types of additional information and genetic measurements. Different from the meta-analysis and integrative analysis framework, strict comparability across studies is not required, which allows more comprehensive information to be included. This advancement of utilizing additional information enables improved performance in identification and interpretation for G-E interaction analysis. Our numerical study shows that the proposed method can outperform multiple direct alternatives. Overall, this study provides an effective and practically meaningful way to incorporate additional information for G-E interaction analysis.

3.2 Methods

Assume N iid subjects. Let $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ and $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_q)$ be $N \times p$ and $N \times q$ matrix of gene expressions and environmental factors. Denote $\mathbf{Y}$ as a length N vector of continuous disease outcome. We consider the regression model for j^{th} gene, $j = 1, 2, \ldots, p$, $Y = \sum_{k=1}^{q} \alpha_{kj} Z_k + \beta_j X_j + \sum_{k=1}^{q} \eta_{kj} Z_k X_j + \epsilon$, where α_k, β_j, and η_{kj} are the coefficients for environmental factors, gene expressions, and their interactions. ϵ is the random errors. We decompose η_{kj} to impose the hierarchical structure of main effects and interactions by using $\eta_{kj} = \beta_j \gamma_{kj}$. Then, the marginal model for j^{th} gene can be written as

$$Y = \sum_{k=1}^{q} \alpha_{kj} Z_k + \beta_j X_j + \sum_{k=1}^{q} \beta_j \gamma_{kj} Z_k X_j + \epsilon.$$

Consider the following objective function

$$Q(\boldsymbol{\theta}) = \frac{1}{p} \sum_{j=1}^{p} \frac{1}{2N} ||Y - \sum_{k=1}^{q} \alpha_{kj} Z_k - \beta_j X_j - \sum_{k=1}^{q} \beta_j \gamma_{kj} Z_k X_j||_2^2 + \sum_{j=1}^{p} \rho(|\beta_j|; \lambda_1, r) + \sum_{j=1}^{p} \sum_{k=1}^{q} \rho(|\gamma_{kj}|; \lambda_1, r) + \frac{1}{2} \lambda_2 \boldsymbol{\beta}' \mathbf{J} \boldsymbol{\beta} + \frac{1}{2} \lambda_2 \sum_{k=1}^{q} \boldsymbol{\gamma}_{\boldsymbol{k}}' \mathbf{J} \boldsymbol{\gamma}_{\boldsymbol{k}}, \tag{3.1}$$

where $\boldsymbol{\theta} = (\alpha_{11}, \ldots, \alpha_{qp}, \boldsymbol{\beta}', \boldsymbol{\gamma}_1', \ldots, \boldsymbol{\gamma}_q')$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$, $\boldsymbol{\gamma}_k = (\gamma_{k1}, \gamma_{k2}, \ldots, \gamma_{kp})'$ for

$k = 1, 2, \ldots, q$. $||\cdot||_2$ is the Euclidean norm, $\rho(|v|; \lambda_1, r) = \lambda_1 \int_0^{|v|} \left(1 - \frac{x}{\lambda_1 r}\right)_+ dx$ is the minimax concave penalty, $\lambda_1, \lambda_2 \geq 0$ are tuning parameters, $r > 0$ is the regularization parameter. $\mathbf{J}$ is a $p \times p$ matrix for Laplacian quadratic penalty, which is tailored for different types of additional information. More details are discussed below. We obtain the proposed estimates that minimizes equation (3.1) as $\hat{\boldsymbol{\theta}} = \arg\min Q(\boldsymbol{\theta})$, and important main genetic effects and interactions are identified by non-zero estimated coefficients.

The proposed objective function is designed under a marginal analysis framework. For each genetic measurement, one regression model is assumed. Decomposing the coefficient of interactions as $\beta_j \gamma_{kj}$, the 'main effects, interactions' hierarchical structure is ensured. Note that environmental factors are pre-selected and have a low dimensionality so that their coefficients are not subject to penalized selection. The proposed method thus differs from the pairwise interaction analysis such as (Choi et al., 2010). The first two minimax concave penalty (MCP) terms in 3.1 is applied, which guarantees that interaction and its corresponding main effects can be selected simultaneously. Without further computational burden, we impose the same tuning parameter across different genes using λ_1 to ensure comparability. In the literature, the MCP-based penalty has been extensively adopted (Kim et al., 2017; Zhang et al., 2010), and many other ways exist for achieving hierarchy, such as the sparse group MCP (Liu et al., 2013). Our investigation suggests the proposed method has computational advantages and satisfactory performance.

We adopt the Laplacian quadratic penalty to incorporate additional information as the last two terms. In this Chapter, we consider two specific examples. (1) Consider the adjacency structure of SNPs as additional information, and assume that SNP measurements are ordered by their physical locations. We adopt the spline type penalty for main genetic effects and interactions as $\sum_{j=2}^{p-1}[(\beta_{j+1} - \beta_j) - (\beta_j - \beta_{j-1})]^2$ and $\sum_{j=2}^{p-1}[(\gamma_{k(j+1)} - \gamma_{kj}) - (\gamma_{jk} - \gamma_{k(j-1)})]^2$. Then, for SNP data, we have $\mathbf{J} = \mathbf{H}'_{(p-2)\times p}\mathbf{H}_{(p-2)\times p}$ where $H_{jj} = H_{j(j+2)} = 1, H_{j(j+1)} = -2$, and 0 otherwise. This penalty encourages smoothness and is analogous to penalize second-order derivatives in spline-based nonparametric estimation. Consequently, the main effects and interactions of physically adjacent SNPs associated with the response are promoted to be similar. Other alternative penalties, such

as the fused lasso and smooth lasso are available. We choose the spline type penalty in this Chapter due to its demonstrated superior performance and computational feasibility (Guo et al., 2016). (2) Consider text-based literature mining data of PubMed as additional information. We adopt PubMatrix (`https://pubmatrix.irp.nia.nih.gov`), which is a web-based tool that allows simple text-based mining of PubMed and has been used in the studies of Wang et al. (2019), Minafra et al. (2018), and many others. PubMatrix uses two lists of keywords and produces a frequency matrix of term co-occurrence as results (Becker et al., 2003). Consider gene expression data and we utilize gene names as keywords. The pairwise frequency matrix is generated by PubMatrix, each element of which suggests not only whether an association exists but also its amount of evidence. Given the fact that the majority of the frequency counts are zero, we construct the adjacency matrix $\mathbf{A} = \{a_{jl}\}_{p\times p}$ using quantiles at $0.2, 0.4, 0.6,$ and 0.8 of the nonzero frequencies. In this way, the magnitude in $\mathbf{A}$ is managed as extreme values are excluded. Then, consider $\mathbf{J} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ where $\mathbf{I}$ is the $p \times p$ identity matrix and $\mathbf{D} = \text{diag}(\sum_{l=1}^{p} a_{1l}, \sum_{l=1}^{p} a_{2l}, \ldots, \sum_{l=1}^{p} a_{pl})$. This penalty promotes similar main genetic effects and interactions for those genes that have demonstrated more co-occurrences as pairs in the existing publications.

We also note that other designs of constructing $\mathbf{J}$ can be tailored given the type of data and additional information. For instance, for gene expression levels, the adjacency matrix can be calculated based on Pearson correlation coefficients, similarity measure such as the Euclidean distance, and others alternative approaches. Recent studies have established the improved performance and effectiveness of the Laplacian quadratic penalty in main effect analysis. However, limited adoption for analyzing G-E interaction exists. We refer Huang et al. (2011a), Shi et al. (2015), and Wu et al. (2019b) for further discussion of construction of Laplacian quadratics.

3.2.1 Computation

To compute the proposed estimates, we adopt an iterative coordinate descent algorithm with fixed tuning parameters. This algorithm minimizes the objective function with respect to one coefficient at each step until convergence. We summarize the algorithm as follows.

1. Start with $\alpha_{kj}^{(0)} = \beta_j^{(0)} = \gamma_{kj}^{(0)} = 0$ for $k = 1, 2, \ldots, q$, and $j = 1, 2, \ldots, p$.

2. Let $t = t+1$. Update $\boldsymbol{\alpha}_j^{(t)} = (\alpha_{1j}^{(t)}, \ldots, \alpha_{qj}^{(t)})'$. Let $\tilde{Y}_j = Y - \beta_j^{(t-1)} X_j - \sum_{k=1}^{q} \beta_j^{(t-1)} \gamma_{kj}^{(t-1)} Z_k X_j$, $\tilde{X} = \mathbf{Z}$, then for $j = 1, 2, \ldots, p$,

$$\boldsymbol{\alpha}_j^{(t)} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y}_j.$$

3. Update $\boldsymbol{\beta}^{(t)}$. Let $\tilde{Y}_j = Y - \sum_{k=1}^{q} \alpha_{kj}^{(t)} Z_k$, $\tilde{X}_j = X_j + \sum_{k=1}^{q} \gamma_{kj}^{(t-1)} Z_k X_j$, then

$$\boldsymbol{\beta}^{(t)} = \arg\min \frac{1}{p} \sum_{j=1}^{p} \frac{1}{2N} ||\tilde{Y}_j - \beta_j \tilde{X}_j||_2^2 + \sum_{j=1}^{p} \rho(|\beta_j|; \lambda_1, r) + \frac{1}{2}\lambda_2 \boldsymbol{\beta}' \mathbf{J} \boldsymbol{\beta}.$$

For $j = 1, 2, \ldots, p$, write $\delta_j^{(t)} = \sum_{l=1}^{j-1} \beta_l^{(t)} J_{jl} + \sum_{l=j+1}^{p} \beta_l^{(t-1)} J_{jl}$, $\chi_j^{(t)} = \frac{1}{N}\tilde{X}_j'\tilde{X}_j$, and $\phi_j^{(t)} = \frac{1}{N}\tilde{X}_j'\tilde{Y}_j$,

$$\beta_j^{(t)} = \begin{cases} \frac{ST\left(\phi_j^{(t)} - \lambda_2 \delta_j^{(t)}, \lambda_1\right)}{\chi_j^{(t)} - \frac{1}{r} + \lambda_2 J_{jj}}, & \text{if } \left|\phi_j^{(t)} - \lambda_2 \delta_j^{(t)}\right| \leq \lambda_1 \left(\chi_j^{(t)} + \lambda_2 J_{jj}\right) \\ \frac{\phi_j^{(t)} - \lambda_2 \delta_j^{(t)}}{\chi_j^{(t)} + \lambda_2 J_{jj}}, & \text{if } \left|\phi_j^{(t)} - \lambda_2 \delta_j^{(t)}\right| > \lambda_1 \left(\chi_j^{(t)} + \lambda_2 J_{jj}\right) \end{cases} \tag{3.2}$$

where $ST(a, b) = sign(a)(|a| - b)_+$ is the soft-thresholding operator.

4. Update $\boldsymbol{\gamma}_{\boldsymbol{k}}^{(t)}$. Let $\tilde{Y}_j = Y - \sum_{k=1}^{q} \alpha_{kj}^{(t)} Z_k - \beta_j^{(t)} X_j$, $\tilde{X}_{kj} = \beta_j^{(t)} Z_k X_j$, then

$$(\boldsymbol{\gamma}_{\mathbf{1}}^{(t)}, \ldots, \boldsymbol{\gamma}_{\boldsymbol{q}}^{(t)}) = \arg\min \frac{1}{p} \sum_{j=1}^{p} \frac{1}{2N} ||\tilde{Y}_j - \sum_{k=1}^{q} \gamma_{kj} \tilde{X}_{kj}||_2^2 + \sum_{j=1}^{p} \sum_{k=1}^{q} \rho(|\gamma_{kj}|; \lambda_1, r) + \frac{1}{2}\lambda_2 \sum_{k=1}^{q} \boldsymbol{\gamma}_{\boldsymbol{k}}' \mathbf{J} \boldsymbol{\gamma}_{\boldsymbol{k}}.$$

For $k = 1, 2, \ldots, q$, $\boldsymbol{\gamma}_k = (\gamma_{k1}, \gamma_{k2}, \ldots, \gamma_{kp})'$ are calculated similar to (3.2), where $\delta_{kj}^{(t)} = \sum_{l=1}^{j-1} \gamma_{kl}^{(t)} J_{jl} + \sum_{l=j+1}^{p} \gamma_{kl}^{(t-1)} J_{jl}$, $\chi_{kj}^{(t)} = \frac{1}{N}\tilde{X}_{kj}'\tilde{X}_{kj}$, and $\phi_{kj}^{(t)} = \frac{1}{N}\tilde{X}_{kj}'(\tilde{Y}_j - \sum_{h=1}^{k-1} \gamma_{hj}^{(t)} \tilde{X}_{hj} - \sum_{h=k+1}^{q} \gamma_{hj}^{(t-1)} \tilde{X}_{hj})$.

5. Compute the relative difference as $\Delta^{(t)} = \frac{|Q(\boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t-1)})|}{|Q(\boldsymbol{\theta}^{(t-1)})|}$. Repeat Step 2-4 until $\Delta^{(t)} < 10^{-4}$.

For model selection, we set r as 3 to reduce computational cost and adopt the extended Bayesian information criterion to choose the values of (λ_1, λ_2) (Chen and Chen, 2008). In the literature, convergence properties of coordinate descent have been well established and

we observe convergence in all of our numerical studies. The computational cost of the proposed method is moderate. We have developed R code and made it publicly available on GitHub.

3.3 Simulation

We set $N = 200$, $p = 1000$, and $q = 5$ for all simulated data. (a) For genetic data, we consider two types. (S1) We generate SNP data with adjacency structure to mimic densely positioned SNPs. Two approaches are adopted to simulate SNP data coded as $(0, 1, 2)$ for genotypes (aa, Aa, AA). For the first approach, we first generate p continuous variables using a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\Sigma = \{\sigma_{jl}\}_{p\times p}$ and then categorize them at q_1 and q_2 quantiles. We consider two correlation structures for Σ. The first one is auto-regressive structure (AR) with $\rho = 0.3$ and 0.5. The second is the banded correlation structure where two scenarios are considered. One has $\sigma_{jl} = 1$ if $j = l$, 0.3 if $|j - l| = 1$, and 0 otherwise (Band1). The other one is $\sigma_{jl} = 1$ if $j = l$, 0.5 if $|j - l| = 1$, 0.3 if $|j - l| = 2$, and 0 otherwise (Band2). We adjust q_1 and q_2 for minor allele frequency (MAF) values and conducer two scenarios. The first one (M1) has MAF = 0.05 with $q_1 = 0.91$ and $q_2 = 0.99$. The second scenario (M2) has MAF = 0.15 with $q_1 = 0.73$ and $q_2 = 0.97$. For the second approach of generating SNP data, we use pairwise LD structure with pairwise correlation $r_{LD} = 0.3$ and 0.5. Specifically, denote p_A and p_B as the MAFs of alleles A and B for two adjacent SNPs. Four haplotypes ab, aB, Ab, and AB have frequencies $(1 - p_A)(1 - p_B) - \phi$, $(1 - p_A)p_B - \phi$, $p_A(1 - p_B) - \phi$, and $p_Ap_B - \phi$ respectively, with $\phi = r_{LD}\sqrt{p_A(1 - p_A)p_B(1 - p_B)}$ and two scenarios of MAFs as in the first approach. (S2) We also simulate gene expression data. Among p expressions, C1 setting has 50 clusters with size 20 and C2 has 10 clusters with size 100. Genes in different clusters are independent whereas within each cluster, gene expressions are generated from a multivariate normal distribution using AR and banded correlation structures. We use same parameter settings for generating covariances. (b) Environmental risk are generated from a multivariate normal distribution with marginal mean 0, variances 1, and AR correlation ($\rho = 0.3$). (c) We set 20 main genetic factors and 40 hierarchical G-E interactions with

nonzero effects, generated from Uniform $(0.75, 1.25)$. The coefficients of environmental risk factors are generated from Uniform $(0.8, 1.2)$. (d) To simulate the response, we assume a joint model as $Y = \sum_{k=1}^{q} \alpha_{kj} Z_k + \sum_{j=1}^{p} \beta_j X_j + \sum_{k=1}^{q} \sum_{j=1}^{p} \beta_j \gamma_{kj} Z_k X_j + \epsilon$, where ϵ follows a standard normal distribution.

We analyze the simulated data using the proposed method. For S1, we consider the spline type penalty. For S2, two types of additional information are considered with different Laplacian quadratics. The first type (J1) is literature mining information. We select top 1000 genes in TCGA SKCM data based on marginal p-values and unitize the pairwise frequency matrix generated by PubMatrix as $\mathbf{J}$. The second type (J2) is the correlation structure of gene expressions, for which we construct $\mathbf{J}$ based Pearson correlation coefficients. We also consider the following alternative approaches for comparison. (1) HierMCP, which excludes the Laplacian quadratic penalty from the proposed objective function. (2)MCP-LP, which uses $Y = \sum_{k=1}^{q} \alpha_{kj} Z_k + \beta_j X_j + \sum_{k=1}^{q} \eta_{kj} Z_k X_j + \epsilon$ without decomposition. To estimate coefficients, MCP and Laplacian quadratics are applied to β_j and η_{kj} as the same as the proposed method. (3) Lasso, which imposes the Lasso penalty under a marginal modeling framework. The tuning parameter is selected by cross-validated mean squared error. (4) MA, which is the benchmark marginal analysis that analyzes one genetic factor at a time. P-values are adjusted by the false discovery rate (FDR) approach and important interactions are selected at FDR=0.1. To assess the identification accuracy of the proposed method in comparison with alternatives, we evaluate the numbers of true positives and false positives for main effects (M:TP and M:FP) and G-E interactions (I:TP and I:FP) respectively. In addition, we report the true positive counts when a total of 60 effects are identified as TP60. These measurements do not take environmental risk factors into account since they are not subject to selection.

Under each setting, we produce 200 datasets. The advantage of the proposed method is obvious in continuous genetic data settings. Summary results for gene expressions are presented in Table 3.1 and 3.2 that incorporates literature mining information J1. For example in Table 3.1 with correlation setting C1, the proposed method has $(18.5, 9.2, 36.2, 7.6)$, compared to $(14.9, 3.0, 19.0, 29.5)$ for HierMCP, $(5.2, 1.9, 18.3, 39.4)$ for MCP-LP, $(2.3, 0, 7.8, 9.8)$ for Lasso, and $(1.5, 0.1, 3.0, 4.6)$ for MA. Compared to HierMCP, the proposed method

yields superior results in identification, which provides strong and direct support to the estimation strategy that incorporates additional information using the Laplacian quadratic penalty. The proposed method also outperforms MCP-LP, which suggests the effectiveness of respecting the hierarchical structure of main effects and interactions by coefficient decomposition. Lasso and MA serve as benchmark analysis that both identify much fewer effects. This indicates that given high-dimensionality and low signal level, traditional approaches that do not address the"main effect, interaction" hierarchy structure nor additional information can lead to misleading discovery. In addition, we uses correlation as additional information and present summary results in Table B.1 and B.2 (Appendix). When zero and nonzero effects are correlated as C2, we also observe the satisfactory performance of the proposed method. Under correlation setting C1, it has TP60= 58.7 compared to 34.4 for HierMCP, 28.2 for MCP-LP, 8.8 for Lasso, and 4.1 for MA. The proposed method remains favorable for SNP data compared to alternative approaches. Summary results for SNP data are presented in Table 3.3 and 3.4. We observe that the proposed method has better or competitive performance in identification accuracy across different. For instance in Table 3.3 with MAF = 0.05 (M1) and AR(0.3), the proposed method has (M:TP, M:FP, I:TP, I:FP)= $(18.9, 9.4, 34.1, 40.5)$, compared to $(13.1, 48.5, 3.0, 7.9)$ for HierMCP, $(10.8, 0.2, 5.4, 30.6)$ for MCP-LP, $(0.1, 0, 3.5, 14.9)$ for Lasso, and $(0.1, 0.5, 1.5, 11.8)$ for MA. Across various settings, the superiority in identification performance of the proposed method demonstrates solid evidence that incorporating additional information using Laplacian quadratics improves accurate selection in the G-E interaction analysis.

3.4 Data analysis

We analyze data on cutaneous melanoma. Data are downloaded from TCGA Provisional using the R package *cgdsr*. The response of interest is the Breslow's depth, which measures the thickness of the tumor and has been extensively studied for the relationship with development and prognosis in melanoma patients (Dickson and Gershenwald, 2011). For environmental risk factors, we include age, sex, Clark level, and American Joint Committee on Cancer (AJCC) nodes pathologic stage (PN), all of which have been shown to be asso-

Table 3.1: Simulation results of S2 under correlation setting C1 and additional information J1. In each cell, mean(sd) based on 200 replicates.

	M:TP	M:FP	I:TP	I:FP	TP60
			AR(0.3)		
Proposed	18.5(1.3)	9.2(7.7)	36.2(2.7)	7.6(6.3)	52.7(4.2)
HierMCP	14.9(1.8)	3.0(2.6)	19.0(1.9)	29.5(2.6)	33.7(5.2)
MCP-LP	5.2(4.7)	1.8(1.5)	18.3(5.3)	39.4(9.5)	21.5(3.7)
Lasso	2.3(1.0)	0(0)	7.8(1.7)	9.8(3.9)	6.5(4.0)
MA	1.5(2.0)	0.1(0.3)	3.0(2.7)	4.6(8.3)	5.2(1.7)
			Band2		
Proposed	19.4(1.0)	2.3(2.6)	37.2(3.1)	5.3(6.3)	57.6(1.3)
HierMCP	15.3(1.6)	0(0)	20.7(2.7)	30.8(4.3)	34.5(1.8)
MCP-LP	5.2(3.9)	0.4(0.8)	22.3(4.3)	21.9(9.7)	38.4(3.9)
Lasso	2(1.0)	0(0)	11.8(2.9)	4.6(2.3)	8.8(4.5)
MA	4.4(3.7)	0.5(0.7)	9.4(4.6)	9.8(8.8)	7.4(1.3)
			LD(0.3)		
Proposed	19.3(1)	10.9(8.1)	37.8(2.6)	5.7(7.5)	52.3(4.7)
HierMCP	14.7(1.7)	4.6(2.7)	17.6(3.3)	30(2.7)	31.8(3.3)
MCP-LP	2.7(2.4)	1.8(2.2)	13.6(4.4)	49.8(25.2)	21.7(4.8)
Lasso	2.2(1.6)	0.4(0.5)	9.8(6.1)	17.6(13.7)	6.8(8.0)
MA	0.8(0.9)	0.1(0.3)	4.3(3.7)	16(25.8)	4.1(1.6)
			LD(0.5)		
Proposed	19.2(1.2)	4(6.0)	37.4(2.4)	5.7(6.8)	56.1(1.5)
HierMCP	15.7(1.4)	0.3(0.6)	20.9(2.0)	30.4(2.4)	35.4(4.3)
MCP-LP	4.7(4.2)	0.7(1.3)	19.1(4.5)	30.3(12.3)	30.8(4.1)
Lasso	3.8(1.5)	0.2(0.4)	11.3(3.1)	4.5(1.9)	8.2(4.8)
MA	4.2(3.6)	0.3(0.6)	7.1(3.5)	9.3(11.4)	7.1(2.1)

Table 3.2: Simulation results of S2 under correlation setting C2 and additional information J1. In each cell, mean(sd) based on 200 replicates.

	M:TP	M:FP	I:TP	I:FP	TP60
			AR(0.3)		
Proposed	19.6(1.0)	7.9(6.5)	37.3(3.1)	5.1(5.7)	53.8(4.0)
HierMCP	15.3(1.8)	1.6(1.8)	19.2(1.9)	30.3(2.3)	35.1(4.9)
MCP-LP	3.1(2.6)	1.4(1.2)	16.6(3.6)	36.7(10.2)	23.2(3.9)
Lasso	1.4(1.1)	0.6(0.5)	12.8(4.0)	19.2(4.1)	7.3(3.7)
MA	1.5(1.6)	0.5(0.8)	5.5(4.9)	14.1(4.2)	4.4(1.8)
			Band2		
Proposed	19.7(1.1)	2.4(3.2)	38(2.5)	4.2(4.7)	56.5(1.7)
HierMCP	15.3(1.2)	0.2(0.4)	19.9(1.6)	29.5(2.8)	35.5(4.6)
MCP-LP	6.4(4.6)	1.2(1.3)	23.7(3.2)	26.2(14.2)	39.4(7.2)
Lasso	3.4(2.3)	0(0)	11(2.9)	2.6(3.6)	9.3(4.6)
MA	4.3(4.1)	0.4(0.9)	9(3.5)	8.2(7.4)	7.4(1.9)
			LD(0.3)		
Proposed	19.1(0.9)	10.2(5.9)	36.8(2.8)	4.5(3.4)	52.9(4.0)
HierMCP	14.5(1.8)	7.5(4.6)	16.1(2.7)	29.3(3.6)	30(4.0)
MCP-LP	3.2(2.9)	2.2(2.1)	13.9(4.3)	49.2(27.2)	20.9(4.1)
Lasso	2(1.6)	1.6(1.5)	6.6(1.1)	23.8(15.3)	4.9(3.1)
MA	0.6(1.2)	0.4(0.7)	2.9(2.8)	10(18.4)	3.7(1.7)
			LD(0.5)		
Proposed	19.6(0.8)	5.9(8.6)	38.6(2.2)	2.6(3.3)	57.7(1.7)
HierMCP	15.6(1.5)	0.2(0.4)	20.7(1.9)	30.7(2.9)	36.3(4.4)
MCP-LP	4.2(3.8)	0.6(1)	20.1(3.2)	30.7(11.4)	31.9(3.2)
Lasso	4(2.9)	0(0)	11.8(3.1)	6.4(3.6)	9.3(4.1)
MA	2.7(3.3)	0.2(0.4)	6.2(5.3)	7.7(10.7)	6.5(2.3)

Table 3.3: Simulation results of S1 under MAF setting M1. In each cell, mean(sd) based on 200 replicates.

	M:TP	M:FP	I:TP	I:FP	TP60
			AR(0.3)		
Proposed	18.9(1.0)	9.4(9.8)	34.1(1.1)	34.7(11.8)	40.5(5.8)
HierMCP	13.1(2.1)	48.5(19.1)	3.0(2.9)	7.9(13.5)	15.2(2.2)
MCP-LP	10.8(4.4)	0.2(0.7)	5.4(4.0)	30.6(5.4)	21.3(4.5)
Lasso	0.1(0.2)	0(0)	3.5(2.6)	14.9(19.2)	3.5(2.6)
MA	0.1(0.2)	0.5(1.1)	1.5(1.7)	11.8(15.3)	5.8(3)
			AR(0.5)		
Proposed	19.0(1.2)	3.6(3.6)	34.6(1.5)	32.9(8.5)	41.4(4.4)
HierMCP	13.9(2.5)	34.6(22.7)	12.9(6.9)	46(32.9)	19.3(3.8)
MCP-LP	14.3(4.3)	0.5(0.5)	5.4(3.4)	27.4(3.9)	26.4(3.9)
Lasso	0.5(0.8)	0.1(0.2)	6.0(3.8)	15.4(13)	6.5(4)
MA	0.2(0.4)	1.7(2.9)	2.4(2.1)	18.7(16.5)	5.7(3.1)
			Band1		
Proposed	18.3(2.1)	6.4(4.9)	32.9(3.5)	30.5(8.6)	39.5(6.6)
HierMCP	12.3(3.3)	51.7(33.3)	8.1(4.1)	30.5(20.8)	17.0(2.8)
MCP-LP	11.7(6.2)	0.4(0.5)	4.5(4.1)	30.8(5.6)	21.9(4)
Lasso	0.1(0.3)	0(0)	2.4(2.3)	7.3(6.4)	2.5(2.3)
MA	0(0)	0.9(1.3)	1.2(1.5)	14.2(15.1)	5.6(2.8)
			Band2		
Proposed	18.9(1)	3.6(4.1)	35.3(1.7)	30.9(12.3)	39.1(5.3)
HierMCP	14.2(2.1)	39.9(24.7)	4.8(3.8)	12.0(15.7)	19.6(3)
MCP-LP	11.4(5.1)	0.5(0.8)	6.1(4.6)	28.7(5.5)	24.5(5.1)
Lasso	0.4(0.6)	0.1(0.2)	5.5(3.3)	20.9(23.7)	5.7(3.1)
MA	0(0)	0.8(1.5)	2.3(2.5)	18.4(22.1)	5.7(3)
			LD(0.3)		
Proposed	19.1(1.1)	2.7(5.8)	33.8(2.2)	30.2(12.3)	42.7(7.1)
HierMCP	14.3(2.1)	52.7(31.7)	2.2(2.7)	4.0(5.2)	19.9(2.5)
MCP-LP	12.5(4.6)	0.3(0.6)	9.0(4.4)	26.5(5.6)	27.1(5.2)
Lasso	0.3(0.6)	0(0)	4.9(3.3)	11.4(14.5)	5.1(3.3)
MA	0(0)	0.4(0.7)	2.7(2.6)	11.8(15.8)	5.6(2.8)
			LD(0.5)		
Proposed	19.2(1.0)	2.6(4.0)	34.0(1.4)	32.8(11.3)	41.2(7.3)
HierMCP	15.4(2.1)	68.8(27.8)	5.3(4.8)	15.9(22)	18.0(1.7)
MCP-LP	13.3(5.7)	0.6(1.1)	7.1(4.4)	25.3(5.7)	27.9(6.6)
Lasso	0.2(0.7)	0(0)	6.7(3.6)	22.3(28.4)	6.6(3.6)
MA	0.3(1.3)	2.4(7.2)	2.9(3)	17.6(22.5)	5.6(2.9)

ciated with melanoma. For genetic factors, we consider the mRNA gene expressions. The level 3 data in TCGA are collected using the IlluminaHiseq RNAseq V2 platform and have

Table 3.4: Simulation results of S1 under MAF setting M2. In each cell, mean(sd) based on 200 replicates.

	M:TP	M:FP	I:TP	I:FP	TP60
			AR(0.3)		
Proposed	18.1(2.3)	22.1(10.9)	32.4(4.4)	30.6(9.6)	34.5(6.4)
HierMCP	12.3(2.1)	19(9.3)	1.6(1.5)	2.2(2.2)	16.5(3.0)
MCP-LP	5.7(4.3)	1(1.4)	4.4(3.3)	39.5(8.7)	13.2(4.4)
Lasso	0(0)	0(0)	1.8(1.7)	6.8(5.4)	1.8(1.7)
MA	0.1(0.2)	0.4(0.7)	1.2(1.3)	11.3(17.9)	4.8(2.9)
			AR(0.5)		
Proposed	18.4(2.3)	13.2(12.9)	33.3(3.5)	31.9(14.1)	37.7(6.8)
HierMCP	13.9(3.0)	16.2(9.1)	4.0(3.0)	7.0(5.4)	20.5(3.3)
MCP-LP	8.7(4.7)	0.4(0.6)	3.3(4.1)	36.8(8)	18.3(6.6)
Lasso	0.4(0.8)	0(0)	5.4(3.2)	20.3(12.8)	5.8(3.2)
MA	0.3(0.7)	2.3(4.9)	1.3(1.7)	11.6(14.2)	4.9(3.1)
			Band1		
Proposed	17.4(2.3)	22(11.4)	30.8(4.2)	30.8(12.3)	35.8(6.2)
HierMCP	11.2(2.0)	19.4(11.7)	2.4(3.9)	5.2(9.9)	17.3(2.8)
MCP-LP	3.8(3.3)	1.9(2.8)	4.7(3.7)	51.2(25.2)	13.9(4.2)
Lasso	0(0)	0(0)	0.8(1.4)	2.8(4.7)	0.8(1.4)
MA	0(0)	0.2(0.4)	1.8(2.5)	13.8(18.6)	5.2(3.0)
			Band2		
Proposed	18.1(1.3)	10.3(7.8)	33.7(2.9)	29.7(14.1)	41.1(6.6)
HierMCP	13.9(1.9)	18.5(11.6)	5.4(4.5)	12.7(16.3)	21.6(4)
MCP-LP	8.5(5.6)	0.9(1.6)	2.3(2.6)	40.3(9.3)	18(4.2)
Lasso	0.1(0.2)	0(0)	2.7(3.0)	9.5(13.4)	2.7(3.1)
MA	0.1(0.2)	1.1(2.9)	2.7(3.2)	19.2(27.3)	4.9(3.0)
			LD(0.3)		
Proposed	18.2(1.2)	12.8(10.8)	33.2(2.2)	29.3(13.5)	42.3(7.4)
HierMCP	14(1.8)	19.1(10.8)	5.9(4.9)	12.9(15.4)	22.8(3.8)
MCP-LP	8.9(5.5)	1.2(2.5)	3.2(3.3)	38.7(9.9)	16.0(4.3)
Lasso	0.1(0.2)	0(0)	2.5(3.1)	9.2(14.2)	2.6(3.1)
MA	0.1(0.2)	0.5(0.9)	2.6(2.7)	17.8(24.1)	4.9(3.0)
			LD(0.5)		
Proposed	18.2(1.2)	12.8(10.8)	32.7(2.1)	29.8(13.4)	41.9(8.1)
HierMCP	13.8(1.7)	18.3(11.2)	6.2(5.0)	13.8(16.0)	23.1(3.8)
MCP-LP	8.9(5.5)	1.2(2.5)	3.2(3.3)	38.7(9.9)	16.1(4.3)
Lasso	0.1(0.2)	0(0)	2.7(3.1)	10.0(14.8)	2.5(3.1)
MA	0.1(0.2)	0.5(0.9)	2.8(2.7)	19.7(24.7)	4.7(3.0)

been lowest-normalized, log-transformed, and median centered. A total of 361 subjects are available with 18,355 measurements of gene expressions. We conduct a prescreening proce-

dure using marginal regression and select the top 1000 genes with the smallest p-values for downstream analysis.

The proposed method identifies 33 main genetic effects and 12 G-E interactions. Details are presented in Table 3.5. Published studies suggest potentially important implications of the findings. For instance, the proposed method identifies the interaction between the pathologic stage with gene TCTEX1D1, and it has been found as one of the differently methylated genes among metastatic melanoma patients. Gene MS4A14 has been found to be consistently altered in expression in cutaneous malignant melanoma patients with multiple in-transit metastases on the limbs. Expressions of LAMP2 cell-surface have been found in different human tumor cell lines. were correlated with better overall survival among gastric cancer patients, Gene MS4A14 has been showed positive expression in gastric cancer and correlated with better overall survival. Gene PLCB4 has been considered to be one of the plausible candidate driver genes of uveal melanoma and a tumor suppressor in cutaneous melanoma. A recurrent mutation in gene PLCB4 has been found to promote uveal melanoma tumorigenesis. We also confirm the identified genes are biologically sensible by enrichment analysis. The selected genes by main effects and interactions are used for enrichment analysis of pathways and diseases by DAVID version 6.8 (Sherman et al., 2009). We entered the identified genes from Table 3.5 into the web application `david.ncifcrf.gov`, with "OFFICIAL_GENE_SYMBOL" as gene identifier and Homo sapiens as species. The identified genes using the proposed approach are also significantly enriched into several GO terms. For example, seven identified genes (ISL1, SFI1, TAF9B, WWC2, CREG1, LPA, PON1) are enriched into negative regulation of cellular metabolic process (GO:0031324, p-value=0.0083). This has been selected to be one of the optimal features of the final characterization of skin cancer-related genes.

3.5 Discussion

In this Chapter, we have developed a new marginal G-E interaction analysis, which adopts a combination of penalization to respect the "main effects, interaction" hierarchical structure and to incorporate additional information. The advantage of the proposed method that

Table 3.5: Analysis of the SKCM data using the proposed method: identified G-E interactions.

		Age	Sex	Clark level	PN
ADK	0.0034				
AMER1	0.0019				
ARHGEF15	-0.0077				
CAMK1	-0.0120				
CDHR5	-0.0102				
COL6A4P2	-0.1560	0.2685	-0.3209	0.4746	-0.5368
CREG1	0.0043				
CYP51A1	-0.0090				
ENOSF1	-0.0102				
FAM107A	-0.0013				
FMC1	-0.0018				
FMO3	-0.0076				
HBS1L	0.0105				
HNRNPA0	0.0003				
HSP90AB2P	0.1462				
ISL1	0.0013				
ITGBL1	-0.0003				
LAMP2	0.0124				
LINC00908	-0.0070				
LPA	-0.0181				
MS4A14	-0.0033				
PLAC9	-0.0029				
PLCB4	-0.0042				
PON1	-0.0077				
REXO1L1P	-0.1557		-0.2845	-0.3137	0.3015
SFI1	-0.0061				
TAF9B	0.0014				
TCTEX1D1	-0.2992	0.1487			0.6563
USHBP1	-0.0163				
USP32P2	-0.1476	0.1439	-0.2570		0.2564
WWC2	0.0062				
YBX3P1	0.0074				
ZSWIM5	-0.0032				

utilizes the additional Laplacian quadratic penalty leads to more accurate identification and efficient computational algorithm. With the goal of improving the identification of G-E interaction, we have transformed the additional information by the adjacency matrix and then Laplacian quadratics. In main effect analysis, the sparse Laplacian shrinkage estimator has been comprehensively investigated and its theoretical properties including statistical inference are well established. Our proposed method adapts this penalization to conduct G-E interaction analysis. We have demonstrated considerable superiority can be achieved over multiple closely related alternatives using simulation under various settings.

In data analysis, biologically sensible findings are made.

Through penalization is more coherent under a joint modeling framework, it is still worth exploring and extending its applicability and performance to marginal models in the interaction analysis. We have witnessed great success in the joint analysis of G-E interaction. It can be of interest to migrate and to advance some of these analysis strategies to marginal models. It can also be of interest to extend our simulation to other types of responses, for example, survival outcomes. In the example of using texted-based mining data, we adopted PubMatrix and other software tools are also available including VxInsight, MedMiner, and others. Although bioinformatics and statistical evaluations have been conducted with the data analysis results, it is crucial to further validate the findings in functional studies of gene annotations.

Chapter 4

Integrating Multidimensional Molecular Data Into Interaction Analysis Using Sparse Biclustering and Lasso-Based Penalization

4.1 Introduction

For the outcomes and phenotypes of cancer, cardiovascular diseases, asthma, mental disorders, and other complex diseases, accumulating evidences have shown that multiple types of molecular changes, environmental risk factors, and their interactions play important roles. For example, the expression of gene IL9 is found to interact with environmental dust mite to increase severe asthma exacerbations in children (Sordillo et al., 2015). In the study of lung cancer genetics, it has been suggested that smoking can act through increasing the CNV (copy number variation) of gene IGF1 to induce its oncogenesis (Huang et al., 2011b). Epigenetic changes have also been investigated. For example, Teschendorff et al. (2015) finds that smoking-associated DNA methylation changes in buccal cells are associated with epithelial cancers. It is observed that in each of the aforementioned and other published studies, only the interactions between a single type of molecular changes and environmental

risk factors have been analyzed.

In recent biomedical studies, multidimensional profiling is becoming popular. In such studies, data on multiple types of molecular changes is collected on the same subjects. Such studies make it possible to not only more deeply understand disease biology but also construct more effective models for disease outcomes and phenotypes. A myriad of novel statistical methods has been developed. For example, Wang et al. (2012b) proposes an integrative Bayesian analysis to identify gene expression and methylation measurements that are associated with clinical outcomes such as survival. Gross and Tibshirani (2015) develops collaborative regression which applies penalization to explicitly accommodate the correlations (overlapping information) as well as independent information between gene expressions and CNVs for marker identification. Zhu et al. (2016a) develops a linear regulatory module-based method using the sparse SVD (singular value decomposition) and penalization techniques to integrate gene expressions and their regulators for cancer outcomes. We refer to Kristensen et al. (2014) and Wu et al. (2019a) for more discussions. The aforementioned and other published studies have convincingly shown that integrating multidimensional molecular data not only is biologically sensible but also improves estimation, marker identification, and prediction. It is noted that these studies have focused on the main effects of molecular changes.

Analyzing multidimensional molecular data as the main effects have provided rich and valuable information in cancer research. However, G-E interaction analysis of multidimensional molecular changes is lacking and relevant statistical methodologies are much underdeveloped (McAllister et al., 2017). In fact, incorporating distinct molecular levels of measurements to select important interactions is not trivial. Rather than immediately appending additional measurements to the existing methods for identifying G-E interactions, genomic regulations among different types of measurements need to be properly accommodated to the model of disease outcomes. Various frameworks have been proposed for analyzing genomic regulations, such as correlation analysis (Langfelder and Horvath, 2008) and network-based analysis (Breitling et al., 2004). Accumulative evidence suggests that it is limited to include only single type data such as gene expression levels as genetic factors in the interaction analysis. Those "one-dimensional" cancer-genomic studies may not be

comprehensive enough in exploiting interactions associated with cancer outcomes.

Motivated by the successes as well as limitations of the existing studies, here we conduct M-E interaction analysis, where M stands for multidimensional molecular changes and E stands for environmental risk factors. The objective is to collectively accommodate multiple types of high-dimensional molecular changes, environmental risk factors, and their interactions in modeling disease outcomes and phenotypes. This analysis is the natural next step of the integrated analysis of the main effects of multidimensional molecular data and studies that conduct the interaction analysis of a single type of molecular changes and environmental risk factors. Beyond the "ordinary" high dimensionality and noisy nature of molecular data, the analysis faces other challenges. Specifically, multiple types of molecular measurements are interconnected, which leads to overlapping information. For example, gene expression levels are regulated by genetic and epigenetic changes. On the other hand, they can also have independent information for disease outcomes (Risch and Plass, 2008). Several techniques, for example built on canonical correlation analysis (Meng et al., 2016) and matrix factorization (Zhang and Zhang, 2019), have been developed to accommodate such overlapping and independent information. In addition, interaction analysis demands respecting the unique "main effects, interactions" hierarchy (Bien et al., 2013; Wu et al., 2019b), for which multiple regularization techniques have been developed.

This study has the potential to significantly expand the gene-environment interaction analysis and multidimensional molecular data analysis paradigms. The proposed approach is designed tailored to the M-E analysis and will significantly advance from the aforementioned ones. With the growing popularity of multidimensional profiling, this study can open a new venue for modeling complex diseases.

4.2 Methods

The proposed approach can accommodate multiples types/combinations of molecular measurements. Without loss of generality and to avoid confusion with terminologies, we use gene expressions and their regulators (for example, genetic and epigenetic changes) as an example in description. Such a combination has been quite popular in published studies

(Wang et al., 2012b; Zhu et al., 2016a). Other combinations, for example proteins and gene expressions, can be analyzed in the same manner. Assume n iid subjects. Denote $\boldsymbol{G} = (\boldsymbol{G}_1, \cdots, \boldsymbol{G}_p)$ and $\boldsymbol{R} = (\boldsymbol{R}_1, \cdots, \boldsymbol{R}_q)$ as the $n \times p$ and $n \times q$ design matrices of p gene expression and q regulator measurements. Denote $\boldsymbol{E} = (\boldsymbol{E}_1, \cdots, \boldsymbol{E}_M)$ as the $n \times M$ design matrix of environmental risk factors, and $\boldsymbol{Y}$ as the length n vector of outcome. We first consider continuous outcomes and will discuss accommodating other types of outcomes later. Assume $\boldsymbol{Y}$ has been properly centered, and $\boldsymbol{E}, \boldsymbol{G}$, and $\boldsymbol{R}$ have been standardized.

4.2.1 M-E interaction analysis

Our goal is to identify important M-E interactions (as well as main effects) and construct a comprehensive outcome model. Overall, the proposed approach consists of the following main steps: (i) identification of the gene expression-regulator regulatory modules, which describe the regulation relationships (overlapping information), (ii) integration of multidimensional molecular measurements within the regulatory modules, and (iii) joint modeling and estimation that respect the "main effects, interactions" hierarchy. The analysis flowchart is provided in Figure 4.1.

Step I We employ a penalized regression to estimate the gene expression-regulator regulations and then sequentially conduct biclustering to identify the regulatory modules. Consider the model $\boldsymbol{G} = \boldsymbol{R}\boldsymbol{\Theta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is the $n \times p$ matrix of random errors and $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_p)$ is the $q \times p$ unknown coefficient matrix. For estimating $\boldsymbol{\Theta}$, consider

$$\hat{\boldsymbol{\Theta}} = \arg\min_{\boldsymbol{\Theta}} \ \frac{1}{2}||\boldsymbol{G} - \boldsymbol{R}\boldsymbol{\Theta}||_F^2 + \lambda \sum_{j=1}^{p} ||\boldsymbol{\theta}_j||_1, \tag{4.1}$$

where $||\cdot||_F$ and $||\cdot||_1$ denote the Frobenius norm of a matrix and L_1 norm of a vector, and $\lambda \geq 0$ is the tuning parameter.

To identify the regulatory modules, we propose conducting biclustering with $\hat{\boldsymbol{\Theta}}$. Here a regulatory module corresponds to a bicluster, which contains a small number of co-expressed gene expressions and their regulators. Specifically, for estimation, we adopt the sparse clustering technique developed in Helgeson et al. (2019), which first introduces weights for gene expressions and then maximizes the weighted between-cluster distance for regulators.

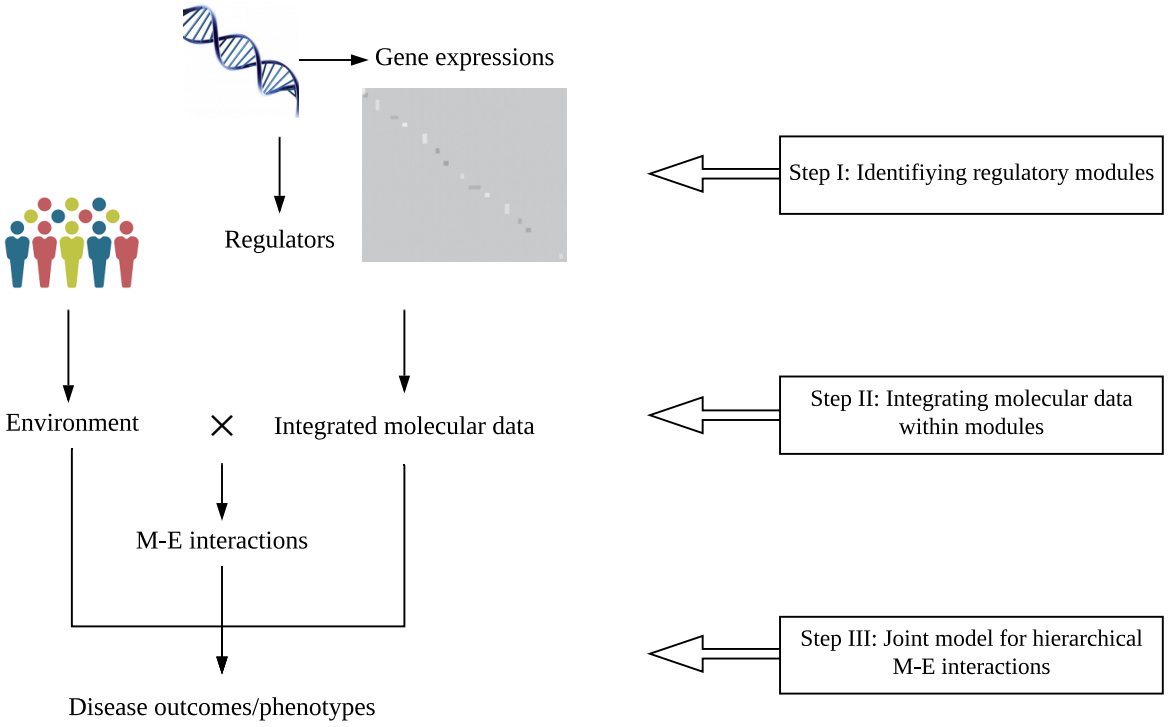


Figure 4.1: Flowchart of the proposed M-E interaction analysis.

The objective function is

$$\max_{\mathcal{C},\bar{\mathcal{C}},\boldsymbol{w}} \sum_{j=1}^{p} w_j \left(\frac{1}{q} \sum_{l=1}^{q} \sum_{l'=1}^{q} d_{l,l',j} - \frac{1}{q_1} \sum_{l,l' \in \mathcal{C}} d_{l,l',j} - \frac{1}{q_2} \sum_{l,l' \in \bar{\mathcal{C}}} d_{l,l',j} \right), \tag{4.2}$$

$$\text{subject to } ||\boldsymbol{w}||^2 \leq 1, ||\boldsymbol{w}||_1 \leq \sqrt{p}, \text{ and } w_j \geq 0 \text{ for } j = 1, \cdots, p,$$

where $\hat{\theta}_{lj}$ is the (l,j)th component of $\hat{\boldsymbol{\Theta}}$, $d_{l,l',j} = \left(\hat{\theta}_{lj} - \hat{\theta}_{l'j}\right)^2$ measures the distance between the lth and l'th regulators, $\mathcal{C}$ and $\bar{\mathcal{C}}$ are the disjoint index sets of regulator clusters, $q_1 = |\mathcal{C}|$ and $q_2 = |\bar{\mathcal{C}}|$ are the cardinalities of $\mathcal{C}$ and $\bar{\mathcal{C}}$ with $q_1 < q_2$ and $q_1 + q_2 = q$, and $\boldsymbol{w} = (w_1, \cdots, w_p)'$ is the weight vector for gene expressions, with a larger weight indicating higher importance for clustering. With the constraints for $\boldsymbol{w}$, each w_j has a nonzero value between 0 and 1. With the estimated weight $\hat{\boldsymbol{w}}$, a two-sample permutation-based Kolmogorov-Smirnov test is conducted to test the significance of the difference between two clusters and select the gene expression set $\mathcal{D}$ with significantly large weights. This process leads to

one regulatory module $\{\mathcal{C}, \mathcal{D}\}$ with regulators in $\mathcal{C}$ and gene expressions in $\mathcal{D}$. To obtain subsequent modules, we update $\hat{\mathbf{\Theta}}$ by subtracting the module just identified and repeat the above procedure. This process is iterated until the Kolmogorov-Smirnov test fails to reject the null hypothesis of no clusters. With the sparsity of $\hat{\mathbf{\Theta}}$, it is expected that only a subset of gene expressions and regulators can form modules. Suppose that there are S identified modules $\{\mathcal{C}_1, \mathcal{D}_1\}, \cdots, \{\mathcal{C}_S, \mathcal{D}_S\}$.

Rationale Linear regression is used to describe the regulations between two types of molecular measurements. Multiple published studies (Shi et al., 2015; Zhu et al., 2016a) have shown that it is a sensible choice, especially considering the high dimensionality. One gene expression is regulated by only a few regulators, and one regulator affects the expressions of only a few genes. As such, $\mathbf{\Theta}$ is assumed to be sparse, and the Lasso penalization is applied for estimation and identification of important regulations.

The concept of regulatory module has been developed in Zhu et al. (2016a) and other studies. A regulatory module consists of a small number of gene expressions and regulators that behave in a coordinated manner. The construction in Zhu et al. (2016a), which is based on sparse SVD, limits each regulatory module to have rank one. Here we lift this inconvenient constraint via biclustering. By construction, each bicluster (regulatory module) consists of gene expressions and regulators sharing similar patterns in $\mathbf{\Theta}$. We adopt the sparse biclustering method developed in Helgeson et al. (2019) because of its favorable numerical performance. Note that here we cluster regulators into two disjoint groups with weighted gene expressions. It is also possible to reverse the roles of gene expressions and regulators, and this leads to similar clustering results in our numerical investigations. With the sequential cluster construction strategy, different regulatory modules may have overlaps. This is desirable as one gene/regulator can participate in multiple biological processes.

Step II We integrates information within each regulatory module $\{\mathcal{C}_s, \mathcal{D}_s\}, s = 1, \cdots, S$, using the PCA (principal component analysis) technique. Given a matrix $\boldsymbol{A}$ and index set $\mathcal{I}$, denote $\boldsymbol{A}_{\mathcal{I}}$ as the columns of $\boldsymbol{A}$ indexed by $\mathcal{I}$. For the sth module, we apply PCA to the stacked matrix $(\boldsymbol{G}_{\mathcal{D}_s}, \boldsymbol{R}_{\mathcal{C}_s})$ and select the top PCs with the cumulative variance contribution rate $\geq 80\%$. Denote the resulted matrix composed of the p_s PCs as $\boldsymbol{X}_s = (\boldsymbol{X}_{s,1}, \cdots, \boldsymbol{X}_{s,p_s})$. In addition, for gene expressions and regulators not involved in any identified modules,

we collect and combine them as $\boldsymbol{Z} = (\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_{p_z}) = (\boldsymbol{G}_{\mathcal{D}^c}, \boldsymbol{R}_{\mathcal{C}^c})$, where $\mathcal{D}^c = \{j \in \{1, \cdots, p\} : j \notin \mathcal{D}_s, s = 1, \cdots, S\}$ and $\mathcal{C}^c = \{j \in \{1, \cdots, q\} : j \notin \mathcal{C}_s, s = 1, \cdots, S\}$. $\boldsymbol{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_S)$ and $\boldsymbol{Z}$ form input for downstream analysis.

Rationale The previous step of analysis does not directly limit the sizes of the modules. Thus, it is possible some modules have moderate to large sizes. In addition, with regulations, measurements within the same modules often times have strong correlations. To reduce dimensionality, remove collinearity, and simplify computation, we apply PCA, which can be replaced by other dimension reduction techniques. Overall, the input for the next step consists of the PCs (representing overlapping information) and the gene expressions and regulators that do not form patterns (representing independent information).

Step III Here we conduct interaction analysis, that respects the "main effects, interactions" hierarchy (Bien et al., 2013). For the continuous outcome, consider the regression model

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{E}\boldsymbol{\alpha} + \sum_{s=1}^{S} \boldsymbol{X}_s \boldsymbol{\beta}_s + \boldsymbol{Z}\boldsymbol{\gamma} + \sum_{m=1}^{M}\sum_{s=1}^{S} (\boldsymbol{E}'_m \odot \boldsymbol{X}'_s)'(\boldsymbol{\beta}_s * \boldsymbol{\eta}_{sm}) + \sum_{m=1}^{M} (\boldsymbol{E}'_m \odot \boldsymbol{Z}')(\boldsymbol{\gamma} * \boldsymbol{\tau}_m) + \boldsymbol{\xi}, \\ &= g(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{E}) + \boldsymbol{\xi}. \end{aligned} \tag{4.3}$$

Here $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_M)'$, $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \cdots, \boldsymbol{\beta}'_S)$, and $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_{p_z})'$ correspond to the main effects of the environmental factors, regulatory modules, and individual molecular measurements (that do note belong to any module), respectively. For the mth environmental factor, $\boldsymbol{\beta}_s * \boldsymbol{\eta}_{sm}$ and $\boldsymbol{\gamma} * \boldsymbol{\tau}_m$ correspond to the interactions with the sth regulatory module and all individual molecular measurements, respectively, with $*$ being the component-wise product. $\odot$ is the "matching column-wise" Khatri-Rao product. $\boldsymbol{\xi}$ is the random error vector. Here, following the literature (Choi et al., 2010), we use the products $E_{im}X_{ij}$ and $E_{im}Z_{ij}$ to describe the interactions for the ith subject. To accommodate the hierarchical structure of interaction analysis, the interaction effects $\beta_{sj}\eta_{smj}$ and $\gamma_j\tau_{mj}$ are decomposed into two components, the first for the corresponding main effects (β_{sj} and γ_j) and the other for the interaction-specific effects (η_{smj} and τ_{mj}).

For the estimation and identification of important interactions (and main effects), we

propose the penalized objective function

$$Q(\boldsymbol{\Phi}) = \frac{1}{2}||\boldsymbol{Y} - g(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{E})||_2^2 + \lambda_1 \sum_{s=1}^{S} \sqrt{p_s} \left(||\boldsymbol{\beta}_s||_2 + \sum_{m=1}^{M} ||\boldsymbol{\eta}_{sm}||_2 \right) + \lambda_2 \left(||\boldsymbol{\gamma}||_1 + \sum_{m=1}^{M} ||\boldsymbol{\tau}_m||_1 \right), \tag{4.4}$$

where $\boldsymbol{\Phi} = (\boldsymbol{\alpha}', \boldsymbol{\beta}_1', \cdots, \boldsymbol{\beta}_S', \boldsymbol{\gamma}', \boldsymbol{\eta}_{11}', \cdots, \boldsymbol{\eta}_{MS}', \boldsymbol{\tau}_1', \cdots, \boldsymbol{\tau}_M')'$, $||\cdot||_2$ is the L_2 norm of a vector, and $\lambda_1, \lambda_2 \geq 0$ are tuning parameters. Gene expressions and regulators that are involved in modules with nonzero estimated $\boldsymbol{\beta}_s$ and $\boldsymbol{\beta}_s * \boldsymbol{\eta}_{sm}$ are identified as having important main effects and M-E interactions, respectively. In addition, for individual molecular measurements, the nonzero components of $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma} * \boldsymbol{\tau}_m$ correspond to important main effects and interactions, respectively.

Rationale A joint model is developed to accommodate all molecular and environmental effects and their interactions. As to be described below, the linear regression model can be replaced by other models. For estimation and selection, we adopt penalization, which has been the choice of quite a few recent interaction studies (Bien et al., 2013; Wu et al., 2019b). For many datasets including those analyzed in this article, the environmental factors are pre-selected based on existing knowledge and usually considered as important, so that their coefficients are not subject to penalized selection. As such, the "main effects, interactions" hierarchy postulates that an identified interaction corresponds to an identified main molecular effect. To achieve this, we decompose the interaction effects into two components and have that $\beta_{sj}\eta_{smj} \neq 0$ only if $\beta_{sj} \neq 0$ and $\gamma_j\tau_{mj} \neq 0$ only if $\gamma_j \neq 0$ (Choi et al., 2010). In (4.4), we employ group Lasso for regulatory modules (where PCs corresponding to the same module form a group) and Lasso for individual molecular measurements to identify M-E interactions and main effects. Here, all PCs corresponding to the same module are in or out simultaneously, which is motivated by the coordinated nature of the molecular measurements in the same module.

Accommodating other types of outcomes With a different type of outcome variable, the lack-of-fit in (4.4) can be replaced by the negative log-likelihood function or an estimating equation-based measure. As an example, consider survival data, which is analyzed be-

low. Denote $\boldsymbol{T}$ as the length n vector of survival times. Consider the AFT (accelerate failure time) model $\log(\boldsymbol{T}) = g(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{E}) + \boldsymbol{\xi}$, where notations have similar implications as above. Denote $\boldsymbol{C}$ as the length n vector of censoring times, then we observe $\boldsymbol{Y} = \log(\min(\boldsymbol{T}, \boldsymbol{C}))$ and $\boldsymbol{\delta} = I(\boldsymbol{T} \leq \boldsymbol{C})$ with $I(\cdot)$ being the indicator function. Assume that data has been sorted according to the observed times from the smallest to the largest. Compute the Kaplan-Meier weights: $\rho_1 = \frac{\delta_1}{n}$, $\rho_i = \frac{\delta_i}{n-i+1} \prod_{i'=1}^{i-1} (\frac{n-i'}{n-i'+1})^{\delta_{i'}}, i = 2, \ldots, n$. Then, we have the weighted penalized objective function

$$\frac{1}{2}||\sqrt{\boldsymbol{\rho}} * (\boldsymbol{Y} - g(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{E}))||_2^2 + \lambda_1 \sum_{s=1}^{S} \sqrt{p_s} \left(||\boldsymbol{\beta}_s||_2 + \sum_{m=1}^{M} ||\boldsymbol{\eta}_{sm}||_2 \right) + \lambda_2 \left(||\boldsymbol{\gamma}||_1 + \sum_{m=1}^{M} ||\boldsymbol{\tau}_m||_1 \right).$$

4.2.2 Computation

The detailed computational algorithms for Steps I and III are provided in Algorithms 1 and 2 (Appendix), respectively. Step II can be realized using existing algorithms and R function `prcomp`. In computation, effort has been made to take advantage of the existing algorithms and software. When not possible, optimization has been based on the CD (coordinate descent) techniques. In the literature, convergence properties of the CD and other techniques used in computation have been well established. Convergence is observed in all of our numerical studies. The two tuning parameters in (4.4) are selected using the extended Bayesian information criterion (Chen and Chen, 2008). The proposed algorithm is computationally feasible. For example, under a standard laptop configuration, it takes less than five minutes for a simulated dataset with 250 subjects, 500 gene expression measurements, and 500 regulator measurements. We have developed R code implementing the proposed approach and made it publicly available at `https://github.com/shuanggema/omics_interaction`.

4.2.3 Heuristic theoretical justifications

Consider the scenario where the number of molecular factors (gene expressions and their regulators) increases and the number of environmental factors is finite as the sample size increases. There are several key estimation procedures and conditions. First, in the step of identifying regulatory modules, the consistency of Lasso estimator $\hat{\boldsymbol{\Theta}}$ is needed. For

each gene expression, with probability at least $1 - \frac{2}{\sqrt{\pi}} q u_n^{-1} e^{-u_n^2/2}$, $\boldsymbol{\theta_j}$ can satisfy the weak oracle property, under mild regularity conditions on the design matrix $\boldsymbol{R}$, signal strengths, Gaussian random error, and $q = o(u_n e^{u_n^2/2})$. Here, the order of u_n can be $o(n^a)$ with $a \in (0, \frac{1}{2}]$, leading to $\log(q) = o(n^{2a})$. Thus, with a total of p gene expressions, to ensure the overall consistency of $\hat{\boldsymbol{\Theta}}$, it is required that $1 - \frac{2}{\sqrt{\pi}} qp u_n^{-1} e^{-u_n^2/2} \to 1$ with the Bonferroni approach. Assume that p and q are of the same order, then we have $\log(q) = \log(p) = o(n^a)$. Second, the adopted biclustering strategy is an "upgrade" of the sparse K-means clustering with an L_2/L_1 penalty. For the sparse K-means with an L_∞/L_0 penalty, it has been shown in Chang et al. (2018) that under certain regularity conditions, the estimated weight $\boldsymbol{w}$ has feature selection consistency. Consistency under an L_2/L_1 penalty is expected to hold with revised norm assumptions, which will lead to consistency of the estimated gene expression clusters. Consistency of the estimated cluster centers of K-means has been well established in Pollard (1981), which can support the consistency of the estimated regulator clusters. Combining such results is expected to lead to the consistency of biclustering. Third, for each regulatory module, PCA is conducted to extract integrated information. With the ratio $n/(|\mathcal{C}_s| + |\mathcal{D}_s|) \to 0$, Jung et al. (2009) shows that if the first few eigenvalues are large enough compared to the others, then the corresponding estimated PC directions are consistent or converge to the appropriate subspace (subspace consistency). Finally, for estimators in interaction analysis with hierarchy, consistency has been established in Choi et al. (2010) and Wu et al. (2019b). As shown in Wu et al. (2019b), under mild regularity conditions on the design matrix, smallest signal, and tuning parameters, the estimator has consistency properties, where the dimensionality $p + q$ is allowed to grow up exponentially faster than the sample size.

4.3 Simulation

We set $p = q = 500$, $M = 5$, and $n = 250$, and generate environmental factors from independent standard normal distributions. In addition, (a) we consider two settings for $\boldsymbol{\Theta}$ to represent different regulation patterns. The first ($\boldsymbol{\Theta}_1$) contains 15 regulatory modules with one overlapping. The corresponding elements are independently generated from normal

distributions with mean ranging from -0.7 to 1.5 and standard deviation 0.1, covering different levels and directions of regulations on average. Each regulatory module contains 12.3 gene expressions and 16.6 regulators. The rest elements of $\boldsymbol{\Theta}_1$ are zero. The second ($\boldsymbol{\Theta}_2$) contains 20 nonzero regulatory modules with one overlapping, and the nonzero values are generated similarly as $\boldsymbol{\Theta}_1$. Those modules consist of 6.0 gene expressions and 8.1 regulators on average. Compared to $\boldsymbol{\Theta}_1$, $\boldsymbol{\Theta}_2$ contains more modules with smaller sizes, representing a different type of regulations. (b) The values of regulators $\boldsymbol{R}$ involved in each regulatory module are generated from a multivariate normal distribution with marginal means 0 and variances 1. We consider three correlation structures. The first (R1) is an auto-regressive structure where the correlation between the jth and lth variables is $(-0.5)^{|j-l|}$. The second (R2) is a banded structure where the correlation between the jth and lth variables is -0.5 if $|j-l|=1$ and 0 otherwise. The third (R3) has a structure where the correlation between the jth and lth variables is $(-1)^{|j-l|}/(|\mathcal{C}_s|+|\mathcal{D}_s|)$. Among them, R1 and R2 are "diagonally dominant", while R3 has all correlations at the same level. The individual regulators that are not involved in any regulatory modules are independently generated from the standard normal distribution. As such, regulators in different modules are independent of each other and also independent of the individual regulators. (c) Gene expression measurements are generated by $\boldsymbol{G}=\boldsymbol{R}\boldsymbol{\Theta}+\boldsymbol{\epsilon}$, where the elements of $\boldsymbol{\epsilon}$ follow independent standard normal distributions. (d) Given $\boldsymbol{G}$, $\boldsymbol{R}$, and $\boldsymbol{\Theta}$, generate the integrated information $\boldsymbol{X}_s$ for each module using the top PCs and $\boldsymbol{Z}$ for the individual molecular units. (e) With $\boldsymbol{X}_s, s=1,\cdots,S$ and $\boldsymbol{Z}$, consider the continuous response under model (5.2). Two types of nonzero coefficient settings are considered, leading to a total of 100 (P1) and 70 (P2) important main molecular effects and M-E interactions, respectively. These nonzero coefficients are generated uniformly from $(0.5, 0.8)$ (B1) or $(0.8, 1.2)$ (B2), representing two signal levels, with the "main effects, interactions" hierarchical structure satisfied. The molecular factors with important effects include gene expressions and regulators involved in the regulatory modules as well as individual molecular measurements. Additional information is provided in the Appendix. Random errors $\boldsymbol{\xi}$ are generated from independent standard normal distributions.

To better appreciate operating characteristics of the proposed module detection procedure, we simulate one dataset under setting $\boldsymbol{\Theta}_1$ and correlation structure R1. We present

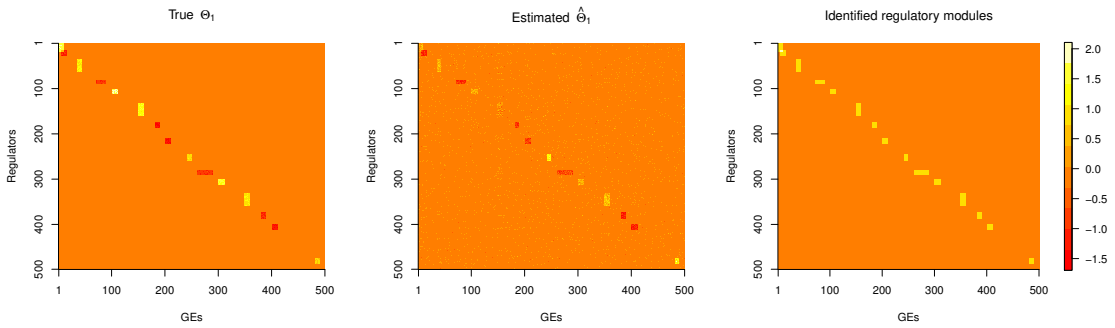


Figure 4.2: Simulation. Left: true values of regulation under setting $\mathbf{\Theta}_1$ and R1; Middle: estimated values; Right: identified regulatory modules.

the true regulation relationships between gene expressions and regulators in Figure 4.2, together with their estimated values and identified regulatory modules. We observe that with moderate associations between small sets of molecular measurements, the estimated $\hat{\mathbf{\Theta}}$ based on Lasso closely reflects the true regulation relationships. Furthermore, biclustering is able to properly identify the regulatory modules based on the estimated regulations.

To be more informative, besides the proposed approach, we also consider the following alternatives which have closely related frameworks. Comparing with these alternatives can directly establish the necessity of the considerations on gene expression-regulator regulations, correlations within regulatory modules, and hierarchical interactions. Specifically, Alt.1 excludes Step II of integration and builds the hierarchical interaction model using gene expressions and regulators directly combined as groups based on the identified regulatory modules. Alt.2 excludes the decomposition of interaction coefficients in Step III, and so the "main effects, interactions" hierarchical structure may be violated. Alt.3 builds the hierarchical joint model directly using the original stacked gene expression and regulator measurements without accounting for the regulations. Alt.4 incorporates the original stacked gene expression and regulator measurements directly in the interaction model. It ignores the regulation relationships and interaction hierarchy. For evaluation, we consider the numbers of true positives (TP) and false positives (FP) for main effects and interactions together.

For each scenario, 200 replicates are simulated. Summary results under settings P1 and P2 are presented in Tables 4.1 and 4.2 respectively. We observe that the proposed ap-

proach achieves better or comparable performance in identification accuracy. For example in Table 4.1 with weak effects (B1), regulation pattern $\mathbf{\Theta}_1$, and correlation structure R1, the proposed approach selects on average 95.94 true positives, compared to 71.90 (Alt.1), 65.20 (Alt.2), 23.15 (Alt.3), and 16.80 (Alt.4). When there are more correlated molecular measurements, the proposed approach remains superior in identification. For instance in Table 4.1 with weak effects (B1), regulation pattern $\mathbf{\Theta}_1$, and correlation structure R3, the proposed approach selects on average 99.70 true positives with 8.50 false positives. In comparison, Alt.1, Alt.2, Alt.3, and Alt.4 select fewer true positives and more false positives with (TP,FP)=(83.68,12.26), (95.90,54.75), (27.30,14.70), and (20.75,136.65), respectively. With a higher signal level under setting B2, all approaches behave better, while with more regulation modules under setting $\mathbf{\Theta}_2$, performance of all approaches decays. Under both settings, the proposed approach still has advantage. It is observed that Alt.1 generally achieves the second best identification performance, and under some scenarios it is competitive in true positive identification compared to the proposed approach, at the cost of larger numbers of false positives. This is because the integration procedure of the proposed approach that uses PCs for the joint interaction model can effectively remove collinearity and reduce false discovery. The proposed approach performs better than Alt.2, suggesting that the hierarchical interaction modeling can lead to more accurate identification. The superior performance of the proposed approach over Alt.3 and Alt.4 provides a direct support to the integrated analysis strategy that accommodating the regulations among multidimensional molecular data in interaction analysis substantially improves identification performance.

4.4 Data analysis

TCGA is one of the largest data resources with multidimensional profiling. TCGA data have been analyzed in interaction analysis with one type of molecular measurements as well as integrated modeling with the main effects of multiple types of molecular measurements. This study is the first to conduct the integrated M-E interaction analysis. We analyze data on lung adenocarcinoma (LUAD) and cutaneous melanoma (SKCM). Data are downloaded from TCGA Provisional using the R package *cgdsr*.

Table 4.1: Summary results for simulation under setting P1 with a total of 100 true positives: mean (sd) from 200 replicates.

			Θ_1		Θ_2	
		Approach	TP	FP	TP	FP
B1	R1	Proposed	95.94(4.63)	11.39(13.83)	80.06(5.37)	6.31(6.02)
		Alt.1	71.90(35.35)	20.45(24.20)	27.06(19.83)	1.94(1.12)
		Alt.2	65.20(28.39)	28.75(14.03)	31.69(15.05)	7.94(18.32)
		Alt.3	23.15(3.47)	7.45(3.90)	20.35(9.10)	19.85(8.43)
		Alt.4	16.80(2.09)	122.20(38.35)	29.85(5.73)	127.40(40.31)
	R2	Proposed	97.30(1.75)	5.70(10.99)	80.72(4.64)	20.89(25.90)
		Alt.1	86.60(30.13)	13.00(16.06)	47.00(16.82)	6.11(4.92)
		Alt.2	85.15(16.11)	39.95(6.87)	33.61(11.44)	19.06(21.78)
		Alt.3	23.35(4.18)	7.65(2.89)	21.05(5.77)	25.79(6.27)
		Alt.4	16.95(2.86)	126.05(46.47)	16.00(9.56)	71.15(65.52)
	R3	Proposed	99.70(0.57)	8.50(14.60)	79.40(3.22)	36.13(38.10)
		Alt.1	83.68(27.69)	12.26(18.29)	51.14(25.72)	8.71(11.69)
		Alt.2	95.90(8.09)	54.75(12.48)	30.50(14.39)	4.50(7.60)
		Alt.3	27.30(1.63)	14.70(17.41)	20.21(7.79)	22.11(8.46)
		Alt.4	20.75(2.65)	136.65(37.06)	20.00(8.55)	103.05(63.77)
B2	R1	Proposed	99.80(0.41)	14.25(18.95)	83.90(4.43)	12.60(10.56)
		Alt.1	99.80(0.41)	57.80(22.75)	32.00(23.43)	14.35(13.92)
		Alt.2	85.80(14.06)	55.55(27.20)	34.75(13.98)	18.25(9.48)
		Alt.3	27.17(2.46)	5.28(1.02)	27.15(8.67)	35.15(11.45)
		Alt.4	21.45(2.98)	142.05(31.31)	30.30(10.98)	110.10(66.72)
	R2	Proposed	99.82(0.39)	4.12(14.69)	77.88(3.67)	20.81(12.93)
		Alt.1	90.80 (27.98)	38.65(21.69)	42.69(22.46)	12.06(8.73)
		Alt.2	77.85(19.63)	47.05(16.62)	21.75(19.49)	19.05(22.55)
		Alt.3	27.37(2.29)	7.79(3.31)	17.45(5.84)	18.95(11.87)
		Alt.4	19.60(2.19)	135.35(36.02)	14.95(9.74)	52.40(47.98)
	R3	Proposed	99.35(0.67)	12.05(17.72)	77.77(2.95)	9.85(6.67)
		Alt.1	96.45(13.77)	50.45(17.38)	35.65(19.45)	21.95(13.06)
		Alt.2	86.45(12.17)	46.45(11.91)	16.50(16.99)	10.00(13.13)
		Alt.3	28.88(3.14)	7.71(2.52)	14.35(4.89)	16.35(7.19)
		Alt.4	21.25(2.71)	140.65(30.84)	14.95(10.79)	65.15(62.45)

4.4.1 Analysis of LUAD data

The response of interest is the reference value for the pre-bronchodilator forced expiratory volume in one second in percent (FEV1). It is an important biomarker for lung capacity, with a lower value suggesting the potentially functional disorder of the lung, and has been shown to be a powerful marker for future morbidity and mortality (Young et al., 2007). It is continuously distributed and ranges from 1.95 to 156 with mean 80.58 and standard deviation 23.55. We focus on the primary tumor samples of the Whites. For environmental

Table 4.2: Summary results for simulation under setting P2 with a total of 70 true positives: mean (sd) from 200 replicates.

			Θ_1		Θ_2	
		Approach	TP	FP	TP	FP
B1	R1	Proposed	68.85(0.88)	0.40(0.50)	67.30(3.26)	2.30(5.65)
		Alt.1	63.30(15.69)	34.80(23.26)	33.95(20.68)	6.68(10.37)
		Alt.2	65.25(4.27)	31.85(8.55)	52.15(10.98)	38.30(36.79)
		Alt.3	22.25(4.46)	7.85(4.49)	20.95(4.32)	23.85(9.28)
		Alt.4	15.30(2.75)	129.45(42.29)	28.10(4.10)	127.65(43.63)
	R3	Proposed	57.15(18.43)	10.50(6.36)	53.90(12.49)	2.65(2.21)
		Alt.1	42.55(28.65)	18.00(25.46)	38.25(15.21)	4.40(8.18)
		Alt.2	42.50(21.19)	24.05(14.57)	28.00(17.26)	3.70(6.14)
		Alt.3	24.50(2.50)	10.90(8.42)	18.00(11.31)	23.00(24.04)
		Alt.4	14.30(1.95)	107.05(30.44)	16.95(5.31)	83.85(51.99)
	R3	Proposed	67.15(6.71)	1.40(3.98)	51.90(13.63)	2.55(2.74)
		Alt.1	61.00(19.74)	27.80(17.56)	34.25(6.54)	8.10(14.49)
		Alt.2	65.35(5.05)	36.35(11.94)	44.45(15.43)	47.00(43.04)
		Alt.3	22.75(2.90)	9.95(7.49)	15.75(4.88)	18.85(12.33)
		Alt.4	15.40(2.26)	117.65(30.91)	19.20(5.69)	105.85(61.27)
B2	R2	Proposed	69.75(0.55)	1.70(6.67)	67.05(5.88)	10.50(11.00)
		Alt.1	69.75(0.44)	32.30(11.68)	65.40(7.38)	28.65(30.47)
		Alt.2	66.40(5.23)	38.15(30.67)	46.00(12.02)	2.30(5.25)
		Alt.3	25.79(5.54)	11.05(17.48)	20.95(4.32)	23.85(9.28)
		Alt.4	16.65(2.21)	136.00(36.41)	36.25(4.27)	150.50(45.17)
	R2	Proposed	67.15(11.35)	1.15(3.77)	57.10(11.11)	4.55(4.08)
		Alt.1	69.80(0.52)	33.70(12.69)	41.85(15.79)	16.35(9.42)
		Alt.2	58.15(9.91)	36.55(11.76)	43.55(15.74)	11.50(13.61)
		Alt.3	26.85(2.89)	6.85(2.13)	18.00(11.31)	23.00(24.04)
		Alt.4	17.50(2.65)	148.40(45.18)	18.90(6.21)	91.40(53.36)
	R3	Proposed	69.85(0.37)	2.80(7.25)	56.77(8.12)	3.69(6.32)
		Alt.1	68.90(4.46)	32.95(9.29)	39.35(12.57)	19.35(18.79)
		Alt.2	66.35(6.67)	35.45(7.99)	50.05(11.98)	7.15(14.18)
		Alt.3	26.70(3.37)	7.40(3.42)	15.75(4.88)	18.85(12.33)
		Alt.4	16.75(1.97)	133.65(32.08)	16.20(6.05)	63.40(43.57)

risk factors, we consider age, American Joint Committee on Cancer (AJCC) tumor pathologic stage (Stage), tobacco smoking history indicator (Smoking), and gender, which have been extensively investigated in the literature. We analyze mRNA gene expression measurements which were collected using the Illumina HiSeq 2000 RNA Sequencing Version 2 analysis platform. For regulators, we include CNV measurements that were collected using the Genome-Wide Human SNP Array 6.0 platform and DNA methylation measurements that were collected using the Illumina Infinium HumanMethylation450 platform. A total of

18,345 gene expression, 23,321 CNV, and 15,288 methylation measurements are available. In principle, the proposed approach can be directly applied. However, considering that only a small number of molecular measurements are potentially associated with the outcome and the analysis may be unstable with the high dimensionality and small sample size, we conduct a prescreening. Specifically, we select the top 1,000 molecular measurements with the smallest p-values using marginal regression. This leads to 164 subjects with 467 gene expression and 533 regulator (316 CNV and 217 methylation) measurements for downstream analysis.

The proposed analysis identifies 20 regulatory modules in Step I, and each module on average contains 11.70 gene expression and 7.35 regulator measurements. The graphical presentation of the modules is provided in Figure C.1 (Appendix), where some overlappings between modules are observed. In interaction analysis, the proposed approach identifies 62 main molecular effects and 29 M-E interactions, among which 50 main effects and 27 interactions belong to six regulatory modules. The identified main effects consist of 41 gene expression, 9 CNV, and 12 methylation measurements, and the identified interactions consist of 20 with gene expressions and 9 with methylations. Detailed estimation results are presented in Table 4.3, where a "group" corresponds to a module or an individual measurement. Literature search suggests that the findings are biologically sensible. For example, Stage and Smoking are shown to be negatively associated with FEV1, which has also been suggested in previous studies. Gene AFF3 is identified along with its interactions with Smoking and gender. A decreased methylation of gene AFF3 in non-small cell lung tumors has been found as one of the key epigenetic changes associated with lung cancer development. Gene PWRN1 has been reported to be involved in the process of spermatogenesis, and its expression level has been shown to be related to tumor size in lung cancer patients. Gene CACNG3 has been identified as an oncogene from a pan-cancer study with somatic mutation data, suggesting its potentially important role for lung adenocarcinoma. Gene PRH1 has been identified as one of the candidate exosomal protein biomarkers for the detection of lung cancer using human saliva and serum. In addition, published studies have shown that gene CACNG6 is significantly upregulated in lung squamous cell carcinoma compared to normal lung tissues. Gene PABPC5 has been found to be hypermethylated

among early-stage non-small cell lung cancer patients compared to controls. Gene MAP4K4 has been demonstrated to be frequently overexpressed in many types of human cancers, relating to transformation, invasiveness, adhesion, and cell migration. Patients with lung adenocarcinoma and high MAP4K4 expressions have been found to have a shorter overall survival. The lower expression levels of gene DRD3 have been found among patients with non-small cell lung cancer.

We take a closer look at the functional and biological connections of genes involved in each identified regulatory module. Specifically, the gene ontology (GO) enrichment analysis is conducted using DAVID version 6.8 (Sherman et al., 2009). It is observed that the identified modules are biologically meaningful with certain significantly enriched GO terms. For example, in regulatory module #1, genes CACNG6 and RYR3 are enriched with calcium channel activity (GO:0005262, p-value= 0.0042) and calcium ion transport (GO:0006816, p-value=0.0072). Biological studies have found calcium controls cell death and proliferation that are relevant to tumorigenesis, and up or down regulations of specific calcium channels and pumps are associated with cancers. As another example, genes ATP8A2 and DGUOK in regulatory module #20 are enriched with purine nucleoside triphosphate (GO:0009144, p-value=0.0053) and purine nucleoside metabolic process (GO:0006163, p-value=0.008), suggesting the functional and biological connections within the identified module.

Analysis is also conducted using the alternative approaches. In Table C.1 (Appendix), we provide the comparison results, including the numbers of identified main effects and interactions, and numbers of overlapping and RV coefficients between the identifications using different approaches. The RV coefficient measures the common information of two data matrices. It lies between 0 and 1, and a larger value indicates a higher degree of overlapping. We observe that different approaches select significantly different sets of main effects and interactions, with moderate overlapping as measured by the RV coefficients. In practical data analysis, it is difficult to objectively evaluate identification performance. To provide an indirect support, we evaluate prediction performance and selection stability. Specifically, for prediction evaluation, we consider the prediction mean squared error (PMSE) based on 200 random resamplings (9/10 training and 1/10 testing samples). The proposed approach demonstrates competitive performance with the average PMSE= 1.02, compared to 1.25

(Alt.1), 1.16 (Alt.2), 1.05 (Alt.3), and 1.02 (Alt.4). We also assess selection stability using the observed occurrence index (OOI) (Huang et al., 2006). For each identified main effect (interaction), OOI computes its selection frequency in the 200 resamplings, and a larger value suggests higher stability. The proposed approach is observed to have much satisfactory stability with the average OOI value being 0.77, compared to 0.53 (Alt.1), 0.45 (Alt.2), 0.26 (Alt.3), and 0.21 (Alt.4).

4.4.2 Analysis of SKCM data

The response of interest is overall survival, which is subject to censoring. We focus on the primary tumor samples of the Whites. We consider age, AJCC tumor pathologic stage (Stage), gender, and Clark level at diagnosis (Clark), all of which have been suggested as associated with melanoma in the literature. A total of 18,925 gene expression, 23,287 CNV, and 15,616 methylation measurements are available. With the same prescreening as in the previous analysis, the data used for downstream analysis contains 314 gene expression and 686 regulator (397 CNV and 289 methylation) measurements on 231 subjects, of which 139 died during follow-up. The observed times range from 2.04 to 357.10 months with median 56.31.

The proposed analysis identifies 17 regulatory modules, which contain on average 7.60 gene expressions and 6.45 regulators. The graphical presentation is provided in Figure C.1 (Appendix). The AFT model is assumed for modeling survival. A total of 28 main effects and 12 interactions are selected by the proposed approach, among which 14 main effects belong to one identified regulatory module and the remaining are related to the individual molecular units. The identified main effects consist of 15 gene expression and 13 methylation measurements, and the identified interactions consist of 9 with gene expressions and 3 with methylations. The estimated coefficients are presented in Table 4.4. Examining the estimated coefficients suggests that melanoma patients with higher levels of age, Stage, and Clark have a shorter survival. Findings on the molecular variables are also sensible. For instance, gene IMP3 has been found to be associated with cell proliferation and considered as an oncofetal protein-related gene. Its expression level has been used as a diagnostic and prognostic marker from surgical pathology in malignant melanoma. TBC1D7 is one

Table 4.3: Analysis of the LUAD data using the proposed method: identified main effects and interactions.

Group	Type	Gene	Main	Age	Stage	Smoking	Gender
				0.010	-0.031	-0.201	-0.067
1	GE	VIT	0.006				
1	GE	PRH1	0.007				
1	GE	NOXRED1	0.006				
1	GE	RYR3	0.007				
1	GE	SERPINB11	0.007				
1	GE	ZNF273	0.004				
1	GE	WRAP53	0.003				
1	GE	SNORA7B	0.006				
1	GE	GUCY2F	0.007				
1	GE	STATH	0.007				
1	GE	CACNG6	0.007				
1	DM	WIPI2	-0.005				
3	GE	LINC00922	-0.059		0.009	0.001	0.004
3	GE	NDP	-0.059		0.008	0.001	0.004
3	GE	TNMD	-0.055		0.008	0.001	0.004
3	GE	IBSP	-0.055		0.008	0.001	0.004
3	GE	PWRN1	-0.053		0.008	0.001	0.004
3	GE	CACNG3	-0.053		0.008	0.001	0.004
3	DM	MIS18A	-0.045		0.007	0.001	0.003
3	DM	RRP1	-0.036		0.005	0.001	0.002
3	DM	ZDHHC2	-0.044		0.006	0.001	0.003
9	GE	ZXDA	-0.014				
9	GE	EXOSC8	0.022				
9	GE	EPSTI1	0.020				
9	GE	UGT2B4	-0.012				
9	CNV	SLC22A10	0.009				
9	CNV	PABPC5	0.018				
9	DM	ATP8A2	0.010				
9	DM	DHX32	0.013				
15	GE	KL	-0.016				
15	CNV	MAP4K4	-0.013				
15	CNV	KCMF1	-0.016				
15	DM	SATB2	-0.013				
16	GE	HIST1H2AA	-0.009				
16	GE	KCNIP3	-0.008				
16	GE	LRRTM3	-0.011				
16	GE	DCLRE1A	-0.012				
16	GE	PPP1R3D	-0.007				
16	GE	NHLRC2	-0.009				
16	GE	NPAP1	-0.010				
16	CNV	MAP4K4	-0.011				

Continued on the next page

of the down-regulated genes that are potentially causal for the induction of loss of proliferative capacity and terminal differentiation in human melanoma cells. The lack of gene A2M expression provides a growth advantage to melanoma cells by interfering with effec-

Table 4.3: Continued from the previous page.

Group	Type	Gene	Main	Age	Stage	Smoking	Gender
20	GE	FTSJ1	0.013				
20	GE	DGUOK	0.012				
20	GE	SESN3	-0.008				
20	GE	CAPZB	0.009				
20	CNV	PABPC5	0.005				
20	CNV	MRGPRD	0.008				
20	DM	IL17D	0.008				
20	DM	ATP8A2	0.006				
20	DM	DHX32	0.004				
21	GE	AFF3	-0.106			0.147	-0.013
27	GE	SGPP2	-0.009				
47	GE	FNIP2	-0.027				
50	GE	C11orf65	0.005				
68	GE	DRD3	0.012				
102	GE	DPRX	0.026				
124	GE	PRIMA1	-0.016				
178	GE	FAM217B	-0.013				
304	CNV	AK4	0.014				
319	CNV	MIR582	-0.024				
423	DM	HOXA1	-0.027				
520	DM	SDE2	0.014				

tive antigen presentation. IL24 is a novel tumor suppressor gene with tumor-apoptotic and immune-activating properties, and one of several genes that are upregulated during terminal differentiation of melanoma cells. The high expression level of gene ZDHHC4 has been observed in NRAS mutant melanoma cell lines. Published studies have also found a statistically significant overexpression of gene BRF2 in cutaneous melanoma compared to normal skin, and suggested it as a potential marker for patients at risk for metastasis. Gene RBP2 has been shown to directly regulate gene transcription in a reporter assay system as a transcriptional regulator with a tumor suppressive potential in melanoma cells. For the identified module, we further conduct the GO enrichment analysis. It is observed that the involved genes share common GO terms. For example, genes A2M, ENOX1, and IL24 are enriched with extracellular space (GO:0005615, p-value=0.0097), for which published studies have suggested that extracellular vesicles released to extracellular space are correlated with genetic tumor progression in human cancer.

We conduct analysis using the alternatives and summarize the comparison results in Table C.1 (Appendix). Similar patterns as in the previous analysis are observed, where different approaches have small numbers of overlapping identifications and moderate RV

coefficients. We also conduct the prediction and selection stability evaluation. With the censored survival response, we adopt the C statistic to measure prediction accuracy (Uno et al., 2011). A larger value of C statistic indicates better prediction. The proposed approach has an average C statistic 0.60, compared to 0.57 (Alt.1), 0.48 (Alt.2), 0.47 (Alt.3), and 0.59 (Alt.4). In addition, it has superior selection stability with an average OOI of 0.74, compared to 0.56 (Alt.1), 0.50 (Alt.2), 0.38 (Alt.3), and 0.26 (Alt.4). These results provide a strong support to the proposed M-E interaction analysis.

Table 4.4: Analysis of the SKCM data using the proposed method: identified main effects and interactions.

Group	Type	Gene	Main	Age	Stage	Gender	Clark
				-0.176	-0.099	0.150	-0.042
14	GE	MYCNOS	0.003				
14	GE	MRGPRX3	0.004				
14	GE	MFSD6L	0.005				
14	GE	IMP3	0.005				
14	GE	TBC1D7	0.003				
14	GE	A2M	0.004				
14	GE	NEURL2	0.005				
14	GE	IL24	0.004				
14	DM	MAU2	0.004				
14	DM	ZDHHC4	0.004				
14	DM	ENOX1	0.002				
14	DM	PTPN12	0.005				
14	DM	BRF2	0.002				
14	DM	SYT6	0.002				
70	GE	DSTYK	-0.054	0.123	-0.164		
71	GE	GLDN	0.044	-0.058	0.012		
82	GE	RBP2	-0.029	-0.026			
124	GE	SATB2	-0.057	-0.032	0.034		
153	GE	RPL36AL	0.003				
204	GE	RNPS1	-0.057	0.112		0.084	
214	GE	ARL6IP1	-0.014				
573	DM	DPY19L3	0.006				
640	DM	RABEP1	-0.080		-0.071	0.010	
647	DM	SLU7	-0.004				
654	DM	KLHL31	-0.023				
696	DM	GLMP	-0.016				
714	DM	BNIP1	-0.025				
759	DM	MS4A15	0.055	-0.045			

4.5 Discussion

Modeling the outcomes and phenotypes of cancer and other complex diseases is an "old" but still widely open problem. In this Chapter, we have developed the M-E interaction analysis, which is the natural next step of the existing literature. In particular, it is built on but advances from the existing gene-environment interaction analysis by incorporating multiple types of molecular measurements (which have overlapping but more importantly independent information in a single analysis). It also advances from the existing multidimensional molecular data analysis by incorporating interactions and respecting the hierarchical structure. The proposed approach has sound biological and statistical basis. Its working characteristics are carefully examined, and simulation and data analysis have demonstrated its satisfactory performance.

It remains an open question how to best accommodate multidimensional molecular data in modeling. The proposed analysis Step I has been motivated by Wang et al. (2012b), Zhu et al. (2016a), and several other studies. Similar to the literature, linear modeling and regularized estimation have been applied for estimating the regulations. Different from the literature, biclustering has been conducted to identify local regulations, where a small number of co-expressed genes are regulated by a small number of regulators in a coordinated manner. It advances from Zhu et al. (2016a) and others by relaxing the rank-one constraint. The Step II of dimension reduction can be conducted by other techniques such as partial least squares, can effectively reduce dimensionality and remove collinearity, and has been shown as effective in numerical study. There are alternative techniques for interaction analysis in Step III. We have chosen penalization for the consistency of analysis framework. It will be of interest to extend by adopting other estimation/selection techniques. We have used gene expressions and their regulators for description. The proposed approach can be directly applied to other and potentially more complex data structures, thus enjoying broad applicability.

Chapter 5

Application of Interaction Analysis for Histopathological Imaging Data

5.1 Introduction

Cancer is extremely complex. Extensive statistical investigations have been conducted, modeling various cancer outcomes/phenotypes. A long array of measurements from different domains have been used in cancer modeling, including clinical/environmental factors, socioeconomic factors, omics (genetic, genomic, epigenetic, proteomic, etc.) measurements, histopathological imaging features, and others. Yet, none of the existing models is completely satisfactory, and it remains a challenging task to develop new ways of cancer modeling.

Imaging has been playing an irreplaceable role in cancer practice and research (Fass, 2008). It is routine for radiologists to use CT, MRI, PET, and other techniques to generate radiological images, which can inform the size, location, and other "macro" features of tumors (Benzaquen et al., 2019). Biopsies are ordered, and pathologists review the slides of representative sections of tissues to make definitive diagnosis. This procedure generates histopathological (diagnostic) images (Gurcan et al., 2009). Through microscopically examining small pieces of tissues, more "micro" features of tumors are obtained. Histopathological images have been used as the gold standard for diagnosis. More recently, histopatho-

logical imaging features have also been used to model other cancer outcomes/phenotypes. For example, in Yuan et al. (2012), they were used for predicting the prognosis of estrogen receptor-negative breast cancer, and a multivariate Cox regression was adopted. In Tabesh et al. (2007), histopathological imaging features were used in a k-nearest neighbor classifier to assign images into different groups of Gleason tumor grading for prostate cancer patients.

With the complexity of cancer, a single domain of measurement is insufficient, and measurements from multiple sources are needed in modeling (Zhong et al., 2019). In the literature, histopathological imaging features and clinical/environmental risk factors have been combined in an additive manner for modeling cancer outcomes. In Wang et al. (2014), for modeling lung cancer prognosis, clinical factors (including age, gender, cancer type, smoking history, and tumor stage) were combined with imaging features in a multivariate Cox regression model. This study and those alike have shown that combining the two sources of information are more informative than a single source. Our literature review suggests that most if not all of the existing studies have considered the additive effects of histopathological imaging features and clinical/environmental factors, and studies that accommodate their interactions (referred to as "I-E" interactions, with "I" and "E" standing for imaging and clinical/environmental factors, in this Chapter) are lacking. Statistically, adding interactions when the main-effect models are not fully satisfactory is "normal". Biologically speaking, incorporating such interactions have been partly motivated by the success of gene-environment interactions. Specifically, in the literature, the biological rationale and practical success of G-E interactions have been well established (Hunter, 2005). Cancer is a genetic disease. Histopathological images reflect essential information on the histological organization and morphological characteristics of tumor cells and their surrounding tumor microenvironment, which are heavily regulated by tumors' molecular features. As such, from G-E interactions, we may naturally derive I-E interactions. It is noted that I-E and G-E interaction analyses cannot replace each other. More specifically, not all genetic information is contained in imaging features, and histopathological features, as reflected in imaging data, are also affected by factors other than molecular changes.

This study has also been partly motivated by the ineffectiveness of techniques adopted in the existing studies. Histopathological images contain rich information, and the number of

extracted features can be quite large, posing analytic challenges. This dimension problem is "brutally" handled in some studies. For example, in Luo et al. (2017), the univariate Cox model was fit to each imaging feature, and those with the strongest marginal effects were selected. Such features were then used along with clinical characteristics, including age, gender, smoking status, and tumor stage, to construct the final prognostic model. When joint modeling is the ultimate goal, the aforementioned approach may miss truly important signals in the first step of screening. To accommodate the high dimensionality in joint modeling, penalization and other regularization techniques have been adopted. For example, in Yu et al. (2016), the elastic net approach, which combines the Lasso and ridge penalties, was used along with Cox regression. With the differences between interactions and main effects, such methods cannot be directly applied to analysis that accommodates I-E interactions. There are also studies that use advanced deep learning techniques. For example, Bychkov et al. (2018) used the CNN (convolutional neural network) technique to predict colorectal cancer prognosis based on images of tumor tissue samples. Other examples also include Zhu et al. (2017) and Coudray et al. (2018). Such deep learning techniques may excel in prediction, however, usually lack interpretations and also suffer from a lack of stability when sample size is small.

The main objective of this article is to explore accommodating I-E interactions in cancer modeling. Although the concept may seem simple, such an interaction analysis has not been conducted in the literature. The adopted statistical methods have been "borrowed" from G-E interaction analysis. With the connectedness between genetic and histopathological imaging features and parallelization of G-E and I-E interaction analysis, such a strategy is sensible. The proposed interaction analysis strategy and methods are demonstrated using the TCGA lung adenocarcinoma data. Overall, this study may suggest an alternative way of utilizing histopathological imaging data and modeling cancer more accurately.

5.2 Data

We demonstrate I-E interaction analysis using the TCGA lung cancer data. TCGA is a collective effort organized by NCI and has published comprehensive data, especially on

outcomes/phenotypes, clinical/environmental measures, and histopathological images, for lung and other cancer types. Lung cancer is the leading cause of cancer death globally (Boolell et al., 2015), and lung adenocarcinoma (LUAD) is the most common histological subtype and has posed increasing public concerns (Network et al., 2014). The TCGA LUAD data has been analyzed in multiple published studies, including Wang et al. (2014) and Luo et al. (2017) that analyzed histopathological images, and Karlsson et al. (2014) and Li et al. (2014) that conducted analysis on clinical/environmental factors. Thus, it is of interest to "continue" these studies on main additive effects and further examine potential I-E interactions with the TCGA LUAD data. It also has the advantage of having a relatively larger sample size, which is critical to achieve meaningful findings. It is noted that the proposed analysis can be directly applied to data on other cancer types.

We acquire 541 whole slide histopathology images from the TCGA data portal (`https://portal.gdc.cancer.gov/projects/TCGA-LUAD`). To extract imaging features, we adopt the following pipeline developed by Luo et al. (2017). First, as the size of the whole slide images, which is from 300Mb up to 2Gb with 110,000×70,000 pixels, is too huge to be analyzed directly, each image is cropped into sub-images with 500×500 pixels and saved as tiff image files using the Openslide Python library. Analyzing all the sub-images (more than 10 million image tiles in total) is still computationally unfeasible. Thus, twenty representative tiff sub-images that contain mostly (>50%) regions of interest are randomly selected as input for the following process. It is expected that the randomly selected sub-images are representative samples for the overall "population" of sub-images. Such cropping and random selection are common steps in whole slide image processing and widely adopted in published imaging studies (Sun et al., 2018a; Yu et al., 2017, 2016; Zhu et al., 2016b). It is noted that randomly selecting sub-images may lead to imaging features with very small differences (and so affect downstream analysis). However, as our main goal is cancer model building, as opposed to feature selection, such small differences may not be of major concern.

Second, we adopt *CellProfiler* (Soliman, 2015), a platform designed for cell image processing and used in quite a few recent publications, to extract quantitative features from each sub-image. Specifically, image colors are separated based on hematoxylin and eosin

staining, and converted to grayscale for extracting regional features. Next, cell nuclei are detected and segmented so that cell-level features can be specifically measured. Other features such as regional occupation, area fraction, and neighboring architecture are also captured. Irrelevant features such as file size and execution information are excluded from analysis. This procedure results in a total of 772 features which are categorized into the texture, geometry, and holistic groups. Specifically, the texture group contains Haralick, Gabor "wavelet", and Granularity features, which are classic image processing features, measure the texture properties of cells and tissues, and have been examined in a large number of imaging studies. The geometry group contains features that describe the geometry properties (such as area, perimeter, and so on), and those extracted by Zernike moments. The holistic group contains holistic statistics that describe overall information, such as the total area, perimeter and number of nuclei, and nuclear staining area fraction.

Third, for each patient, the features of images are normalized using sample mean at the patient level. Missing values (with a missing rate lower than 20%) are imputed using sample medians.

For clinical/environmental risk factors, we consider age, American Joint Committee on Cancer tumor pathologic stage, tobacco smoking history indicator, and sex. These variables have been suggested as associated with multiple lung cancer outcomes/phenotypes, including those analyzed in this article (Westcott et al., 2015). In particular, Nordquist et al. (2004) found that the mean age at diagnosis of lung adenocarcinoma among never-smokers was significantly higher than that among current smokers, and the never-smokers with lung adenocarcinoma were predominantly female. Studies have shown that tobacco smoking is responsible for 90% of lung cancer (Bryant and Cerfolio, 2007), and has been identified as a negative prognostic factor for lung adenocarcinoma (Landi et al., 2008). In addition, these factors have also been considered in G-E interaction analysis (Wu et al., 2017).

Multiple outcome variables have been analyzed in the literature (Wang et al., 2014). In this article, we consider two important response variables: (a) FEV1: the reference value for the pre-bronchodilator forced expiratory volume in one second in percent. It is an important biomarker for lung capacity. It is continuously distributed, with mean 80.28 and interquartile range [67.00, 96.25]. Data is available for 132 subjects; and (b) overall survival,

which is subject to right censoring. Data is available for 271 subjects, among whom 102 died during follow-up. The mean observed time is 27.47 months, with interquartile range [14.06, 35.00].

Remarks The adopted feature extraction process follows Luo et al. (2017), where the extracted imaging features were used to predict lung cancer prognosis. Similar processes have also been adopted in other publications (Yu et al., 2017, 2016). Different from limited histopathological features recognized visually by pathologists, CellProfiler extracted features are morphological features of tissue texture, cells, nuclei, and neighboring architecture. These features are extracted and measured by comprehensive computer algorithms, and are impossible to be assessed by human eyes. As demonstrated in Luo et al. (2017), quantitative imaging features provide objective and rich information contained in images that can reveal hidden information to decode tumor development and progression in lung cancer. Following the literature (Luo et al., 2017; Sun et al., 2018a; Zhu et al., 2016b), we adopt feature names automatically assigned by CellProfiler, as can be partly seen in Tables 5.1-5.4. These names provide a brief description of the extracted information with the general form "Compartment_FeatureGroup_Feature_Channel_Parameters". For example, features "AreaShape_MedianRadius" and "AreaShape_MaximumRadius" measure the median and maximum radius of the identified tissue, respectively. As in some recent studies (Luo et al., 2017; Sun et al., 2018a; Zhu et al., 2016b), in this Chapter, our goal is not to identify specific imaging features as markers and make biological interpretations. Instead, we aim to conduct better cancer modeling by incorporating I-E interactions. As such, although they may not have simple, explicit biological interpretations, these features are sensible for our analysis.

5.3 Methods

In parallel to G-E interaction analysis (Wu and Ma, 2019), we conduct two types of I-E interaction analysis, namely marginal and joint analysis. The overall flowchart of analysis is provided in Figure 5.1. In marginal analysis, one imaging feature, one clinical/environmental variable (or multiple such variables), and their interaction are analyzed at a time. In joint

analysis, all imaging features, all clinical/environmental variables, and their interactions are analyzed in a single model. The two types of analysis have their own pros and cons and cannot replace each other. We refer to the literature (Witten and Tibshirani, 2010; Zhang et al., 2011) for more detailed discussions on the two types of analysis.

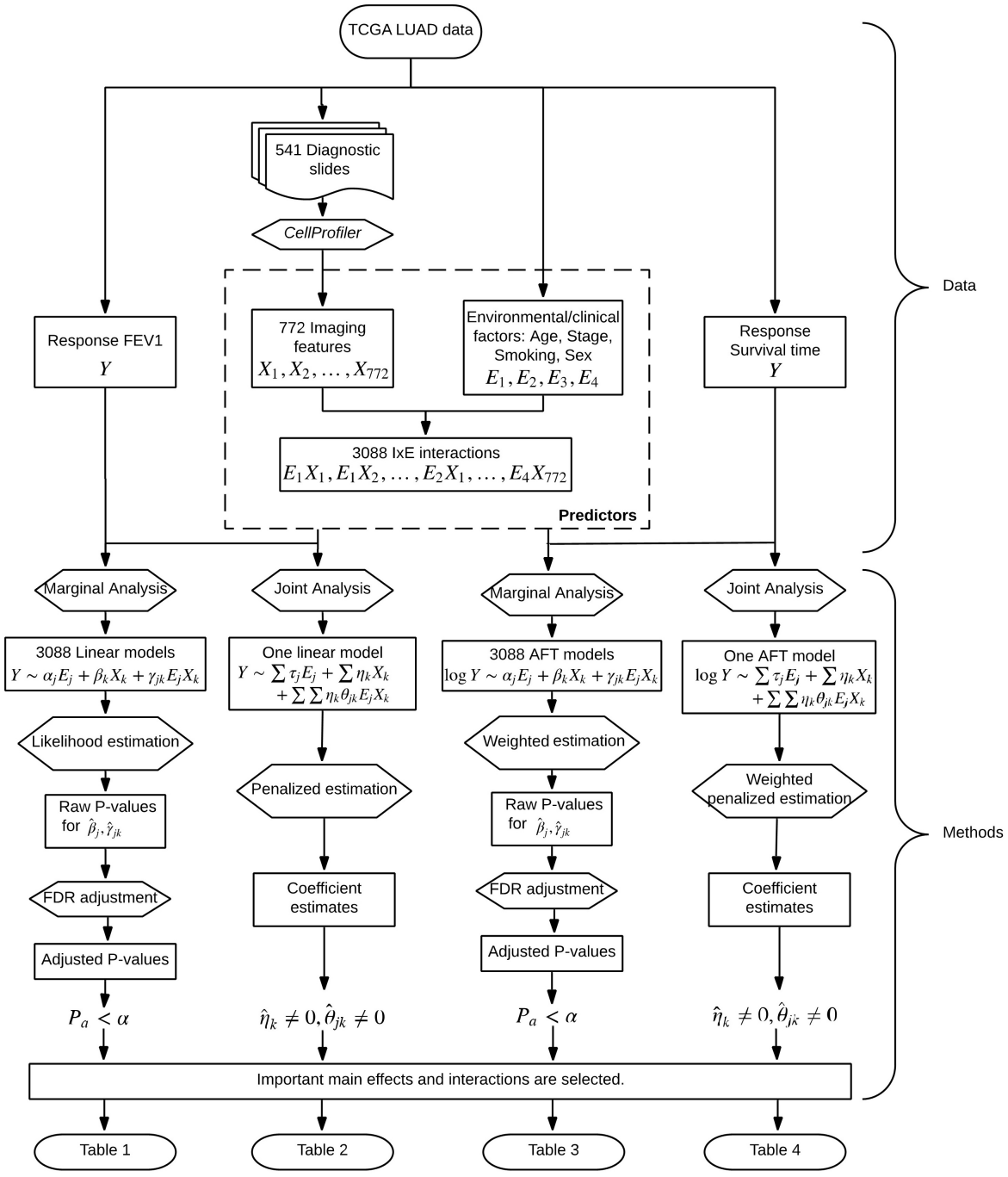


Figure 5.1: Flowchart of the I-E interaction analysis of TCGA LUAD data.

First consider a continuous cancer outcome, which matches the FEV1 analysis. Denote Y as the length N vector of outcome, where N is the sample size. Denote $\mathbf{E} = [E_1, \cdots, E_J]$ as the $N \times J$ matrix of clinical/environmental variables, and $\mathbf{X} = [X_1, \cdots, X_K]$ as the $N \times K$ matrix of imaging features. As represented by the LUAD data, usually clinical/environmental variables are pre-selected and low-dimensional, and imaging features are high-dimensional.

5.3.1 Marginal analysis

Detailed discussions of marginal G-E interaction analysis are available in Xu et al. (2019) and other recent literature. The marginal I-E interaction analysis proceeds as follows. First assume that Y, $\mathbf{E}$, and $\mathbf{X}$ have been properly centered.

(a) For $j = 1, \ldots, J$ and $k = 1, \ldots, K$, consider the linear regression model

$$Y = \alpha_j E_j + \beta_k X_k + \gamma_{jk} E_j X_k + \epsilon, \tag{5.1}$$

where α_j and β_k respectively represent the main effects of the j^{th} clinical/environmental factor and the k^{th} imaging feature, γ_{jk} is the interactive effect, and ϵ is the random error. A total of $J \times K$ models are built.

(b) As each model has a low dimension, estimates can be obtained using standard likelihood based approaches and existing software. P-values can be obtained accordingly.

(c) Interactions (and main effects) with small P-values are identified as important. When more definitive conclusions are needed, the FDR (false discovery rate) or Bonferroni approach can be applied.

It is noted that in Step (a), one clinical/environmental variable is analyzed in each model, which follows Xu et al. (2019). It is also possible to accommodate all clinical/environmental variables in each model. In Step (c), discoveries can be made on interactions only or interactions and main effects combined. Advantages of marginal analysis include its computational simplicity and stability. On the negative side, with the complexity of cancer, an outcome/phenotype is usually associated with multiple imaging features and clini-

cal/environmental variables. As such, each marginal model can be "mis-specified" or "sub-optimal". In addition, there is a lack of attention to the differences between interactions and main effects.

5.3.2 Joint analysis

Joint analysis can tackle some limitations of marginal analysis, and is getting increasingly popular in statistical and bioinformatics literature. It proceeds as follows.

(a) Consider the joint model

$$Y = \sum_{j=1}^{J} \tau_j E_j + \sum_{k=1}^{K} \eta_k X_k + \sum_{j=1}^{J} \sum_{k=1}^{K} \eta_k \theta_{jk} E_j X_k + \epsilon, \tag{5.2}$$

where τ_j and η_k are the main effects of the j^{th} environmental factor and the k^{th} imaging feature, respectively, and the product of η_k and θ_{jk} corresponds to the interaction.

(b) For estimation, consider the Lasso penalization

$$\min_{\eta_k, \theta_{jk}} ||Y - f(\mathbf{E}, \mathbf{X})||^2 + \lambda_1 \sum_k |\eta_k| + \lambda_2 \sum_j \sum_k |\theta_{jk}|, \tag{5.3}$$

where $f(\mathbf{E}, \mathbf{X}) = \sum_j \tau_j E_j + \sum_k \eta_k X_k + \sum_j \sum_k \eta_k \theta_{jk} E_j X_k$, and $\lambda_1, \lambda_2 > 0$ are tuning parameters. In numerical study, we select the tuning parameters using the extended Bayesian information criterion (Chen and Chen, 2008).

(c) Interactions (and main effects) with nonzero estimates are identified as being associated with the outcome.

5.3.3 Accommodating survival outcomes

Consider cancer survival. Denote T as the N-vector of survival times. Below we describe joint analysis, and marginal analysis can be conducted accordingly. We adopt the AFT (accelerated failure time) model, under which

$$\log(T) = \sum_{j=1}^{J} \tau_j E_j + \sum_{k=1}^{K} \eta_k X_k + \sum_{j=1}^{J} \sum_{k=1}^{K} \eta_k \theta_{jk} E_j X_k + \epsilon, \tag{5.4}$$

where notations have similar implications as in the above section. With high-dimensional data, the AFT model has been widely adopted because of its lucid interpretation and more importantly computational simplicity (Huang et al., 2006). Under right censoring, denote C as the N-vector of censoring times, $Y = \log(\min(T, C))$, and $\delta = I(T \leq C)$, where operations are taken component-wise. To accommodate censoring, a weighted approach is adopted. Assume that data have been sorted according to Y_i's from the smallest to the largest. The Kaplan-Meier weights can be computed as $w_1 = \frac{\delta_1}{N}$, $w_i = \frac{\delta_i}{N-i+1} \prod_{j=1}^{i-1} \left(\frac{N-j}{N-j+1}\right)^{\delta_j}$, $i = 2, \ldots, N$. Similar to (5.3), consider the penalized estimation

$$\min_{\eta_k, \theta_{jk}} ||\sqrt{w} \times (Y - f(\mathbf{E}, \mathbf{X}))||^2 + \lambda_1 \sum_k |\eta_k| + \lambda_2 \sum_j \sum_k |\theta_{jk}|, \tag{5.5}$$

where the square root and multiplication are taken component-wise. Interpretations and other operations are the same as for continuous outcomes.

In joint analysis, the most prominent challenge is the high dimensionality. Here the penalization technique is adopted, which can simultaneously accommodate high dimensionality and identify relevant interactions/main effects. Another feature of this analysis that is worth highlighting is that it respects the "main effects, interactions" hierarchy. That is, if an I-E interaction is identified, the corresponding main imaging feature effect is automatically identified. It has been suggested that, statistically and biologically, it is critical to respect this hierarchy (Choi et al., 2010). We refer to the literature (Bien et al., 2013; Liu et al., 2013) for alternative penalization and other joint interaction analysis methods. Compared to marginal analysis, joint analysis can be computationally more challenging, and well-developed software packages are still limited. In addition, the analysis results can be less stable.

The proposed analysis can be effectively realized. To facilitate data analysis within and beyond this study, we have developed R code and made it publicly available at `www.github.com/shuanggema`.

Table 5.1: Marginal analysis of FEV1: identified main effects and interactions, with raw P-values P_r.

Feature group	Feature name		Estimate	P_r
Geometry	AreaShape_Zernike_2_2	Main	0.270	0.002
Geometry	AreaShape_Zernike_5_3	Main	-0.319	0.001
Geometry	Mean_Identifyhemasub2_AreaShape_Zernike_9_9	Main	-0.259	0.004
Geometry	Median_Identifyhemasub2_AreaShape_Zernike_7_1	Main	-0.249	0.005
Geometry	Median_Identifyhemasub2_AreaShape_Zernike_8_6	Main	-0.272	0.003
Texture	StDev_Identifyeosinprimarycytoplasm_Texture_Correlation_maskosingray_3_01	Main	0.280	0.002
Geometry	StDev_Identifyhemasub2_AreaShape_Zernike_8_8	Main	-0.251	0.005
Geometry	StDev_Identifyhemasub2_AreaShape_Zernike_9_1	Main	-0.259	0.004
Geometry	StDev_Identifyhemasub2_AreaShape_Center_Y	Sex	0.291	0.002
Geometry	StDev_Identifyhemasub2_AreaShape_Zernike_8_2	Sex	0.304	0.001
Geometry	StDev_Identifyhemasub2_Location_Center_Y	Sex	0.294	0.002

Table 5.2: Joint analysis of FEV1: identified main effects and interactions.

Feature group	Feature name	Main	Age	Stage	Smoking	Sex
			-0.049	-0.052	-0.002	0.006
Geometry	AreaShape_Zernike_2_2	0.163	0.040	-0.014	-0.185	
Geometry	AreaShape_Zernike_5_3	-0.053				
Geometry	AreaShape_Zernike_6_0	-0.034				
Texture	Granularity_10_ImageAfterMath	0.137	0.110	-0.020		0.064
Geometry	Location_Center_X	0.002				
Geometry	Mean_Identifyeosinprimarycytoplasm_Location_Center_X	0.005				
Geometry	Median_Identifyhemasub2_AreaShape_Zernike_7_1	-0.127	-0.073		0.072	0.003
Geometry	StDev_Identifyhemasub2_AreaShape_Zernike_8_2	-0.170		-0.083		0.188
Texture	StDev_Identifyhemasub2_Granularity_6_ImageAfterMath	-0.029				
Texture	Texture_AngularSecondMoment_ImageAfterMath_3_00	-0.044				
Texture	Texture_AngularSecondMoment_ImageAfterMath_3_03	-0.010				

5.4 Results

5.4.1 Analysis of FEV1

Marginal analysis After the FDR adjustment, none of the main effects or interactions is statistically significant. In Table 5.1, we present the main effects and interactions with the smallest (unadjusted) P-values. The top ranked main effects are from the Geometry and Texture groups, and the top ranked interactions are from the Geometry group and with sex.

Based on the analysis results, we conduct a power calculation. First assume the current levels of estimated effects and their variations. Then with a sample size of 224, the top ranked I-E interactions can be identified as significant with target FDR 0.1. Second, consider the current sample size and levels of variations. Then an effect of -0.35 can be identified as significant with target FDR 0.1.

For comparison, we conduct the analysis of main effects (without interactions). The top

eight main effects (with the smallest P-values) have four overlaps with those in Table 5.1, suggesting that accommodating interactions can lead to different findings.

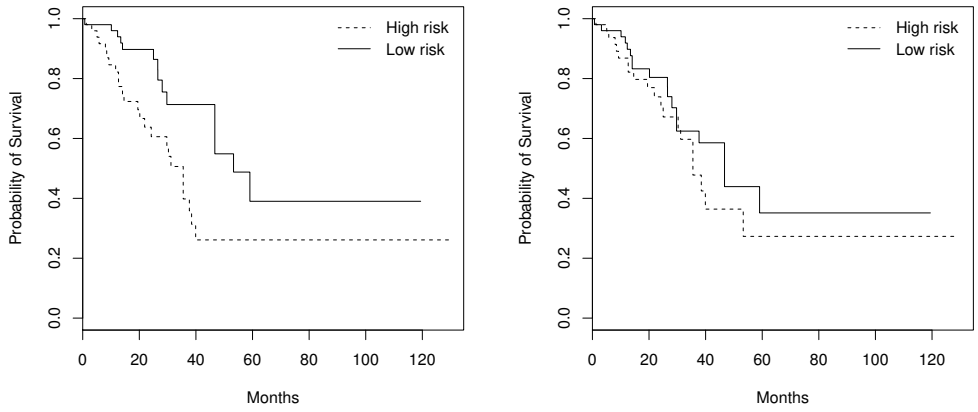
Joint analysis The analysis results are provided in Table 5.2. A total of 11 imaging features are identified, representing the Geometry and Texture groups. A total of 11 interactions are identified, with all four clinical/environmental variables.

For comparison, we consider the joint model with all clinical/environmental variables and imaging features but no interactions. Lasso penalization is applied for selection and estimation. A total of eight imaging features are identified, with one overlapping with those in Table 5.2. We further compute the RV coefficient, which may more objectively quantify the amount of "overlapping information" between two analyses. Specifically, it measures the "correlation" between two data matrices of important effects identified by two different approaches, with a larger value indicating higher similarity. The RV coefficient is 0.24, suggesting a mild level of overlapping.

A significant advantage of joint analysis is that it can lead to a predictive model for the outcome variable. We conduct the evaluation of prediction based on a resampling procedure, which may provide support to the validity of analysis. Specifically, we split data into a training and a testing set, generate estimates using the training data, and make prediction for the testing set subjects. The PMSE (prediction mean squared error) is then computed. This procedure is repeated 100 times, and the mean PMSE is computed. The I-E interaction model has a mean PMSE of 0.84, whereas the main-effect-only model has a mean PMSE of 1.12. This significant improvement suggests the benefit of accommodating interactions.

5.4.2 Analysis of overall survival

Marginal analysis The analysis results are provided in Table 5.3, where we present estimates, raw P-values, as well as the FDR adjusted P-values. Three imaging features from the Holistic group have the FDR adjusted P-values < 0.1. And 36 imaging features from the Geometry group and 24 features from the Texture group are identified as having interactions with Smoking, the most important environmental factor for lung cancer. Compared to the above analysis, more "signals" are identified. Note that the effective sample size is smaller



(a) (b)

Figure 5.2: Kaplan-Meier curves of high and low risk groups identified by the approach that accommodates interactions (left; logrank test P-value 0.007) and the one with main effects only (right; logrank test P-value 0.320).

than that above. As such, the smaller P-values are likely to be caused by stronger signals.

For comparison, we conduct the analysis of main effects. One imaging feature is identified as having FDR adjusted P-value <0.1, which is also identified in Table 5.3. With the complexity of lung cancer prognosis, the interaction analysis, which identifies more effects, can be more sensible.

Joint analysis The analysis results are provided in Table 5.4. A total of 31 imaging features are identified, representing the three feature groups. Two imaging features are identified as interacting with two and four clinical/environmental variables, respectively.

The analysis of main effects is conducted using the Lasso penalization. A total of two imaging features are identified, with one overlapping with those in Table 5.4. The RV coefficient is computed as 0.40, representing a moderate level of overlapping. As with FEV1, prediction evaluation is also conducted based on resampling. For the testing set, subjects are classified into low and high risk groups with equal sizes based on the predicted survival times, where subjects with predicted survival times larger than the median are classified into the low risk group. For one resampling of training and testing sets, in Figure 2, we plot the Kaplan-Meier curves estimated using the observed survival times for the predicted low and high risk groups, along with those generated under the additive main-effect model. Com-

Table 5.3: Marginal analysis of overall survival: identified main effects and interactions, with raw P-values P_r and FDR adjusted P-values P_a.

Feature group	Feature name		Estimate	P_r	P_a
Holistic	Threshold_FinalThreshold_Identifyeosinprimarycytoplasm	Main	-0.301	0	0.095
Holistic	Threshold_OrigThreshold_Identifyeosinprimarycytoplasm	Main	-0.301	0	0.095
Holistic	Threshold_WeightedVariance_identifyhemaprimarynuclei	Main	-0.360	0	0.077
Geometry	AreaShape_Area	Smoking	0.253	0.004	0.078
Geometry	AreaShape_MaximumRadius	Smoking	0.266	0.004	0.074
Geometry	AreaShape_MeanRadius	Smoking	0.265	0.005	0.079
Geometry	AreaShape_MedianRadius	Smoking	0.266	0.005	0.079
Geometry	AreaShape_MinFeretDiameter	Smoking	0.257	0.003	0.073
Geometry	AreaShape_MinorAxisLength	Smoking	0.264	0.002	0.07
Geometry	AreaShape_Zernike_4_4	Smoking	-0.241	0.005	0.079
Geometry	AreaShape_Zernike_7_3	Smoking	-0.308	0	0.027
Geometry	AreaShape_Zernike_8_4	Smoking	-0.242	0.007	0.096
Geometry	AreaShape_Zernike_8_6	Smoking	-0.252	0.005	0.079
Geometry	AreaShape_Zernike_9_1	Smoking	-0.303	0	0.027
Texture	Granularity_13_ImageAfterMath.1	Smoking	-0.317	0.001	0.054
Texture	Mean_Identifyeosinprimarycytoplasm_Texture_Correlation_maskosingray_3_03	Smoking	0.232	0.005	0.079
Geometry	Mean_Identifyhemasub2_AreaShape_Area	Smoking	0.297	0.001	0.049
Geometry	Mean_Identifyhemasub2_AreaShape_MaximumRadius	Smoking	0.318	0.001	0.049
Geometry	Mean_Identifyhemasub2_AreaShape_MeanRadius	Smoking	0.318	0.001	0.049
Geometry	Mean_Identifyhemasub2_AreaShape_MedianRadius	Smoking	0.308	0.002	0.054
Geometry	Mean_Identifyhemasub2_AreaShape_MinFeretDiameter	Smoking	0.299	0.001	0.049
Geometry	Mean_Identifyhemasub2_AreaShape_MinorAxisLength	Smoking	0.310	0.001	0.045
Geometry	Mean_Identifyhemasub2_AreaShape_Zernike_4_4	Smoking	-0.263	0.003	0.07
Geometry	Mean_Identifyhemasub2_AreaShape_Zernike_5_1	Smoking	-0.268	0.002	0.07
Geometry	Mean_Identifyhemasub2_AreaShape_Zernike_8_2	Smoking	-0.277	0.003	0.073
Geometry	Mean_Identifyhemasub2_AreaShape_Zernike_8_8	Smoking	-0.290	0.003	0.073
Geometry	Mean_Identifyhemasub2_AreaShape_Zernike_9_1	Smoking	-0.226	0.004	0.074
Texture	Mean_Identifyhemasub2_Granularity_13_ImageAfterMath	Smoking	-0.325	0.001	0.054
Texture	Mean_Identifyhemasub2_Texture_Correlation_ImageAfterMath_3_01	Smoking	0.330	0	0.039
Texture	Mean_Identifyhemasub2_Texture_Correlation_ImageAfterMath_3_02	Smoking	0.297	0.002	0.07
Texture	Mean_Identifyhemasub2_Texture_Correlation_ImageAfterMath_3_03	Smoking	0.397	0	0.01
Texture	Mean_Identifyhemasub2_Texture_SumVariance_ImageAfterMath_3_02	Smoking	0.258	0.007	0.093
Texture	Median_Identifyeosinprimarycytoplasm_Texture_Correlation_maskosingray_3_03	Smoking	0.233	0.004	0.079
Geometry	Median_Identifyhemasub2_AreaShape_Area	Smoking	0.344	0	0.027
Geometry	Median_Identifyhemasub2_AreaShape_MaxFeretDiameter	Smoking	0.242	0.005	0.079
Geometry	Median_Identifyhemasub2_AreaShape_MaximumRadius	Smoking	0.323	0.001	0.049
Geometry	Median_Identifyhemasub2_AreaShape_MeanRadius	Smoking	0.323	0.001	0.049
Geometry	Median_Identifyhemasub2_AreaShape_MedianRadius	Smoking	0.266	0.005	0.079
Geometry	Median_Identifyhemasub2_AreaShape_MinFeretDiameter	Smoking	0.346	0	0.027
Geometry	Median_Identifyhemasub2_AreaShape_MinorAxisLength	Smoking	0.342	0	0.027
Geometry	Median_Identifyhemasub2_AreaShape_Perimeter	Smoking	0.247	0.006	0.085
Geometry	Median_Identifyhemasub2_AreaShape_Zernike_4_4	Smoking	-0.242	0.002	0.059
Geometry	Median_Identifyhemasub2_AreaShape_Zernike_5_1	Smoking	-0.256	0.003	0.073
Texture	Median_Identifyhemasub2_Granularity_13_ImageAfterMath	Smoking	-0.311	0.001	0.049
Texture	Median_Identifyhemasub2_Texture_Correlation_ImageAfterMath_3_01	Smoking	0.319	0.001	0.049
Texture	Median_Identifyhemasub2_Texture_Correlation_ImageAfterMath_3_02	Smoking	0.274	0.005	0.081
Texture	Median_Identifyhemasub2_Texture_Correlation_ImageAfterMath_3_03	Smoking	0.394	0	0.01
Texture	StDev_Identifyeosinprimarycytoplasm_Texture_SumAverage_maskosingray_3_00	Smoking	0.272	0.003	0.073
Texture	StDev_Identifyeosinprimarycytoplasm_Texture_SumAverage_maskosingray_3_01	Smoking	0.273	0.003	0.073
Texture	StDev_Identifyeosinprimarycytoplasm_Texture_SumAverage_maskosingray_3_02	Smoking	0.270	0.004	0.074
Texture	StDev_Identifyeosinprimarycytoplasm_Texture_SumAverage_maskosingray_3_03	Smoking	0.275	0.003	0.073
Geometry	StDev_identifyhemaprimarynuclei_Location_Center_Y	Smoking	-0.245	0.007	0.093
Geometry	StDev_Identifyhemasub2_AreaShape_Zernike_8_4	Smoking	-0.280	0.001	0.045
Geometry	StDev_Identifyhemasub2_AreaShape_Zernike_8_8	Smoking	-0.236	0.007	0.094
Texture	StDev_Identifyhemasub2_Texture_SumVariance_ImageAfterMath_3_01	Smoking	0.266	0.007	0.096
Texture	StDev_Identifyhemasub2_Texture_SumVariance_ImageAfterMath_3_02	Smoking	0.283	0.005	0.079
Texture	StDev_Identifyhemasub2_Texture_SumVariance_ImageAfterMath_3_03	Smoking	0.283	0.006	0.084
Geometry	StDev_identifytissueregion_Location_Center_Y	Smoking	-0.289	0.002	0.059
Texture	Texture_Correlation_ImageAfterMath_3_01	Smoking	0.252	0.004	0.078
Texture	Texture_Correlation_ImageAfterMath_3_03	Smoking	0.329	0	0.027
Texture	Texture_Correlation_maskosingray_3_03	Smoking	0.237	0.004	0.074
Texture	Texture_Entropy_ImageAfterMath_3_01	Smoking	0.220	0.007	0.093
Texture	Texture_Entropy_ImageAfterMath_3_03	Smoking	0.233	0.004	0.074

pared to the main-effect model, it is obvious that the two risk groups identified by the I-E interaction model have a much clearer separation of the survival functions, indicating better prediction performance. To be more rigorous, we further conduct a logrank test, which is a nonparametric test for comparing the survival distributions of two subject groups. With 100 resamplings, the average logrank statistics are 7.28 (I-E interaction model, P-value=0.007) and 0.99 (main-effect model, P-value=0.320), respectively. The superior prediction performance of the I-E interaction models suggests that incorporating interactions can lead to clinically more powerful models, justifying the value of the proposed analysis.

Table 5.4: Joint analysis of overall survival: identified main effects and interactions.

Feature group	**Feature name**	**Main**	**Age**	**Stage**	**Smoking**	**Sex**
			-0.024	-0.317	-0.038	-0.088
Geometry	AreaShape_Zernike_6_0	-0.038				
Geometry	AreaShape_Zernike_6_4	-0.019				
Geometry	AreaShape_Zernike_6_6	0.052				
Geometry	AreaShape_Zernike_9_3	0.027				
Geometry	AreaShape_Zernike_9_5	0.153				
Texture	Granularity_10_ImageAfterMath.1	-0.033				
Texture	Granularity_9_ImageAfterMath	0.081				
Geometry	Mean_Identifyhemasub2_AreaShape_Center_X	0.002				
Geometry	Mean_Identifyhemasub2_AreaShape_Zernike_5_1	0.013				
Geometry	Mean_Identifyhemasub2_AreaShape_Zernike_6_2	-0.002				
Geometry	Mean_Identifyhemasub2_AreaShape_Zernike_6_4	-0.010				
Geometry	Mean_Identifyhemasub2_AreaShape_Zernike_9_9	-0.146				
Geometry	Mean_Identifyhemasub2_Location_Center_X	0.002				
Geometry	Mean_identifytissueregion_Location_Center_X	0.056				
Geometry	Median_Identifyeosinprimarycytoplasm_Location_Center_X	-0.071				
Geometry	Median_Identifyhemasub2_AreaShape_Zernike_4_0	0.023				
Geometry	Median_Identifyhemasub2_AreaShape_Zernike_7_3	0.083				
Geometry	Median_Identifyhemasub2_AreaShape_Zernike_8_4	-0.120				
Geometry	Median_Identifyhemasub2_AreaShape_Zernike_8_6	-0.098				
Geometry	Median_Identifyhemasub2_AreaShape_Zernike_9_1	-0.044				
Geometry	Median_identifytissueregion_Location_Center_Y	-0.063				
Holistic	Neighbors_SecondClosestDistance_Adjacent	-0.170		-0.072	0.002	
Geometry	StDev_Identifyeosinprimarycytoplasm_Location_Center_Y	0.095				
Texture	StDev_Identifyeosinprimarycytoplasm_Texture _DifferenceVariance_maskosingray_3_00	0.036				
Geometry	StDev_Identifyhemasub2_AreaShape_Orientation	-0.159				
Geometry	StDev_Identifyhemasub2_AreaShape_Zernike_8_8	-0.146				
Texture	StDev_Identifyhemasub2_Granularity_12_ImageAfterMath	-0.101				
Texture	StDev_Identifyhemasub2_Granularity_13_ImageAfterMath	0.327	0.130	0.072	-0.189	0.174
Texture	StDev_Identifyhemasub2_Granularity_9_ImageAfterMath	0.003				
Texture	StDev_Identifyhemasub2_Texture_SumVariance _ImageAfterMath_3_01	-0.034				
Geometry	StDev_identifytissueregion_Location_Center_Y	0.016				

5.4.3 Simulation

Comparatively, joint analysis is newer and has been less conducted. To gain more insights into the validity of findings from our joint interaction analysis, we conduct a set of data-

based simulation. Specifically, the observed imaging features and clinical/environmental factors are used. To generate variations across simulation replicates, we use resampling, with sample sizes set as 200. The "signals" and their levels are set as those in Tables 5.2 and 5.4, respectively. For both the continuous and (log) survival outcomes, we generate random errors from $N(0, 1)$. For the survival setting, we generate the censoring times from randomly sampling the observed. The Lasso-based penalization approach is then applied, with tuning parameters selected using the extended BIC approach. To evaluate identification, TP (true positive) and FP (false positive) values are computed. Summary statistics are computed based on 100 replicates. Under the continuous outcome setting, there are 11 true main effects and 11 I-E interactions. For main effects, the TP and FP values are 9.75 (1.65) and 3.15 (1.39), respectively, where numbers in "()" are standard deviations. For interactions, the TP and FP values are 7.35 (0.99) and 0.05 (0.22), respectively. Under the censored survival outcome setting, there are 31 true main effects and 6 I-E interactions. For main effects, the TP and FP values are 24.41 (3.98) and 13.90 (2.47), respectively. For interactions, the TP and FP values are 3.24 (0.21) and 0.24 (0.12), respectively. Overall, at the estimated signal levels and with the observed feature distributions, the joint analysis is capable of identifying the majority of true interactions and main effects, with a moderate number of false discoveries. This provides a high level of confidence to the joint interaction analysis.

5.5 Discussion

Histopathological imaging analysis has been routine in cancer diagnosis, and recently, its application in the analysis of cancer biomarkers, outcomes, and phenotypes has been explored. This study has taken a natural next step and conducted the imaging-environment interaction analysis. Statistically and biologically speaking, the analysis has been partly motivated by G-E interaction analysis. It is noted that the statistical methods themselves have been almost fully "translated" from G-E interaction analysis. As I-E interaction analysis has not been conducted in published cancer modeling studies, it is sensible to first employ well-developed methods, and in the future, methods that are more tailored to imaging data may be developed. We also note that in cancer modeling and other biomedical fields, it

is not uncommon to apply methods well developed in one field to other new fields. The proposed I-E interaction analysis, especially joint analysis, may seem considerably more complex than some cancer modeling approaches. With the complexity of cancer, models with a few variables and simple statistical analysis are getting increasingly insufficient. Published studies have suggested that advanced statistical techniques and complex models are needed. Recent developments for lung cancer, including the elastic net-Cox analysis (Yu et al., 2016), deep convolutional neural network (Coudray et al., 2018), and deep network based on convolutional and recurrent architectures (Bychkov et al., 2018), have comparable or higher levels of complexity compared to the proposed analysis. Artificial intelligence (AI) techniques, which have been recently used for cancer modeling in particular including the radiomics analysis of non-small-cell lung cancer (Hosny et al., 2018; Thrall et al., 2018), have even higher levels of complexity. We conjecture that such complexity will also be needed for future developments in cancer modeling using imaging data. The increasing complexity in cancer modeling seems to be an inevitable trend, and domain specific expertise is a must for such analysis.

We have analyzed the TCGA LUAD data with a continuous and a censored survival outcome. This choice has been motivated by the clinical importance of lung adenocarcinoma as well as data availability (a larger sample size). It is noted that the proposed analysis and R program will be directly applicable to the analysis of data on other cancer types. I-E interactions have been identified in both marginal and joint analysis, for both FEV1 and overall survival. There is one prominent difference between imaging and genetic/clinical data. With extensive investigations and functional experiments, the biological and biomedical implications of most clinical/environmental factors and genes are at least partially known. It is thus possible to evaluate whether G-E interactions are biologically sensible. The circumstance is significantly different for histopathological imaging features. The rationale and algorithms for feature extraction have been made clear in the developments of CellProfiler and other software. However, the identified features do not have lucid biological interpretations. As such, we are not able to objectively assess the biological implications of the findings in Table 5.1 -5.4. It is noted that this limitation is also shared by recently published imaging studies Luo et al. (2017); Sun et al. (2018a); Zhu et al. (2016b), which have

unambiguously demonstrated the great value of such imaging features in cancer modeling. It is also noted that imaging features derived from computer-aided pathological analysis have the unique advantage of being objective and comprehensive, and can reveal hidden information contained in histopathological images that cannot be recognized or assessed by pathologists. Our statistical evaluations, including the prediction evaluation and data-based simulation, can provide support to the analysis results to a great extent. In general, more investigations into the biological implications of the computer-program-extracted imaging features will be needed.

This study has suggested a new venue for cancer modeling. Although findings made on LUAD may not be applicable to other cancers, the analysis technique and R program will be broadly applicable. Following the flowchart in Figure 5.1 and detailed steps described in this article, and using the publicly available R program, cancer biostatisticians and clinicians should be able to carry out the proposed analysis with their own data. More specifically, with their own clinical/environmental and imaging data, they will be able to construct models for prognosis and other outcomes/phenotypes. Such models, as other cancer models (for example those using omics data), can be used to assist clinical decision making. Overall, this study may help advance the challenging field of cancer modeling.

Chapter 6

Concluding Remarks

In sum, we proposed four statistical methods for analyzing G-E interactions and presented one application using both marginal and joint models for imaging data. In simulation studies, improved identification and prediction performance was produced in comparison with multiple alternatives. Besides numerical studies of the proposed methods, we also conducted data analyses using TCGA data on multiple cancer types. Sensible findings of important G-E interactions with superior stability and prediction were made and interpreted using the published literature. In addition, we developed R code for the proposed methods and made it publicly available for researchers.

Two generic paradigms of marginal and joint modeling frameworks in G-E interaction analysis are extensively explored and compared. On the one hand, marginal models enjoy computational simplicity and are straightforward to understand. Most of the existing studies are based on a marginal framework, yet marginal models are not able to predict the outcomes, limited to marker identifications. On the other hand, joint modeling that requires penalized estimation becomes increasingly popular in the literature. Though computational cost is relatively higher compared to marginal models, interpretable selection and accurate prediction can be achieved by the joint modeling. In this dissertation, we do not reach a definitive conclusion in the competition of marginal and joint modeling frameworks for G-E interaction analysis.

The first approach in Chapter 2.2 was built on the quantile regression technique, used weights to easily accommodate censoring, and adopted partial correlation to identify impor-

tant interactions while properly controlling for the main genetic and environmental effects. The second approach in Chapter 2.3 employed the trimmed regression under joint modeling, applied penalization and stability selection to identify important G-E interactions, and respected the "main effects, interactions" hierarchical structure. These two proposed methods can accommodate prognostic outcomes and demonstrated that robust methods are capable of improving identification accuracy. The third approach in Chapter 3 utilized penalization under a marginal analysis framework. We constructed the penalty terms for incorporating multiple types of additional information and for selecting hierarchal interactions. The hierarchical structure was enforced by coefficient decomposition and tailored penalization. The last proposed approach in Chapter 4 integrated multidimensional molecular measurements and sufficiently accounted for their overlapping as well as independent information. The proposed joint estimation was based on the penalization technique and had solid statistical properties, leading to improved estimation and prediction.

6.1 Limitations

Though methodological advancement was made by the proposed novel and useful approaches, this dissertation of G-E interaction analysis inevitably has limitations. Following linear regression, the effects of G-E interactions were uniformly described by their coefficients. Many advantages were introduced by this regression framework, including the coefficient decomposition strategy that enforces the hierarchical structure. Nonetheless, there exist several other schemes for analyzing interactions. For instance, Ren et al. (2019) proposed a semiparametric Bayesian model that includes linear and non-linear G-E interactions simultaneously. A partially linear varying coefficient model was adopted where a smoothing varying coefficient function of the environmental risk factor was used for describing the non-linear interactions. Other penalization techniques for selecting hierarchical interactions include the hierarchical Lasso (Bien et al., 2013), the sparse group Lasso (Simon et al., 2013), and many others (Hao and Zhang, 2017). In this study, the simplicity and interpretability brought by the linear regression framework were well appreciated and we note its potential for extending to more complex models in future work.

Another limitation of this dissertation is that statistical inference was not comprehensively discussed. The proposed methods focused more on methodological development and numerical examinations. Theoretical derivation for the proposed methods under high-dimensional settings was much limited to draw more definitive conclusions. In the dissertation, theoretical justifications were made in a heuristic manner. In fact, the statistical techniques that we adopted in these studies have been extensively investigated and the relevant statistical properties has already been examined in the literature. Although more rigorous justifications were not readily available for more sophisticated approaches, we note that the building blocks of the proposed methods have established grounds in statistical properties to well support this dissertation.

This study is also limited to the TCGA data in real data analysis. As one of the largest publicly available and high-quality data sources for cancer genomic studies, we chose the TCGA data and focused mainly on lung adenocarcinoma and cutaneous melanoma datasets. Additionally, our proposed methods included different sets of environmental risk factors, for example, Breslow's depth in Chapter 2.2 as one of the E factors whereas in Chapter 2.3 as the response of interest. Such intensive investigation of the TCGA datasets is widely accepted and especially common among the publications for demonstrating methodological advancement. We also note that different results were made across the proposed methods. Due to the scope of the dissertation, an explicit comparison between those findings was not discussed.

6.2 Future work

Overall, we recognized the limitations of the existing methodologies for G-E interaction analysis, conducted comprehensive investigations for novel and useful methods to tackle practical problems and to advance in methodological development, and analyzed simulated and real datasets to validate the superiority and applicability of the proposed approaches. It is challenging to identify important G-E interactions for complex diseases. The proposed methods presented in this dissertation suggest new venues of interaction analysis and allow future advancements to build on. For example, one can extend the goodness-of-fit term

in the proposed objective functions to other loss functions, such as the absolute-value-based ones. All of the proposed methods assume linear regression models, which can be replaced by non-linear and nonparametric models. As the most fundamental elements in G-E interaction analysis were thoroughly addressed in this dissertation, future advancement in methodology that can be added to the proposed methods becomes natural.

Besides, several analysis strategies, such as coefficient decomposition, which tackled the essential complications in interaction analysis, can be further adapted and extended to other research fields. Inspired by our application to histopathological imaging data, the proposed methods are generally applicable to interaction analysis for high-dimensional data and can be regarded as statistical methods for interaction selection. In fact, the analysis of the TCGA data can serve as a prototype and applications to data on complex diseases other than cancer are desired. Future work in analyzing data on various cancer types and other complex diseases will potentially contribute to better understanding of the underlying biological mechanisms of disease development. Extension from genetic factors to other high-dimensional measurements is flexible and will also add value to the dissertation.

Appendix A

Chapter 2

A.1 Censored Quantile Partial Correlation for Cancer Prognosis

Figure A.1: Plot of pROC under setting C1 with $\rho = 0.3$ and Error 3.

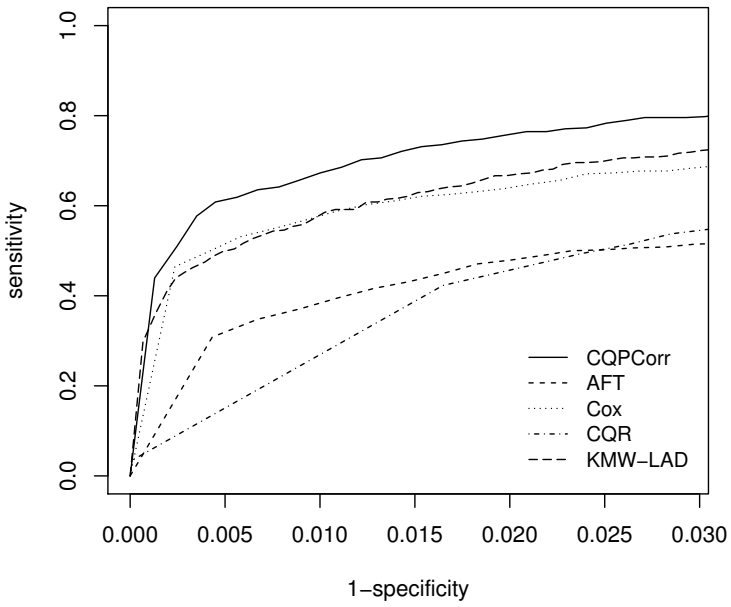


Table A.1: Simulation results for setting C1 with the AR correlation structure ($\rho = 0.3$), Error 1 and various values of sample size. In each cell, mean (sd) based on 200 replicates.

n	Method	TP20	TP40	pAUC	TP.FDR	FP.FDR	E.FDR
250	AFT	10.8(1.5)	11.8(1.4)	0.83(0.06)	12.6(1.7)	83.0(79.4)	0.78(0.15)
	Cox	11.1(1.7)	12.3(1.7)	0.88(0.05)	11.0(1.8)	12.5(18.5)	0.41(0.23)
	CQR	3.8(1.5)	5.6(1.6)	0.69(0.04)	9.3(1.8)	125.1(34.1)	0.93(0.02)
	KMW-LAD	8.4(1.7)	10.4(2.1)	0.86(0.05)	3.9(2.0)	1.3(1.5)	0.21(0.19)
	CQPCorr	10.8(1.6)	12.4(1.7)	0.91(0.04)	8.1(2.4)	1.9(1.8)	0.17(0.14)
300	AFT	11.2(1.4)	12.5(1.5)	0.85(0.05)	13.1(1.6)	67.1(48.0)	0.79(0.11)
	Cox	11.5(1.6)	12.8(1.5)	0.89(0.04)	12.0(1.9)	14.9(14.8)	0.46(0.21)
	CQR	4.6(1.5)	6.4(1.7)	0.72(0.04)	10.7(1.9)	123.8(29.7)	0.92(0.02)
	KMW-LAD	9.8(1.7)	11.5(1.5)	0.88(0.04)	6.4(3.0)	1.6(1.9)	0.19(0.19)
	CQPCorr	11.9(1.6)	13.2(1.6)	0.93(0.04)	10.0(2.1)	2.3(2.4)	0.16(0.12)
350	AFT	11.7(1.6)	12.9(1.7)	0.86(0.05)	13.4(1.8)	57.8(37.4)	0.76(0.12)
	Cox	11.8(1.4)	13.3(1.6)	0.91(0.04)	12.3(1.9)	13.4(9.9)	0.46(0.16)
	CQR	5.8(2.0)	7.7(2.2)	0.75(0.04)	11.6(1.4)	130.5(39.0)	0.91(0.02)
	KMW-LAD	11.2(1.4)	12.7(1.4)	0.92(0.04)	8.9(2.1)	2.3(2.6)	0.18(0.15)
	CQPCorr	12.9(1.6)	14.1(1.5)	0.95(0.03)	11.5(2.0)	2.3(1.5)	0.16(0.09)
400	AFT	12.4(1.6)	13.7(1.6)	0.88(0.04)	14.2(1.5)	61.4(33.4)	0.78(0.09)
	Cox	12.8(1.5)	14.2(1.5)	0.93(0.04)	13.3(1.7)	14.0(10.7)	0.46(0.15)
	CQR	6.6(2.0)	8.8(2.2)	0.77(0.05)	12.6(1.9)	130.4(39.9)	0.91(0.03)
	KMW-LAD	12.1(1.2)	13.9(1.3)	0.95(0.03)	10.0(2.0)	2.4(2.0)	0.17(0.12)
	CQPCorr	13.8(1.3)	14.9(1.0)	0.97(0.02)	12.8(1.7)	2.4(2.0)	0.14(0.11)

Table A.2: Simulation results for setting C1 with the AR correlation structure ($\rho = 0.3$), Error 2 and various values of sample size. In each cell, mean (sd) based on 200 replicates.

n	Method	TP20	TP40	pAUC	TP.FDR	FP.FDR	E.FDR
250	AFT	6.2(1.9)	7.5(2.0)	0.75(0.06)	4.6(2.6)	6.7(9.8)	0.42(0.25)
	Cox	8.1(1.9)	9.4(2.2)	0.82(0.06)	6.0(2.5)	3.7(4.6)	0.30(0.22)
	CQR	3.2(1.6)	4.7(1.9)	0.66(0.05)	8.2(2.6)	124.2(36.6)	0.94(0.02)
	KMW-LAD	7.3(1.7)	9.2(1.8)	0.82(0.05)	2.4(1.9)	0.8(1.1)	0.20(0.26)
	CQPCorr	9.2(2.0)	10.9(2.0)	0.87(0.05)	5.6(2.5)	1.5(1.5)	0.19(0.16)
300	AFT	7.8(1.8)	9.1(1.7)	0.80(0.05)	6.2(2.5)	6.1(8.8)	0.34(0.23)
	Cox	9.3(1.9)	10.6(2.0)	0.85(0.06)	8.0(2.3)	3.8(5.4)	0.24(0.19)
	CQR	4.3(1.5)	6.1(1.8)	0.71(0.05)	9.3(2.2)	112.5(36.1)	0.92(0.03)
	KMW-LAD	8.8(1.8)	10.5(2.2)	0.86(0.05)	4.5(2.8)	0.9(1.2)	0.11(0.13)
	CQPCorr	10.7(1.7)	12.0(1.8)	0.90(0.05)	7.6(2.5)	1.2(1.3)	0.11(0.11)
350	AFT	8.1(1.5)	9.2(1.5)	0.80(0.04)	6.4(2.5)	6.6(7.9)	0.38(0.24)
	Cox	9.8(1.7)	11.1(1.6)	0.87(0.05)	8.3(1.9)	3.2(3.7)	0.23(0.18)
	CQR	4.8(1.3)	6.7(1.4)	0.71(0.04)	10.0(1.8)	119.9(40.3)	0.92(0.03)
	KMW-LAD	9.6(1.5)	11.6(1.6)	0.88(0.04)	6.0(2.2)	1.8(2.2)	0.18(0.15)
	CQPCorr	11.2(1.4)	12.7(1.7)	0.92(0.04)	8.9(2.2)	1.7(1.7)	0.14(0.11)
400	AFT	8.8(1.6)	10.0(1.7)	0.83(0.05)	7.3(2.4)	6.6(9.4)	0.33(0.22)
	Cox	10.9(1.8)	12.3(1.8)	0.89(0.05)	9.6(2.2)	3.5(3.7)	0.23(0.14)
	CQR	5.6(2.0)	7.6(2.2)	0.75(0.05)	11.7(2.1)	122.3(39.0)	0.91(0.02)
	KMW-LAD	10.7(1.5)	12.4(1.6)	0.91(0.04)	7.1(2.3)	1.7(1.9)	0.16(0.17)
	CQPCorr	12.4(1.6)	13.7(1.5)	0.94(0.04)	10.4(2.4)	2.1(1.4)	0.16(0.09)

Table A.3: Simulation results for setting C1 with the AR correlation structure ($\rho = 0.3$), Error 3 and various values of sample size. In each cell, mean (sd) based on 200 replicates.

n	Method	TP20	TP40	pAUC	TP.FDR	FP.FDR	E.FDR
250	AFT	4.3(2.2)	5.5(2.3)	0.69(0.07)	2.0(2.3)	5.1(9.8)	0.43(0.39)
	Cox	7.1(1.9)	8.5(2.1)	0.79(0.06)	4.7(2.2)	2.7(5.0)	0.24(0.23)
	CQR	3.0(1.4)	4.5(1.8)	0.65(0.05)	7.1(2.6)	108.4(37.8)	0.93(0.03)
	KMW-LAD	6.5(1.8)	8.5(2.0)	0.80(0.06)	2.0(1.6)	0.4(0.6)	0.17(0.28)
	CQPCorr	8.3(1.8)	9.9(2.1)	0.84(0.06)	4.3(2.0)	1.1(1.3)	0.16(0.15)
300	AFT	4.5(2.3)	5.7(2.3)	0.70(0.06)	2.5(2.1)	4.9(9.1)	0.39(0.36)
	Cox	7.4(2.2)	8.7(1.8)	0.79(0.06)	5.0(2.6)	2.5(4.2)	0.24(0.23)
	CQR	3.7(1.5)	5.2(1.5)	0.69(0.04)	8.3(2.1)	104.0(29.9)	0.92(0.02)
	KMW-LAD	8.0(1.8)	9.6(2.0)	0.83(0.05)	3.0(2.5)	0.7(1.1)	0.18(0.25)
	CQPCorr	9.2(1.8)	10.8(1.9)	0.87(0.05)	6.0(2.2)	1.3(1.9)	0.14(0.16)
350	AFT	5.4(2.0)	6.7(2.2)	0.73(0.06)	3.0(2.7)	4.1(6.8)	0.35(0.33)
	Cox	8.7(1.5)	9.9(1.5)	0.83(0.05)	6.1(1.9)	1.7(2.1)	0.17(0.16)
	CQR	4.5(1.2)	5.9(1.1)	0.70(0.04)	9.2(2.0)	106.8(27.8)	0.92(0.02)
	KMW-LAD	9.9(1.8)	11.3(1.9)	0.88(0.05)	4.4(2.4)	0.6(1.0)	0.10(0.20)
	CQPCorr	10.9(1.5)	12.4(1.3)	0.91(0.04)	8.0(2.0)	1.0(0.8)	0.10(0.07)
400	AFT	6.4(1.9)	7.9(2.0)	0.76(0.06)	3.8(2.7)	3.8(6.9)	0.36(0.31)
	Cox	9.6(1.5)	10.9(1.6)	0.86(0.05)	7.4(2.3)	2.1(1.9)	0.19(0.14)
	CQR	5.1(2.0)	7.1(2.1)	0.74(0.04)	10.7(1.9)	111.6(33.2)	0.91(0.02)
	KMW-LAD	10.2(1.8)	11.9(1.7)	0.90(0.05)	5.8(3.2)	0.9(1.1)	0.10(0.11)
	CQPCorr	11.7(1.5)	12.9(1.7)	0.92(0.04)	9.2(2.0)	1.2(1.2)	0.10(0.09)

Table A.4: Simulation results for setting C3 with the AR correlation structure ($\rho = 0.5$). In each cell, mean (sd) based on 200 replicates.

Error	Method	TP20	TP40	pAUC	TP.FDR	FP.FDR	E.FDR
1	AFT	10.8(1.5)	12.0(1.8)	0.86(0.05)	12.0(2.0)	52.5(76.5)	0.64(0.22)
	Cox	9.3(1.8)	11.1(1.7)	0.86(0.05)	7.2(3.3)	3.3(3.1)	0.25(0.18)
	CQR	4.2(1.8)	6.2(1.9)	0.71(0.06)	10.8(2.0)	141.9(45.9)	0.92(0.02)
	KMW-LAD	9.1(1.7)	10.8(1.9)	0.86(0.05)	6.0(3.1)	1.0(1.7)	0.13(0.19)
	CQPCorr	9.9(1.8)	11.4(2.0)	0.89(0.04)	6.8(2.8)	1.1(1.0)	0.12(0.09)
2	AFT	5.6(2.5)	7.0(2.6)	0.74(0.07)	3.8(2.7)	8.1(19.9)	0.45(0.29)
	Cox	4.7(2.3)	6.1(2.5)	0.71(0.08)	2.5(2.7)	2.0(3.7)	0.18(0.26)
	CQR	3.6(1.8)	5.4(1.8)	0.69(0.07)	8.5(2.8)	107.9(25.7)	0.92(0.03)
	KMW-LAD	7.5(2.4)	9.0(2.3)	0.81(0.07)	3.2(2.2)	0.4(0.7)	0.10(0.22)
	CQPCorr	7.6(2.2)	9.2(2.2)	0.82(0.06)	4.0(2.5)	0.7(0.8)	0.11(0.11)
3	AFT	3.9(2.0)	5.2(2.1)	0.69(0.07)	1.2(2.4)	4.1(12.7)	0.31(0.41)
	Cox	3.5(2.0)	4.5(2.3)	0.67(0.08)	0.8(1.3)	1.0(2.6)	0.16(0.28)
	CQR	3.0(1.5)	4.5(1.8)	0.66(0.04)	7.5(2.0)	112.0(33.6)	0.93(0.02)
	KMW-LAD	6.8(1.9)	8.5(1.8)	0.80(0.05)	2.2(1.8)	0.7(1.5)	0.11(0.23)
	CQPCorr	7.1(1.8)	8.8(1.8)	0.80(0.06)	3.5(2.4)	0.7(0.9)	0.12(0.14)

Table A.5: Simulation results for setting C4 with the AR correlation structure ($\rho = 0.5$). In each cell, mean (sd) based on 200 replicates.

Error	Method	TP20	TP40	pAUC	TP.FDR	FP.FDR	E.FDR
1	AFT	12.1(2.0)	13.8(1.6)	0.90(0.07)	13.6(2.3)	41.6(59.2)	0.61(0.19)
	Cox	10.0(2.3)	11.6(2.2)	0.88(0.05)	7.4(4.0)	3.4(3.9)	0.25(0.18)
	CQR	4.3(1.7)	6.1(2.0)	0.71(0.06)	9.8(3.0)	130.8(38.4)	0.93(0.02)
	KMW-LAD	9.2(1.6)	11.8(1.2)	0.90(0.03)	5.9(2.6)	2.3(2.8)	0.21(0.19)
	CQPCorr	10.1(2.3)	12.0(1.8)	0.91(0.05)	7.5(3.7)	2.9(2.9)	0.20(0.15)
2	AFT	1.2(1.6)	1.6(1.7)	0.56(0.06)	0.6(1.1)	1.6(2.4)	0.33(0.43)
	Cox	4.2(2.1)	5.7(3.0)	0.70(0.10)	1.6(1.8)	1.2(2.7)	0.21(0.33)
	CQR	3.4(1.5)	4.8(2.3)	0.67(0.06)	7.9(2.1)	123.2(30.0)	0.94(0.02)
	KMW-LAD	7.2(3.3)	9.8(2.8)	0.83(0.08)	2.8(1.8)	0.8(1.3)	0.15(0.19)
	CQPCorr	7.6(1.9)	10.4(2.0)	0.85(0.06)	3.8(2.7)	1.1(0.9)	0.14(0.13)
3	AFT	0.4(1.4)	0.6(1.5)	0.52(0.05)	0.2(1.0)	4.5(14.9)	0.28(0.45)
	Cox	2.0(2.1)	2.9(2.7)	0.60(0.09)	0.4(0.7)	0.5(0.8)	0.24(0.38)
	CQR	2.6(2.0)	4.4(2.1)	0.65(0.07)	7.2(2.7)	106.5(41.3)	0.93(0.03)
	KMW-LAD	6.0(2.3)	7.8(2.4)	0.79(0.06)	1.1(1.1)	0.4(0.7)	0.11(0.21)
	CQPCorr	6.2(2.8)	8.2(2.8)	0.83(0.08)	2.0(1.9)	0.5(0.8)	0.10(0.15)

Table A.6: Simulation results for setting C5 with the AR correlation structure ($\rho = 0.5$). In each cell, mean (sd) based on 200 replicates.

Error	Method	TP20	TP40	pAUC	TP.FDR	FP.FDR	E.FDR
1	AFT	10.0(0.2)	10.0(0.2)	0.77(0.03)	10.0(0.0)	62.2(66.0)	0.76(0.15)
	Cox	9.9(0.3)	10.0(0.4)	0.79(0.02)	9.9(0.4)	18.4(16.6)	0.56(0.20)
	CQR	5.0(1.7)	6.4(1.9)	0.69(0.04)	8.5(1.7)	112.4(35.3)	0.93(0.02)
	KMW-LAD	8.8(0.9)	9.6(0.8)	0.80(0.02)	6.8(2.0)	1.2(1.7)	0.12(0.12)
	CQPCorr	9.8(0.5)	10.1(0.7)	0.81(0.02)	8.9(1.0)	1.4(1.0)	0.12(0.08)
2	AFT	7.9(1.6)	8.4(1.7)	0.76(0.04)	7.3(2.1)	9.3(10.2)	0.44(0.24)
	Cox	8.6(1.4)	8.9(1.5)	0.77(0.04)	7.5(1.6)	4.3(7.2)	0.26(0.20)
	CQR	3.6(1.7)	5.0(1.5)	0.67(0.03)	7.7(1.8)	112.1(35.7)	0.93(0.02)
	KMW-LAD	7.8(1.6)	8.5(1.5)	0.77(0.04)	4.4(2.2)	1.0(1.8)	0.11(0.18)
	CQPCorr	8.6(1.1)	9.1(1.0)	0.80(0.02)	7.2(1.6)	1.0(1.0)	0.11(0.10)
3	AFT	6.0(2.2)	7.2(2.3)	0.74(0.06)	4.3(3.2)	8.0(18.3)	0.28(0.30)
	Cox	7.6(1.9)	8.3(1.8)	0.77(0.05)	5.8(2.5)	2.4(3.0)	0.23(0.20)
	CQR	3.4(1.4)	4.9(1.5)	0.67(0.05)	7.0(1.6)	102.6(34.2)	0.93(0.02)
	KMW-LAD	6.9(1.4)	8.3(1.5)	0.77(0.05)	3.1(2.2)	0.6(1.0)	0.12(0.16)
	CQPCorr	8.2(1.6)	9.2(1.2)	0.80(0.04)	5.4(2.1)	0.9(1.0)	0.11(0.10)

Settings with banded correlation structure and binary E factors

Under coefficient settings C1 and C2, the following additional scenarios are examined. For G factors, besides the AR correlation structure, we also consider the banded correlation structure. Here two scenarios are considered. Under the first scenario (Band 1), the jth and lth G variables have correlation coefficient 0.5 if $|j-l| = 1$ and 0 otherwise. Under the second scenario (Band 2), the jth and lth G variables have correlation coefficient 0.7 if $|j-l| = 1$, 0.4 if $|j-l| = 2$, 0.1 if $|j-l| = 3$, and 0 otherwise. For E factors, the other scenario (E2) dichotomizes two of the continuous E factors at 0 and create two binary variables. Under coefficient settings C1 and C2 with the AR correlation structure ($\rho = 0.5$), we examine another datasets with a higher censoring rate (35%). Summary results are provided in Tables A.7-A.14.

Table A.7: Simulation results for setting C1 with the AR correlation structure and E2. In each cell, mean (sd) based on 200 replicates.

	Error	Method	TP20	TP40	pAUC	TP.FDR	FP.FDR	E.FDR
$\rho = 0.3$	1	AFT	8.5(1.0)	9.4(1.1)	0.74(0.04)	9.5(1.2)	73.0(58.7)	0.83(0.11)
		Cox	8.6(1.4)	9.4(1.3)	0.79(0.04)	7.9(1.5)	7.0(5.1)	0.41(0.19)
		CQR	3.1(1.5)	4.1(1.5)	0.65(0.04)	6.0(1.7)	81.4(20.7)	0.93(0.02)
		KMW-LAD	5.9(1.7)	7.1(1.7)	0.75(0.04)	3.1(1.6)	0.6(0.9)	0.15(0.23)
		CQPCorr	7.9(1.3)	8.8(1.5)	0.80(0.04)	5.2(2.0)	1.1(0.9)	0.14(0.11)
	2	AFT	5.2(1.9)	5.8(2.0)	0.70(0.05)	3.6(2.5)	6.0(9.0)	0.43(0.32)
		Cox	6.5(1.4)	7.6(1.5)	0.75(0.03)	4.2(2.4)	1.5(1.8)	0.20(0.18)
		CQR	3.0(1.5)	4.5(1.4)	0.65(0.04)	6.0(1.7)	87.5(26.9)	0.93(0.03)
		KMW-LAD	5.8(1.6)	7.1(1.7)	0.74(0.05)	2.5(1.9)	0.4(0.7)	0.09(0.18)
		CQPCorr	6.8(1.8)	8.1(2.0)	0.77(0.05)	2.8(1.9)	0.4(0.7)	0.09(0.13)
	3	AFT	1.6(1.8)	2.0(2.1)	0.60(0.07)	1.0(1.2)	2.3(4.3)	0.48(0.48)
		Cox	4.6(1.8)	6.1(1.2)	0.70(0.04)	1.9(2.0)	1.9(3.0)	0.30(0.41)
		CQR	2.3(1.8)	3.0(1.6)	0.62(0.03)	4.6(1.3)	86.4(32.0)	0.94(0.03)
		KMW-LAD	4.4(1.7)	6.0(1.5)	0.70(0.03)	1.3(1.7)	0.4(0.5)	0.21(0.37)
		CQPCorr	5.7(2.4)	7.0(1.9)	0.74(0.05)	1.7(1.5)	0.7(1.3)	0.20(0.39)
$\rho = 0.5$	1	AFT	10.0(0.6)	10.3(0.8)	0.78(0.04)	10.4(0.8)	47.2(35.1)	0.75(0.15)
		Cox	9.8(0.7)	10.4(0.8)	0.82(0.03)	9.8(0.7)	13.7(16.8)	0.46(0.24)
		CQR	5.1(1.8)	6.3(1.9)	0.72(0.05)	8.6(1.7)	81.7(24.6)	0.90(0.03)
		KMW-LAD	9.4(1.2)	10.3(1.4)	0.83(0.04)	7.1(2.0)	1.3(1.5)	0.12(0.13)
		CQPCorr	9.8(0.9)	10.6(1.0)	0.85(0.03)	8.9(1.0)	1.3(1.4)	0.11(0.11)
	2	AFT	8.4(1.2)	9.2(1.1)	0.78(0.03)	7.5(2.4)	13.8(26.8)	0.48(0.20)
		Cox	8.6(1.4)	9.4(1.0)	0.79(0.02)	8.0(2.0)	5.6(4.4)	0.34(0.18)
		CQR	4.7(1.5)	6.1(1.6)	0.71(0.06)	8.1(2.5)	83.2(30.3)	0.90(0.03)
		KMW-LAD	8.0(1.8)	9.3(1.8)	0.81(0.05)	5.1(2.5)	0.9(1.1)	0.15(0.22)
		CQPCorr	9.3(1.6)	10.3(1.7)	0.84(0.04)	7.5(2.1)	1.5(1.5)	0.15(0.12)
	3	AFT	7.2(1.9)	8.4(1.5)	0.76(0.04)	6.4(3.1)	9.6(14.9)	0.39(0.26)
		Cox	8.9(1.1)	9.4(0.9)	0.79(0.03)	7.5(2.2)	3.8(4.9)	0.25(0.20)
		CQR	4.6(2.1)	6.2(1.8)	0.70(0.05)	7.2(1.7)	69.5(24.7)	0.89(0.05)
		KMW-LAD	7.6(1.5)	8.6(1.6)	0.79(0.05)	3.9(2.3)	0.5(0.7)	0.10(0.15)
		CQPCorr	8.6(1.5)	9.7(1.3)	0.82(0.04)	7.0(1.9)	0.9(0.8)	0.10(0.09)

Table A.8: Simulation results for setting C1 with the banded correlation structure and E1. In each cell, mean (sd) based on 200 replicates.

	Error	Method	TP20	TP40	pAUC	TP.FDR	FP.FDR	E.FDR
Band1	1	AFT	10.8(1.7)	11.8(2.1)	0.81(0.06)	12.8(1.8)	103.9(96.1)	0.83(0.11)
		Cox	11.1(1.8)	12.2(1.8)	0.88(0.05)	11.6(1.8)	14.4(11.8)	0.49(0.16)
		CQR	4.0(1.6)	6.0(1.7)	0.70(0.05)	9.8(2.0)	130.0(39.0)	0.93(0.02)
		KMW-LAD	9.2(1.6)	10.9(1.9)	0.86(0.05)	5.9(2.7)	1.7(1.9)	0.20(0.17)
		CQPCorr	11.0(1.5)	12.3(1.5)	0.91(0.04)	8.6(2.0)	2.0(1.7)	0.16(0.12)
	2	AFT	6.8(2.2)	8.2(1.9)	0.76(0.06)	5.9(2.6)	16.4(21.1)	0.52(0.28)
		Cox	8.2(2.1)	9.5(1.8)	0.81(0.05)	6.6(2.7)	4.9(4.7)	0.35(0.20)
		CQR	3.3(1.9)	4.5(1.9)	0.67(0.05)	8.2(2.0)	125.6(50.2)	0.93(0.02)
		KMW-LAD	7.5(1.8)	9.0(2.1)	0.80(0.05)	3.8(2.8)	1.4(1.4)	0.22(0.19)
		CQPCorr	9.0(1.8)	10.4(2.1)	0.85(0.05)	5.6(2.3)	1.1(1.2)	0.15(0.15)
	3	AFT	4.9(2.6)	6.3(2.4)	0.71(0.07)	3.3(2.7)	11.0(15.9)	0.46(0.39)
		Cox	7.4(1.8)	8.6(2.0)	0.79(0.06)	5.0(2.3)	3.8(3.8)	0.34(0.25)
		CQR	4.0(1.2)	5.3(1.6)	0.67(0.05)	7.7(2.5)	112.3(43.6)	0.93(0.03)
		KMW-LAD	7.2(2.1)	9.2(2.2)	0.82(0.06)	3.7(2.2)	0.5(0.7)	0.11(0.14)
		CQPCorr	8.9(2.0)	10.4(2.2)	0.85(0.06)	5.4(2.5)	0.7(0.9)	0.10(0.11)
Band2	1	AFT	12.0(1.0)	14.0(1.7)	0.85(0.08)	15.3(1.2)	109.7(67.6)	0.85(0.08)
		Cox	12.3(0.6)	15.0(0.0)	0.95(0.03)	14.7(0.6)	37.0(32.1)	0.65(0.16)
		CQR	7.3(3.1)	10.3(2.5)	0.84(0.04)	15.3(1.2)	136.0(9.5)	0.90(0.01)
		KMW-LAD	12.7(0.6)	15.0(1.0)	0.97(0.01)	12.0(1.0)	4.0(1.7)	0.25(0.08)
		CQPCorr	14.0(1.0)	15.3(1.2)	0.97(0.02)	13.3(0.6)	4.3(1.2)	0.24(0.05)
	2	AFT	10.1(1.4)	11.4(1.8)	0.85(0.06)	11.9(2.0)	53.5(55.0)	0.73(0.15)
		Cox	11.2(2.0)	12.7(2.0)	0.90(0.05)	11.9(2.3)	16.0(10.3)	0.52(0.16)
		CQR	6.2(1.8)	8.5(2.1)	0.77(0.06)	11.7(2.1)	126.3(53.0)	0.90(0.05)
		KMW-LAD	11.1(1.6)	12.8(2.0)	0.92(0.05)	9.3(2.7)	3.9(3.3)	0.25(0.15)
		CQPCorr	12.5(1.9)	14.0(1.7)	0.94(0.05)	11.5(2.5)	4.0(2.2)	0.24(0.10)
	3	AFT	9.1(1.2)	10.5(1.8)	0.82(0.07)	10.0(1.9)	38.5(61.8)	0.63(0.21)
		Cox	11.2(1.8)	12.4(2.1)	0.89(0.07)	11.1(1.7)	11.5(9.3)	0.45(0.18)
		CQR	6.0(1.8)	8.1(1.9)	0.77(0.04)	11.8(1.4)	120.1(36.9)	0.90(0.03)
		KMW-LAD	10.4(1.5)	12.2(2.1)	0.90(0.06)	8.1(2.2)	3.0(2.2)	0.25(0.13)
		CQPCorr	11.8(1.7)	13.4(1.8)	0.93(0.05)	10.4(2.2)	3.4(1.8)	0.24(0.09)

Table A.9: Simulation results for setting C1 with the banded correlation structure and E2. In each cell, mean (sd) based on 200 replicates.

	Error	Method	TP20	TP40	pAUC	TP.FDR	FP.FDR	E.FDR
Band1	1	AFT	9.5(0.5)	9.8(0.4)	0.75(0.03)	9.9(0.5)	96.3(84.4)	0.85(0.10)
		Cox	9.4(0.7)	9.9(0.7)	0.79(0.02)	9.4(0.9)	14.8(12.2)	0.54(0.16)
		CQR	4.2(1.7)	6.1(1.8)	0.70(0.05)	7.5(1.9)	69.9(18.8)	0.90(0.03)
		KMW-LAD	8.1(1.2)	9.1(1.1)	0.80(0.04)	4.8(1.7)	1.2(1.6)	0.15(0.16)
		CQPCorr	9.1(1.2)	10.2(1.0)	0.83(0.03)	7.6(1.7)	1.4(1.0)	0.14(0.08)
	2	AFT	6.5(1.8)	7.8(1.8)	0.74(0.05)	5.6(2.8)	10.5(12.1)	0.50(0.28)
		Cox	7.4(1.4)	8.4(1.5)	0.76(0.03)	6.3(1.8)	4.0(4.4)	0.30(0.24)
		CQR	3.5(1.2)	4.8(1.3)	0.66(0.04)	6.2(2.3)	79.8(30.2)	0.92(0.02)
		KMW-LAD	7.1(1.7)	8.2(1.6)	0.78(0.04)	3.2(1.8)	0.9(0.9)	0.20(0.23)
		CQPCorr	8.1(1.5)	9.2(1.4)	0.80(0.04)	5.6(2.3)	1.4(1.6)	0.17(0.17)
	3	AFT	4.4(2.2)	5.1(2.5)	0.68(0.07)	2.2(2.4)	4.9(9.3)	0.42(0.40)
		Cox	6.8(2.0)	7.4(2.1)	0.74(0.05)	4.0(2.9)	1.5(2.9)	0.16(0.22)
		CQR	3.4(1.4)	4.3(1.1)	0.65(0.04)	5.8(2.1)	79.8(22.5)	0.93(0.02)
		KMW-LAD	5.5(1.4)	6.7(1.3)	0.73(0.05)	2.1(2.0)	0.4(1.3)	0.04(0.12)
		CQPCorr	7.4(1.1)	8.3(1.2)	0.78(0.03)	4.1(2.0)	0.3(0.5)	0.04(0.07)
Band2	1	AFT	10.1(0.5)	10.4(1.0)	0.76(0.07)	10.9(1.0)	126.6(113.8)	0.86(0.12)
		Cox	10.2(0.6)	10.5(0.8)	0.82(0.04)	10.4(0.7)	24.9(16.6)	0.64(0.17)
		CQR	6.6(1.7)	8.4(1.7)	0.77(0.05)	10.5(1.6)	99.4(25.3)	0.90(0.02)
		KMW-LAD	10.1(1.6)	10.9(1.5)	0.85(0.05)	8.5(1.2)	2.4(1.8)	0.21(0.13)
		CQPCorr	10.8(1.5)	11.3(1.5)	0.86(0.04)	10.1(1.3)	2.8(1.7)	0.20(0.10)
	2	AFT	9.5(0.8)	9.8(0.7)	0.79(0.03)	9.6(0.8)	23.2(21.1)	0.60(0.23)
		Cox	9.8(0.6)	10.5(1.1)	0.82(0.04)	9.4(0.7)	10.7(13.7)	0.42(0.22)
		CQR	6.4(1.1)	8.2(1.5)	0.77(0.04)	10.2(1.8)	79.1(26.2)	0.88(0.05)
		KMW-LAD	9.5(1.2)	10.0(1.1)	0.84(0.04)	7.7(2.0)	1.5(1.3)	0.15(0.12)
		CQPCorr	10.2(1.0)	10.8(0.9)	0.86(0.04)	9.1(0.9)	1.6(1.3)	0.14(0.10)
	3	AFT	8.5(1.1)	9.1(0.9)	0.76(0.02)	8.2(1.5)	26.9(37.1)	0.52(0.30)
		Cox	9.4(0.8)	9.8(0.4)	0.80(0.02)	9.3(0.9)	9.1(14.4)	0.30(0.29)
		CQR	5.9(1.6)	7.4(1.6)	0.75(0.03)	9.4(1.3)	69.9(15.7)	0.88(0.03)
		KMW-LAD	9.6(1.3)	10.6(1.2)	0.84(0.05)	7.0(1.6)	0.8(1.1)	0.08(0.10)
		CQPCorr	9.8(0.8)	10.8(1.2)	0.84(0.03)	8.4(1.1)	0.6(0.5)	0.07(0.06)

Table A.10: Simulation results for setting C2 with the AR correlation structure and E2. In each cell, mean (sd) based on 200 replicates.

	Error	Method	TP20	TP40	pAUC	TP.FDR	FP.FDR	E.FDR
$\rho = 0.3$	1	AFT	8.5(1.4)	9.5(1.4)	0.78(0.04)	8.6(2.2)	31.9(37.2)	0.61(0.24)
		Cox	7.6(1.6)	8.5(1.6)	0.77(0.05)	5.7(2.6)	3.2(3.5)	0.25(0.20)
		CQR	3.0(1.5)	4.1(1.6)	0.65(0.04)	5.9(1.9)	82.2(33.0)	0.93(0.02)
		KMW-LAD	5.7(1.8)	7.0(1.8)	0.74(0.05)	2.5(2.0)	0.9(1.3)	0.20(0.25)
		CQPCorr	6.8(1.5)	8.0(1.6)	0.77(0.04)	3.8(2.2)	1.2(1.4)	0.19(0.18)
	2	AFT	2.6(1.2)	3.6(1.3)	0.62(0.04)	0.5(0.7)	1.7(3.0)	0.42(0.43)
		Cox	4.0(1.7)	5.0(2.0)	0.68(0.06)	1.6(2.0)	0.8(1.4)	0.15(0.24)
		CQR	2.1(1.0)	3.3(1.6)	0.62(0.05)	4.4(2.3)	72.9(20.7)	0.94(0.03)
		KMW-LAD	4.4(1.4)	5.7(1.4)	0.71(0.04)	1.0(1.1)	0.1(0.3)	0.03(0.08)
		CQPCorr	5.1(1.5)	6.5(1.3)	0.73(0.04)	1.6(1.3)	0.1(0.4)	0.04(0.09)
	3	AFT	1.4(1.5)	1.9(1.5)	0.57(0.05)	0.2(0.5)	1.1(2.0)	0.37(0.47)
		Cox	3.1(2.1)	3.7(2.6)	0.63(0.08)	0.8(1.4)	0.3(0.7)	0.12(0.28)
		CQR	2.1(1.3)	3.3(1.7)	0.62(0.05)	4.4(1.7)	62.7(27.1)	0.93(0.03)
		KMW-LAD	3.9(1.4)	5.0(1.9)	0.67(0.06)	0.3(0.6)	0.1(0.2)	0.03(0.11)
		CQPCorr	3.8(1.8)	5.2(2.1)	0.69(0.07)	1.3(1.4)	0.2(0.4)	0.07(0.15)
$\rho = 0.5$	1	AFT	9.7(0.8)	10.4(1.0)	0.78(0.06)	10.3(1.3)	55.8(65.9)	0.69(0.25)
		Cox	9.4(0.9)	10.0(1.1)	0.81(0.04)	8.8(1.3)	7.9(13.9)	0.28(0.26)
		CQR	5.1(1.8)	6.5(2.1)	0.72(0.06)	8.8(1.7)	91.3(24.0)	0.91(0.03)
		KMW-LAD	8.8(1.7)	10.0(1.9)	0.84(0.06)	6.4(1.8)	1.0(1.3)	0.11(0.13)
		CQPCorr	9.7(2.1)	10.4(2.1)	0.84(0.06)	7.3(2.7)	1.1(1.1)	0.10(0.09)
	2	AFT	6.5(2.0)	7.9(1.8)	0.75(0.05)	4.1(3.2)	3.9(6.8)	0.29(0.28)
		Cox	7.9(2.1)	8.6(2.0)	0.78(0.06)	5.7(3.3)	2.7(3.5)	0.22(0.22)
		CQR	4.5(2.3)	6.5(2.2)	0.73(0.06)	8.0(2.1)	69.2(29.3)	0.88(0.05)
		KMW-LAD	8.7(1.3)	9.9(1.9)	0.82(0.06)	3.9(2.0)	0.3(0.5)	0.08(0.12)
		CQPCorr	9.4(1.7)	10.5(1.9)	0.85(0.05)	6.8(2.4)	0.7(0.8)	0.07(0.08)
	3	AFT	5.5(3.0)	6.2(3.2)	0.71(0.09)	3.4(3.5)	3.5(7.1)	0.30(0.37)
		Cox	6.3(3.0)	7.0(2.9)	0.73(0.09)	4.7(3.8)	2.0(2.8)	0.25(0.28)
		CQR	4.3(1.6)	6.0(1.8)	0.71(0.05)	7.8(2.4)	67.0(20.3)	0.89(0.03)
		KMW-LAD	8.6(1.8)	9.5(1.5)	0.82(0.05)	3.1(2.1)	0.2(0.4)	0.05(0.11)
		CQPCorr	8.8(1.8)	9.6(1.7)	0.83(0.05)	5.5(2.3)	0.3(0.5)	0.04(0.06)

Table A.11: Simulation results for setting C2 with the banded correlation structure and E1. In each cell, mean (sd) based on 200 replicates.

	Error	Method	TP20	TP40	pAUC	TP.FDR	FP.FDR	E.FDR
Band1	1	AFT	10.9(1.5)	12.2(1.9)	0.85(0.06)	12.8(2.1)	62.8(58.3)	0.74(0.17)
		Cox	10.8(2.1)	11.8(2.0)	0.88(0.06)	10.1(2.8)	8.9(9.0)	0.38(0.20)
		CQR	4.3(1.9)	5.9(2.0)	0.70(0.05)	10.4(2.4)	142.5(38.4)	0.93(0.02)
		KMW-LAD	8.7(2.1)	10.5(2.2)	0.86(0.05)	5.1(2.4)	1.3(1.3)	0.18(0.15)
		CQPCorr	10.2(2.2)	11.8(1.9)	0.89(0.05)	6.6(2.5)	1.7(1.7)	0.17(0.14)
	2	AFT	4.7(2.0)	6.0(2.0)	0.71(0.06)	3.4(3.0)	16.2(32.7)	0.55(0.35)
		Cox	7.3(2.1)	8.6(2.3)	0.79(0.08)	5.2(2.5)	4.2(5.3)	0.32(0.26)
		CQR	3.8(1.8)	5.0(1.9)	0.68(0.05)	8.6(2.3)	120.0(32.0)	0.93(0.02)
		KMW-LAD	7.7(1.4)	9.4(1.7)	0.83(0.05)	3.5(2.0)	0.8(1.3)	0.15(0.23)
		CQPCorr	8.5(1.7)	10.6(2.0)	0.86(0.06)	5.1(2.1)	1.0(1.1)	0.15(0.14)
	3	AFT	2.7(1.4)	3.8(2.0)	0.65(0.07)	1.1(1.2)	5.2(9.7)	0.50(0.44)
		Cox	5.5(2.1)	6.9(2.7)	0.73(0.09)	3.0(2.7)	1.7(2.4)	0.27(0.30)
		CQR	3.0(1.6)	4.3(1.5)	0.65(0.05)	7.5(2.0)	119.5(32.4)	0.94(0.03)
		KMW-LAD	6.9(1.8)	8.4(1.9)	0.80(0.05)	2.2(2.2)	0.6(1.4)	0.10(0.16)
		CQPCorr	7.5(1.9)	8.8(2.2)	0.81(0.07)	3.3(2.2)	0.5(0.7)	0.09(0.12)
Band2	1	AFT	12.7(1.2)	14.3(1.5)	0.87(0.07)	15.2(1.2)	106.3(97.0)	0.80(0.12)
		Cox	13.1(1.3)	14.7(1.2)	0.94(0.04)	14.5(1.7)	36.0(43.7)	0.60(0.17)
		CQR	6.8(2.0)	9.2(2.3)	0.80(0.04)	13.3(1.8)	151.4(55.7)	0.91(0.03)
		KMW-LAD	12.5(1.1)	14.1(1.2)	0.95(0.03)	11.5(1.6)	5.5(3.4)	0.29(0.15)
		CQPCorr	13.4(1.1)	14.9(1.3)	0.97(0.02)	13.1(1.6)	5.3(2.3)	0.28(0.09)
	2	AFT	10.4(1.6)	11.9(1.3)	0.87(0.04)	11.0(2.5)	26.1(31.7)	0.60(0.17)
		Cox	11.6(1.9)	13.1(2.2)	0.91(0.06)	11.6(2.4)	16.8(22.6)	0.43(0.24)
		CQR	6.7(1.4)	8.6(1.3)	0.76(0.05)	12.7(1.8)	142.9(40.0)	0.91(0.03)
		KMW-LAD	11.2(1.3)	13.2(1.3)	0.92(0.04)	8.7(2.5)	3.2(2.7)	0.22(0.15)
		CQPCorr	12.8(1.3)	14.2(1.2)	0.96(0.03)	11.1(2.2)	3.1(1.9)	0.20(0.10)
	3	AFT	8.2(2.7)	9.6(2.7)	0.82(0.09)	6.3(4.4)	11.9(17.9)	0.45(0.32)
		Cox	9.5(2.7)	11.4(2.8)	0.88(0.08)	7.8(3.7)	5.3(5.6)	0.38(0.25)
		CQR	5.4(1.9)	7.4(2.3)	0.74(0.06)	11.4(2.6)	132.8(35.5)	0.92(0.02)
		KMW-LAD	10.9(2.0)	12.9(2.4)	0.92(0.05)	8.3(2.9)	3.0(2.7)	0.23(0.14)
		CQPCorr	11.8(1.5)	13.3(1.7)	0.94(0.05)	9.9(2.9)	3.1(1.8)	0.22(0.09)

Table A.12: Simulation results for setting C2 with the banded correlation structure and E2. In each cell, mean (sd) based on 200 replicates.

	Error	Method	TP20	TP40	pAUC	TP.FDR	FP.FDR	E.FDR
Band1	1	AFT	9.0(1.2)	9.4(1.3)	0.76(0.04)	9.4(1.6)	45.8(52.2)	0.69(0.23)
		Cox	8.7(1.6)	9.4(1.4)	0.78(0.04)	8.2(2.0)	8.9(21.2)	0.30(0.26)
		CQR	4.5(1.7)	5.8(2.0)	0.69(0.06)	7.9(2.3)	83.6(29.5)	0.91(0.04)
		KMW-LAD	7.4(2.0)	8.6(2.1)	0.79(0.07)	4.5(2.3)	1.0(1.1)	0.15(0.15)
		CQPCorr	8.1(1.9)	9.2(2.2)	0.81(0.07)	5.8(2.0)	1.1(1.0)	0.14(0.12)
	2	AFT	4.7(1.8)	5.6(1.9)	0.69(0.05)	2.6(2.6)	13.7(33.9)	0.49(0.40)
		Cox	5.7(2.1)	6.6(2.2)	0.71(0.06)	3.7(2.6)	2.3(3.5)	0.27(0.22)
		CQR	3.3(1.9)	4.8(2.1)	0.67(0.06)	6.5(2.3)	77.3(22.1)	0.92(0.03)
		KMW-LAD	6.2(2.0)	7.7(2.3)	0.76(0.07)	2.9(1.8)	0.8(1.4)	0.15(0.24)
		CQPCorr	7.2(2.2)	7.9(2.4)	0.78(0.07)	4.2(2.4)	0.8(1.0)	0.14(0.16)
	3	AFT	2.9(2.0)	3.6(2.1)	0.63(0.07)	0.9(1.4)	2.6(5.9)	0.44(0.44)
		Cox	5.1(1.7)	5.9(1.7)	0.70(0.05)	2.4(2.3)	1.3(2.4)	0.14(0.24)
		CQR	2.8(1.4)	4.1(1.7)	0.64(0.04)	5.3(1.4)	75.1(28.9)	0.93(0.03)
		KMW-LAD	5.6(1.7)	7.1(2.1)	0.74(0.06)	1.6(1.4)	0.3(0.5)	0.09(0.18)
		CQPCorr	6.0(1.6)	7.1(1.8)	0.75(0.05)	2.9(1.6)	0.4(1.1)	0.08(0.15)
Band2	1	AFT	10.1(0.6)	10.3(0.5)	0.80(0.05)	10.2(0.4)	35.7(31.5)	0.64(0.29)
		Cox	10.0(0.9)	10.4(1.0)	0.81(0.04)	10.1(0.3)	31.7(53.5)	0.52(0.25)
		CQR	6.7(1.7)	8.2(1.8)	0.75(0.05)	10.0(1.8)	101.8(37.0)	0.90(0.03)
		KMW-LAD	9.7(1.2)	10.8(1.4)	0.86(0.05)	8.6(1.6)	2.5(1.9)	0.21(0.11)
		CQPCorr	9.9(0.6)	11.2(0.8)	0.88(0.04)	9.2(1.4)	2.4(1.0)	0.20(0.05)
	2	AFT	9.3(0.9)	9.9(0.6)	0.80(0.02)	8.6(1.7)	8.0(9.3)	0.33(0.27)
		Cox	9.4(1.3)	10.2(1.3)	0.82(0.04)	8.9(1.4)	4.9(4.9)	0.29(0.19)
		CQR	6.3(1.6)	7.6(1.5)	0.75(0.05)	9.4(1.7)	78.2(26.2)	0.88(0.04)
		KMW-LAD	9.9(1.4)	11.4(1.9)	0.87(0.05)	7.7(1.4)	1.2(2.1)	0.09(0.14)
		CQPCorr	10.2(1.2)	11.2(1.6)	0.87(0.04)	8.4(1.6)	1.8(1.1)	0.17(0.10)
	3	AFT	7.0(2.2)	8.2(2.3)	0.77(0.07)	5.0(3.5)	5.2(8.1)	0.29(0.26)
		Cox	8.9(1.4)	9.8(1.4)	0.80(0.05)	6.9(2.7)	4.0(3.9)	0.32(0.25)
		CQR	5.5(1.4)	7.4(1.4)	0.74(0.05)	9.2(1.9)	77.8(35.2)	0.88(0.04)
		KMW-LAD	9.0(1.5)	10.5(1.5)	0.85(0.05)	6.4(2.5)	1.5(2.2)	0.12(0.14)
		CQPCorr	9.8(1.3)	10.9(1.4)	0.86(0.04)	7.4(2.2)	1.2(1.2)	0.12(0.10)

Table A.13: Simulation results for setting C1 with the AR correlation structure ($\rho = 0.5$) and 35% censoring rate. In each cell, mean (sd) based on 200 replicates.

Error	Method	TP20	TP40	pAUC	TP.FDR	FP.FDR	E.FDR
1	AFT	10.0(0.2)	10.0(0.2)	0.77(0.03)	10.0(0.0)	62.2(66.0)	0.76(0.15)
	Cox	9.9(0.3)	10.0(0.4)	0.79(0.02)	9.9(0.4)	18.4(16.6)	0.56(0.20)
	CQR	5.0(1.7)	6.4(1.9)	0.69(0.04)	8.5(1.7)	112.4(35.3)	0.93(0.02)
	KMW-LAD	8.8(0.9)	9.6(0.8)	0.80(0.02)	6.8(2.0)	1.2(1.7)	0.12(0.12)
	CQPCorr	9.8(0.5)	10.1(0.7)	0.81(0.02)	8.9(1.0)	1.4(1.0)	0.12(0.08)
2	AFT	7.9(1.6)	8.4(1.7)	0.76(0.04)	7.3(2.1)	9.3(10.2)	0.44(0.24)
	Cox	8.6(1.4)	8.9(1.5)	0.77(0.04)	7.5(1.6)	4.3(7.2)	0.26(0.20)
	CQR	3.6(1.7)	5.0(1.5)	0.67(0.03)	7.7(1.8)	112.1(35.7)	0.93(0.02)
	KMW-LAD	7.8(1.6)	8.5(1.5)	0.77(0.04)	4.4(2.2)	1.0(1.8)	0.11(0.18)
	CQPCorr	8.6(1.1)	9.1(1.0)	0.80(0.02)	7.2(1.6)	1.0(1.0)	0.11(0.10)
3	AFT	6.0(2.2)	7.2(2.3)	0.74(0.06)	4.3(3.2)	8.0(18.3)	0.28(0.30)
	Cox	7.6(1.9)	8.3(1.8)	0.77(0.05)	5.8(2.5)	2.4(3.0)	0.23(0.20)
	CQR	3.4(1.4)	4.9(1.5)	0.67(0.05)	7.0(1.6)	102.6(34.2)	0.93(0.02)
	KMW-LAD	6.9(1.4)	8.3(1.5)	0.77(0.05)	3.1(2.2)	0.6(1.0)	0.12(0.16)
	CQPCorr	8.2(1.6)	9.2(1.2)	0.80(0.04)	5.4(2.1)	0.9(1.0)	0.11(0.10)

Table A.14: Simulation results for setting C2 with the AR correlation structure ($\rho = 0.5$) and 35% censoring rate. In each cell, mean (sd) based on 200 replicates.

Error	Method	TP20	TP40	pAUC	TP.FDR	FP.FDR	E.FDR
1	AFT	9.7(1.0)	10.2(1.2)	0.79(0.05)	10.6(1.7)	64.6(82.3)	0.72(0.20)
	Cox	9.6(1.3)	10.3(1.4)	0.82(0.04)	9.2(1.6)	7.3(7.3)	0.37(0.17)
	CQR	4.0(1.8)	5.6(2.1)	0.69(0.06)	8.7(2.0)	121.3(36.7)	0.93(0.03)
	KMW-LAD	7.9(1.7)	9.3(2.1)	0.80(0.06)	3.7(2.0)	0.9(1.0)	0.19(0.22)
	CQPCorr	8.8(1.2)	9.7(1.4)	0.82(0.04)	7.3(1.8)	2.0(2.0)	0.18(0.15)
2	AFT	5.1(2.8)	6.1(2.8)	0.71(0.08)	2.8(3.1)	4.7(8.3)	0.46(0.41)
	Cox	6.6(2.2)	7.5(1.8)	0.75(0.06)	4.0(2.0)	2.2(2.5)	0.25(0.21)
	CQR	3.0(1.1)	4.6(1.4)	0.67(0.04)	7.2(1.6)	102.7(37.7)	0.93(0.03)
	KMW-LAD	5.8(1.6)	6.6(2.3)	0.73(0.07)	1.6(1.6)	0.1(0.3)	0.03(0.09)
	CQPCorr	6.8(2.1)	7.8(2.4)	0.76(0.07)	2.9(2.6)	0.4(0.8)	0.04(0.08)
3	AFT	2.6(2.1)	3.7(2.4)	0.63(0.06)	0.9(1.7)	2.0(3.8)	0.47(0.47)
	Cox	4.8(2.0)	6.2(2.0)	0.71(0.07)	2.4(2.2)	1.3(1.9)	0.33(0.35)
	CQR	2.1(1.1)	3.1(1.5)	0.62(0.04)	5.5(1.7)	110.7(28.8)	0.95(0.02)
	KMW-LAD	5.2(1.6)	6.0(1.4)	0.70(0.05)	0.8(1.2)	0.2(0.6)	0.13(0.29)
	CQPCorr	6.3(2.1)	7.2(1.9)	0.75(0.06)	3.1(2.1)	0.9(1.3)	0.15(0.16)

Table A.15: Simulation results for setting C1 with the AR correlation structure ($\rho = 0.5$) and various values of τ. In each cell, mean (sd) based on 200 replicates.

Error	Method	τ	TP20	TP40	pAUC	TP.FDR	FP.FDR	E.FDR
1	CQR	0.2	4.9(1.7)	7.1(2.0)	0.73(0.04)	11.5(2.0)	140.9(48.5)	0.92(0.03)
		0.35	4.9(1.6)	6.8(1.9)	0.72(0.05)	11.2(2.0)	141.1(44.8)	0.92(0.03)
		0.5	4.7(1.6)	6.9(1.8)	0.74(0.06)	11.5(2.0)	133.0(33.1)	0.92(0.02)
		0.65	5.3(1.9)	7.2(1.9)	0.72(0.04)	11.1(1.9)	137.8(43.0)	0.92(0.02)
		0.8	4.7(1.6)	6.8(1.9)	0.72(0.04)	11.2(1.8)	140.2(44.8)	0.92(0.02)
	CQPCorr	0.2	9.6(1.7)	10.8(1.8)	0.86(0.05)	11.5(2.5)	4.3(2.7)	0.30(0.13)
		0.35	11.4(2.0)	12.8(1.6)	0.92(0.04)	9.8(2.1)	3.6(2.7)	0.24(0.15)
		0.5	12.2(1.6)	13.8(1.5)	0.94(0.03)	10.9(2.0)	3.3(2.5)	0.21(0.12)
		0.65	10.5(1.5)	12.4(1.5)	0.91(0.03)	8.9(1.8)	3.3(2.4)	0.24(0.15)
		0.8	9.2(1.4)	10.3(1.4)	0.85(0.05)	8.4(1.7)	4.0(3.3)	0.28(0.17)
2	CQR	0.2	5.6(1.7)	7.1(1.9)	0.73(0.06)	9.7(1.9)	110.1(39.9)	0.91(0.03)
		0.35	5.3(1.4)	7.3(1.8)	0.72(0.06)	9.5(1.8)	107.3(42.3)	0.91(0.04)
		0.5	5.2(1.4)	7.1(1.5)	0.73(0.04)	9.6(2.1)	108.4(23.8)	0.91(0.02)
		0.65	5.3(1.7)	7.3(1.9)	0.73(0.05)	9.9(2.0)	108.4(40.6)	0.91(0.03)
		0.8	5.1(1.8)	7.1(2.2)	0.73(0.06)	10.2(1.6)	110.2(40.3)	0.91(0.03)
	CQPCorr	0.2	7.2(1.5)	8.0(1.6)	0.77(0.03)	9.7(1.9)	2.6(1.7)	0.32(0.15)
		0.35	10.3(1.4)	11.4(1.8)	0.88(0.05)	8.3(1.7)	1.7(1.8)	0.15(0.13)
		0.5	10.4(1.7)	12.0(2.0)	0.89(0.05)	8.0(2.4)	1.6(1.2)	0.16(0.09)
		0.65	10.1(1.9)	11.6(1.6)	0.88(0.04)	8.0(2.8)	2.1(2.1)	0.18(0.14)
		0.8	7.9(1.8)	9.2(2.0)	0.81(0.05)	5.9(2.5)	2.3(2.1)	0.24(0.15)
3	CQR	0.2	4.8(1.6)	6.4(1.8)	0.71(0.04)	9.0(2.0)	99.6(36.7)	0.91(0.03)
		0.35	4.6(1.6)	6.3(1.5)	0.71(0.05)	8.8(2.4)	97.7(37.3)	0.91(0.04)
		0.5	4.5(1.6)	6.3(1.7)	0.70(0.05)	9.0(1.9)	105.8(44.7)	0.92(0.03)
		0.65	4.8(1.6)	6.1(2.0)	0.70(0.05)	8.6(2.2)	97.6(36.3)	0.91(0.03)
		0.8	4.9(1.6)	6.3(1.8)	0.71(0.05)	9.0(2.2)	101.6(38.1)	0.91(0.03)
	CQPCorr	0.2	6.1(2.1)	7.5(2.2)	0.75(0.06)	9.0(2.5)	1.3(1.6)	0.19(0.18)
		0.35	9.2(1.8)	10.6(1.6)	0.85(0.05)	6.8(2.1)	1.4(1.3)	0.16(0.12)
		0.5	10.7(1.7)	12.2(1.9)	0.90(0.06)	7.5(2.4)	1.3(1.4)	0.14(0.11)
		0.65	9.6(1.5)	11.0(1.6)	0.87(0.04)	6.4(2.2)	1.4(1.3)	0.16(0.13)
		0.8	6.8(1.7)	8.1(2.0)	0.77(0.06)	3.7(2.0)	1.5(1.9)	0.22(0.20)

Table A.16: Data analysis: numbers of overlapping interactions (RV-coefficients) identified by different methods. Upper panel: results based on FDR control. Lower panel: results based on (roughly) top forty lists.

LUAD	AFT	Cox	CQR	KMW-LAD	CQPCorr
AFT	8(1.00)	2(0.71)	4(0.77)	0(0.41)	2(0.85)
Cox		29(1.00)	16(0.87)	0(0.47)	13(0.84)
CQR			620(1.00)	2(0.64)	15(0.89)
KMW-LAD				4(1.00)	0(0.50)
CQPCorr					48(1.00)
SKCM	AFT	Cox	CQR	KMW-LAD	CQPCorr
AFT	17(1.00)	16(0.74)	11(0.60)	0(0.00)	6(0.77)
Cox		573(1.00)	101(0.53)	1(0.00)	44(0.72)
CQR			741(1.00)	5(0.02)	20(0.81)
KMW-LAD				20(1.00)	0(0.00)
CQPCorr					80(1.00)

LUAD	AFT	Cox	CQR	KMW-LAD	CQPCorr
AFT	40(1.00)	14(1.00)	1(0.40)	1(0.27)	4(0.20)
Cox		40(1.00)	1(0.39)	0(0.25)	5(0.19)
CQR			40(1.00)	1(0.71)	1(0.44)
KMW-LAD				40(1.00)	3(0.40)
CQPCorr					46(1.00)
SKCM	AFT	Cox	CQR	KMW-LAD	CQPCorr
AFT	40(1.00)	12(0.87)	5(0.14)	0(0.01)	7(0.53)
Cox		40(1.00)	0(0.10)	0(0.00)	7(0.48)
CQR			47(1.00)	2(0.02)	3(0.14)
KMW-LAD				45(1.00)	0(0.01)
CQPCorr					43(1.00)

A.2 Penalized Trimmed Estimation and Selection for Joint Interaction Analysis

Table A.17: Outlier detection results under simulation scenarios with continuous G factors and AR structure under linear model. TP: true positive outliers. FP: false positive outliers. In each cell, mean (sd) based on 200 replicates.

μ	D1: $N(0,1)$		D2: $0.9N(0,1)+$ $0.1Cauchy(0,5)$		D3: $0.9N(0,1)+$ $0.1N(20,1)$		D4: $N(0,1)$ and with leverage points		D5: $0.9N(0,1)+$ $0.1Cauchy(0,5)$ and with leverage points	
	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
1.0	0.0(0.0)	98.6(4.5)	17.9(2.1)	80.2(4.9)	25.0(0.1)	71.5(3.9)	12.9(2.1)	87.4(4.1)	19.2(2.0)	80.3(5.2)
1.1	0.0(0.0)	85.0(4.5)	17.5(2.2)	68.8(4.6)	25.0(0.1)	61.6(4.3)	11.9(2.1)	76.4(4.7)	18.9(2.2)	68.3(5.0)
1.2	0.0(0.0)	75.0(4.2)	17.0(2.4)	59.6(4.8)	25.0(0.0)	53.1(4.5)	11.0(2.0)	65.8(4.1)	18.3(2.0)	58.6(4.6)
1.3	0.0(0.0)	65.2(5.8)	16.7(2.5)	50.9(5.1)	25.0(0.0)	45.4(4.7)	10.2(2.0)	56.8(4.3)	18.0(2.2)	49.4(4.8)
1.4	0.0(0.0)	55.9(6.4)	16.5(2.4)	42.9(5.8)	25.0(0.0)	37.5(5.2)	9.7(1.8)	47.6(4.5)	17.8(2.2)	41.7(5.1)
1.5	0.0(0.0)	47.4(6.4)	16.1(2.6)	36.1(5.5)	25.0(0.0)	30.9(5.6)	9.0(1.8)	40.0(3.7)	17.5(2.3)	34.1(5.1)
1.6	0.0(0.0)	40.3(7.1)	15.9(2.6)	30.1(5.4)	25.0(0.0)	25.2(5.5)	8.2(1.6)	33.4(4.7)	17.2(2.2)	27.6(4.7)
1.7	0.0(0.0)	33.7(6.8)	15.7(2.8)	24.7(5.4)	25.0(0.0)	20.1(5.1)	7.9(1.5)	27.2(4.4)	16.9(2.3)	22.5(4.3)
1.8	0.0(0.0)	28.3(6.8)	15.3(2.8)	19.5(4.8)	25.0(0.0)	15.9(4.7)	7.5(1.4)	23.1(4.8)	16.7(2.4)	17.6(4.3)
1.9	0.0(0.0)	23.4(6.4)	15.1(2.8)	15.7(4.5)	25.0(0.0)	12.7(4.4)	7.3(1.3)	18.2(4.4)	16.4(2.4)	14.3(4.0)
2.0	0.0(0.0)	19.1(5.7)	14.8(2.7)	12.6(3.9)	25.0(0.0)	10.0(4.2)	7.0(1.3)	14.6(3.9)	16.1(2.5)	11.0(3.6)
2.1	0.0(0.0)	15.7(5.4)	14.4(2.7)	10.0(4.0)	25.0(0.0)	7.9(3.9)	6.6(1.2)	11.4(3.9)	15.7(2.5)	8.7(3.2)
2.2	0.0(0.0)	12.8(5.4)	14.2(2.7)	7.8(3.6)	25.0(0.0)	5.8(2.9)	6.3(1.1)	9.0(3.5)	15.4(2.5)	6.9(3.3)
2.3	0.0(0.0)	10.4(4.7)	13.9(2.7)	5.8(3.2)	25.0(0.0)	4.4(2.3)	6.3(1.0)	6.9(2.9)	15.1(2.6)	5.2(2.9)
2.4	0.0(0.0)	8.5(4.0)	13.6(2.8)	4.5(2.7)	25.0(0.0)	3.4(2.2)	6.1(0.9)	5.3(2.7)	14.8(2.7)	3.6(2.4)
2.5	**0.0(0.0)**	**6.9(3.5)**	**13.2(2.9)**	**3.4(2.3)**	**25.0(0.0)**	**2.5(1.9)**	**5.8(0.9)**	**4.1(2.1)**	**14.4(2.6)**	**2.5(1.8)**
2.6	0.0(0.0)	5.4(3.2)	12.9(2.9)	2.5(2.0)	25.0(0.0)	1.9(1.6)	5.7(0.8)	3.3(2.0)	14.0(2.7)	2.0(1.7)
2.7	0.0(0.0)	4.2(2.9)	12.7(2.8)	2.0(1.9)	25.0(0.0)	1.5(1.5)	5.5(0.8)	2.4(1.8)	13.7(2.8)	1.5(1.7)
2.8	0.0(0.0)	3.2(2.5)	12.4(2.9)	1.5(1.7)	25.0(0.0)	1.0(1.3)	5.2(0.9)	1.9(1.6)	13.2(2.8)	1.2(1.2)
2.9	0.0(0.0)	2.4(2.0)	12.1(2.9)	1.2(1.4)	25.0(0.0)	0.7(1.1)	5.2(0.9)	1.4(1.4)	13.0(2.6)	1.0(1.1)
3.0	0.0(0.0)	1.8(1.7)	11.9(2.8)	0.9(1.2)	25.0(0.0)	0.4(0.7)	5.0(0.9)	1.0(1.2)	12.3(2.6)	0.7(1.0)
3.1	0.0(0.0)	1.4(1.4)	11.6(2.9)	0.8(1.1)	25.0(0.0)	0.3(0.6)	5.0(0.6)	0.8(1.1)	12.1(2.6)	0.6(1.0)
3.2	0.0(0.0)	1.1(1.3)	11.4(2.8)	0.6(1.0)	24.6(2.6)	0.2(0.5)	5.0(0.7)	0.6(0.9)	11.9(2.6)	0.4(0.7)
3.3	0.0(0.0)	0.9(1.1)	11.1(2.9)	0.4(0.8)	24.1(3.6)	0.2(0.5)	4.9(0.7)	0.6(1.0)	11.4(2.5)	0.4(0.7)
3.4	0.0(0.0)	0.7(1.0)	10.8(3.0)	0.3(0.7)	23.1(5.6)	0.2(0.4)	4.9(0.6)	0.4(0.7)	11.0(2.7)	0.2(0.5)
3.5	0.0(0.0)	0.6(0.9)	10.6(3.0)	0.3(0.6)	21.0(8.1)	0.1(0.4)	4.8(0.6)	0.3(0.6)	10.9(2.7)	0.1(0.4)
3.6	0.0(0.0)	0.4(0.7)	10.4(2.9)	0.2(0.5)	15.5(10.0)	0.1(0.4)	4.7(0.7)	0.2(0.6)	10.6(2.6)	0.1(0.4)
3.7	0.0(0.0)	0.4(0.7)	10.0(2.9)	0.1(0.4)	12.1(10.4)	0.0(0.2)	4.6(0.8)	0.2(0.7)	10.4(2.5)	0.1(0.4)
3.8	0.0(0.0)	0.2(0.5)	9.9(2.8)	0.1(0.3)	8.3(9.2)	0.0(0.1)	4.6(0.8)	0.1(0.5)	9.9(2.5)	0.0(0.2)
3.9	0.0(0.0)	0.2(0.5)	9.7(2.7)	0.1(0.3)	5.1(7.4)	0.0(0.1)	4.6(0.8)	0.1(0.3)	9.4(2.3)	0.0(0.2)
4.0	0.0(0.0)	0.1(0.3)	9.5(2.7)	0.1(0.3)	2.8(4.4)	0.0(0.1)	4.6(0.9)	0.1(0.2)	9.4(2.4)	0.0(0.2)

Table A.18: Summary results under simulation scenarios with continuous G factors and Band structure under linear model. In each cell, mean (sd) based on 200 replicates.

	M:TP	M:FP	I:TP	I:FP	RSSE	PMSE
			D1: $N(0,1)$			
LTS-MCP-Hier	7.8(0.4)	0.8(1.9)	12.0(2.0)	1.0(1.1)	2.34(0.53)	0.99(0.53)
LS-MCP	5.5(0.9)	3.9(4.4)	10.9(0.9)	12.9(11.1)	2.92(0.45)	1.31(0.62)
LAD-Lasso	8.0(0.1)	11.4(5.6)	13.1(1.5)	30.4(11.6)	1.78(0.39)	1.47(0.56)
RLARS	7.4(0.7)	0.7(1.1)	7.5(1.8)	11.4(7.2)	3.25(0.44)	2.29(0.80)
LTS-MCP	6.0(1.0)	6.5(3.3)	10.7(1.0)	25.8(8.9)	2.63(0.49)	1.34(0.25)
LS-MCP-Hier	8.0(0.1)	0.5(1.7)	12.7(1.1)	0.5(0.7)	1.79(0.35)	0.81(0.19)
			D2: $0.9N(0,1)+0.1Cauchy(0,5)$			
LTS-MCP-Hier	7.8(0.4)	0.9(2.9)	11.4(2.0)	1.1(1.0)	2.28(0.50)	1.14(0.60)
LS-MCP	2.1(1.8)	18.2(8.0)	2.3(2.4)	71.9(11.6)	30.32(40.12)	547.58(2145.90)
LAD-Lasso	7.5(0.6)	2.8(2.2)	7.0(2.4)	7.5(3.4)	3.13(0.35)	4.11(1.22)
RLARS	7.1(0.8)	0.8(1.1)	6.0(1.8)	11.4(6.8)	3.67(0.50)	3.33(1.23)
LTS-MCP	5.8(0.9)	8.4(3.9)	10.4(1.1)	29.5(10.3)	2.80(0.43)	1.37(0.36)
LS-MCP-Hier	5.8(1.4)	150.7(118.1)	2.4(3.2)	27.2(74.9)	28.76(42.09)	1181.23(5245.02)
			D3: $0.9N(0,1)+0.1N(20,1)$			
LTS-MCP-Hier	7.8(0.4)	0.9(2.3)	11.7(1.8)	1.0(1.1)	2.15(0.47)	1.05(0.50)
LS-MCP	2.7(1.0)	24.1(5.2)	2.7(1.4)	67.9(5.2)	9.96(0.68)	33.43(7.57)
LAD-Lasso	7.3(0.8)	3.0(1.8)	5.4(2.1)	8.0(2.7)	3.39(0.34)	5.09(1.51)
RLARS	5.8(1.2)	1.4(1.4)	3.7(1.8)	11.4(4.9)	4.31(0.48)	5.55(2.15)
LTS-MCP	6.0(0.9)	7.3(3.7)	10.7(1.0)	26.4(8.9)	2.67(0.49)	1.20(0.28)
LS-MCP-Hier	6.1(0.9)	94.6(7.1)	2.4(1.5)	5.3(5.2)	8.79(0.62)	33.32(6.65)
			D4: $N(0,1)$ and with leverage points			
LTS-MCP-Hier	7.0(1.3)	6.8(13.7)	9.9(3.5)	2.8(2.2)	2.91(0.90)	1.26(2.64)
LS-MCP	1.3(0.9)	22.2(5.2)	3.1(1.9)	69.0(6.5)	7.21(0.86)	18.85(6.15)
LAD-Lasso	3.9(1.2)	4.1(2.3)	4.0(1.6)	12.8(3.5)	4.05(0.33)	9.23(2.16)
RLARS	7.1(0.8)	0.8(1.1)	6.9(1.9)	12.1(7.9)	3.42(0.43)	2.88(0.96)
LTS-MCP	5.8(1.2)	8.9(4.4)	10.4(1.3)	31.7(11.1)	2.78(0.58)	3.12(0.68)
LS-MCP-Hier	5.2(1.2)	56.4(35.5)	3.4(2.7)	5.0(3.3)	5.44(1.27)	14.17(8.50)
			D5: $0.9N(0,1)+0.1Cauchy(0,5)$ and with leverage points			
LTS-MCP-Hier	7.2(1.2)	5.6(12.7)	9.8(2.9)	2.7(2.2)	2.75(0.87)	1.09(2.28)
LS-MCP	0.5(0.7)	18.0(9.6)	1.5(1.5)	69.0(16.7)	25.10(32.12)	258.07(680.53)
LAD-Lasso	3.6(1.4)	4.4(2.1)	4.0(2.1)	12.4(3.4)	4.06(0.36)	9.17(2.26)
RLARS	6.7(0.9)	1.1(1.3)	5.9(1.6)	13.3(7.4)	3.77(0.43)	3.65(1.24)
LTS-MCP	6.0(1.0)	9.3(4.0)	10.7(1.2)	32.7(8.5)	2.67(0.52)	2.55(0.35)
LS-MCP-Hier	4.3(1.6)	154.0(110.7)	1.0(1.8)	26.0(61.5)	27.86(39.79)	1019.76(4540.05)

Table A.19: Summary results under simulation scenarios with continuous G factors and Band structure under AFT model. In each cell, mean (sd) based on 200 replicates.

	M:TP	M:FP	I:TP	I:FP	RSSE	Cstat
			D1: $N(0,1)$			
LTS-MCP-Hier	7.8(0.5)	8.0(9.5)	10.1(2.7)	0.8(1.0)	2.67(0.50)	0.90(0.03)
LS-MCP	6.2(1.0)	13.0(5.4)	11.0(0.9)	38.9(10.2)	2.59(0.54)	0.92(0.02)
LAD-Lasso	7.4(0.9)	13.8(7.9)	6.6(4.0)	31.1(16.6)	3.32(0.61)	0.83(0.06)
RLARS	7.2(0.8)	10.5(2.8)	3.1(1.4)	22.2(4.4)	4.22(0.35)	0.78(0.05)
LTS-MCP	5.8(1.0)	14.8(4.6)	6.6(1.7)	57.8(7.4)	3.36(0.31)	0.85(0.03)
LS-MCP-Hier	8.0(0.2)	2.6(4.4)	11.7(1.3)	0.8(1.1)	2.04(0.35)	0.92(0.02)
			D2: $0.9N(0,1)+0.1Cauchy(0,5)$			
LTS-MCP-Hier	7.7(0.6)	7.5(7.6)	9.2(2.7)	1.2(1.1)	2.88(0.58)	0.88(0.03)
LS-MCP	1.1(1.2)	12.9(7.8)	1.2(1.5)	61.1(8.9)	46.82(93.14)	0.56(0.07)
LAD-Lasso	5.8(1.5)	4.9(2.4)	1.8(1.5)	12.4(3.4)	4.08(0.35)	0.74(0.06)
RLARS	6.1(1.6)	7.3(3.9)	1.5(1.1)	24.1(5.9)	5.42(5.41)	0.72(0.06)
LTS-MCP	5.7(1.2)	16.0(4.1)	5.7(1.7)	58.5(6.1)	3.66(0.35)	0.83(0.03)
LS-MCP-Hier	5.5(1.4)	193.2(160.4)	2.1(2.4)	73.5(236.8)	57.08(118.64)	0.58(0.07)
			D3: $0.9N(0,1)+0.1N(20,1)$			
LTS-MCP-Hier	8.0(0.1)	3.8(6.7)	11.4(1.7)	0.9(1.0)	2.12(0.42)	0.92(0.01)
LS-MCP	2.5(1.0)	26.6(5.4)	2.6(1.3)	70.5(6.3)	10.76(0.69)	0.63(0.04)
LAD-Lasso	6.6(1.1)	4.3(2.2)	2.9(1.9)	11.0(3.0)	3.77(0.33)	0.78(0.05)
RLARS	6.3(1.1)	4.2(2.9)	1.4(1.2)	12.0(5.7)	4.40(0.39)	0.77(0.04)
LTS-MCP	6.0(1.1)	11.2(4.0)	9.2(1.8)	47.4(9.4)	2.93(0.50)	0.89(0.02)
LS-MCP-Hier	5.9(1.1)	101.8(7.3)	2.5(1.6)	7.0(6.3)	9.68(0.59)	0.66(0.04)
			D4: $N(0,1)$ and with leverage points			
LTS-MCP-Hier	6.8(1.2)	10.3(9.7)	9.1(3.4)	1.7(1.5)	3.46(0.71)	0.84(0.07)
LS-MCP	3.1(1.1)	15.2(4.5)	4.9(2.1)	53.2(6.0)	4.98(0.62)	0.74(0.05)
LAD-Lasso	6.0(1.3)	7.2(5.8)	3.4(2.5)	18.8(10.2)	3.92(0.38)	0.76(0.05)
RLARS	6.8(1.0)	12.1(4.0)	2.8(1.6)	21.8(4.7)	4.41(0.39)	0.76(0.04)
LTS-MCP	5.4(1.3)	15.8(3.9)	5.5(1.9)	61.7(5.9)	3.73(0.41)	0.81(0.04)
LS-MCP-Hier	5.8(1.1)	42.5(23.6)	4.5(2.4)	2.9(2.1)	4.18(0.64)	0.77(0.05)
			D5: $0.9N(0,1)+0.1Cauchy(0,5)$ and with leverage points			
LTS-MCP-Hier	7.1(1.1)	10.1(12.0)	9.5(3.2)	1.8(1.3)	3.30(0.75)	0.84(0.06)
LS-MCP	0.9(1.0)	13.1(7.1)	1.1(1.3)	57.1(10.2)	36.77(71.77)	0.56(0.06)
LAD-Lasso	5.2(1.7)	4.6(2.3)	1.8(1.5)	12.9(3.7)	4.16(0.33)	0.73(0.06)
RLARS	6.2(1.3)	8.5(4.6)	2.5(1.5)	22.6(6.7)	6.82(16.01)	0.73(0.06)
LTS-MCP	5.8(1.1)	15.9(4.3)	5.1(1.6)	61.1(5.2)	3.74(0.42)	0.81(0.04)
LS-MCP-Hier	5.0(1.7)	173.4(152.1)	2.0(2.3)	66.5(240.4)	52.85(129.70)	0.57(0.06)

Table A.20: Summary results under simulation scenarios with categorical G factors and AR structure under linear model. In each cell, mean (sd) based on 200 replicates.

	M:TP	M:FP	I:TP	I:FP	RSSE	PMSE
			D1: $N(0,1)$			
LTS-MCP-Hier	7.9(0.3)	0.3(1.0)	12.3(1.5)	0.6(0.9)	2.03(0.42)	0.95(0.44)
LS-MCP	6.2(1.1)	4.5(4.2)	11.3(1.3)	14.3(11.5)	2.48(0.54)	1.13(0.37)
LAD-Lasso	8.0(0.0)	10.7(5.7)	13.4(1.0)	27.9(10.1)	1.68(0.31)	1.37(0.41)
RLARS	4.1(1.2)	13.2(5.8)	2.3(1.6)	8.4(5.0)	4.73(0.44)	8.87(3.01)
LTS-MCP	6.7(0.9)	7.0(3.3)	11.3(1.1)	27.6(8.1)	2.19(0.49)	1.19(0.25)
LS-MCP-Hier	8.0(0.0)	0.3(0.8)	13.1(0.9)	0.5(0.7)	1.66(0.27)	0.80(0.17)
			D2: $0.9N(0,1)+0.1Cauchy(0,5)$			
LTS-MCP-Hier	8.0(0.2)	1.0(2.8)	11.8(2.0)	0.9(1.1)	2.13(0.47)	1.07(0.45)
LS-MCP	1.9(1.8)	21.1(8.4)	2.2(2.4)	74.7(11.4)	35.47(49.17)	712.17(2635.35)
LAD-Lasso	7.8(0.4)	2.4(1.7)	7.8(2.3)	7.2(3.5)	3.02(0.34)	4.05(1.16)
RLARS	4.0(1.1)	11.4(5.3)	1.9(1.4)	7.8(4.4)	4.85(0.41)	9.86(3.38)
LTS-MCP	6.5(1.0)	8.5(3.5)	11.1(1.2)	32.1(7.9)	2.35(0.52)	1.24(0.28)
LS-MCP-Hier	6.0(1.5)	153.5(116.8)	2.4(3.2)	24.5(72.4)	28.86(41.92)	1214.97(5318.06)
			D3: $0.9N(0,1)+0.1N(20,1)$			
LTS-MCP-Hier	8.0(0.2)	0.5(1.4)	12.3(1.6)	0.8(0.9)	1.94(0.41)	0.93(0.41)
LS-MCP	2.8(1.1)	25.0(5.4)	2.7(1.4)	67.2(6.0)	9.90(0.71)	34.37(7.31)
LAD-Lasso	7.5(0.6)	2.8(1.9)	5.6(2.2)	8.4(2.6)	3.33(0.33)	5.04(1.54)
RLARS	3.8(1.1)	10.1(3.8)	0.9(0.9)	6.3(3.2)	5.14(0.50)	11.87(3.95)
LTS-MCP	6.7(1.1)	7.6(3.5)	11.4(1.1)	28.0(7.5)	2.21(0.52)	1.03(0.22)
LS-MCP-Hier	6.4(1.0)	94.4(7.5)	2.2(1.5)	5.1(5.6)	8.64(0.53)	31.64(6.01)
			D4: $N(0,1)$ and with leverage points			
LTS-MCP-Hier	7.7(0.6)	7.7(12.3)	10.1(3.0)	1.3(1.1)	2.50(0.61)	1.71(1.08)
LS-MCP	4.4(1.3)	21.5(5.7)	6.8(1.9)	55.4(7.3)	4.95(0.79)	6.95(2.98)
LAD-Lasso	7.0(0.9)	7.6(4.0)	4.7(2.4)	9.0(4.1)	3.51(0.37)	6.35(1.97)
RLARS	5.7(1.1)	13.7(6.6)	2.7(1.7)	8.7(4.8)	4.46(0.42)	7.11(2.26)
LTS-MCP	6.2(1.1)	11.9(5.1)	10.5(1.6)	34.4(7.0)	2.67(0.59)	2.17(0.56)
LS-MCP-Hier	7.7(0.6)	52.4(20.8)	6.5(2.2)	2.4(2.4)	3.61(0.67)	4.38(1.90)
			D5: $0.9N(0,1)+0.1Cauchy(0,5)$ and with leverage points			
LTS-MCP-Hier	7.9(0.4)	1.7(4.7)	11.2(2.3)	1.0(1.1)	2.27(0.53)	1.56(0.78)
LS-MCP	1.4(1.4)	22.2(7.7)	1.6(2.0)	76.1(10.6)	39.92(54.72)	771.74(2428.31)
LAD-Lasso	7.2(0.9)	4.5(2.6)	4.6(2.2)	8.0(2.9)	3.51(0.31)	6.12(1.65)
RLARS	5.6(1.1)	11.0(5.6)	2.7(1.8)	9.3(5.8)	4.50(0.40)	7.18(2.21)
LTS-MCP	6.4(0.9)	10.1(4.2)	11.1(1.4)	32.8(7.5)	2.43(0.55)	1.86(0.38)
LS-MCP-Hier	5.4(1.5)	164.5(117.5)	1.7(2.4)	29.5(75.9)	31.75(42.57)	1196.18(4354.42)

Table A.21: Summary results under simulation scenarios with categorical G factors and AR structure under AFT model. In each cell, mean (sd) based on 200 replicates.

	M:TP	M:FP	I:TP	I:FP	RSSE	Cstat
			D1: $N(0,1)$			
LTS-MCP-Hier	7.8(0.4)	6.4(10.0)	10.1(2.7)	0.8(1.0)	2.43(0.53)	0.90(0.03)
LS-MCP	7.0(1.0)	12.3(4.6)	12.1(1.2)	38.1(8.7)	1.92(0.67)	0.92(0.01)
LAD-Lasso	7.5(0.7)	14.6(8.2)	8.0(4.1)	33.2(16.6)	3.16(0.60)	0.85(0.05)
RLARS	2.6(1.3)	3.5(2.8)	0.7(0.9)	34.4(6.8)	5.65(0.54)	0.61(0.07)
LTS-MCP	6.1(1.0)	15.2(4.3)	6.7(1.7)	57.4(8.5)	3.27(0.32)	0.85(0.03)
LS-MCP-Hier	7.9(0.3)	1.3(3.0)	12.3(1.2)	0.5(0.8)	1.89(0.36)	0.92(0.02)
			D2: $0.9N(0,1)+0.1Cauchy(0,5)$			
LTS-MCP-Hier	7.8(0.4)	6.2(5.4)	9.3(3.3)	1.1(1.1)	2.64(0.59)	0.88(0.04)
LS-MCP	1.1(1.4)	16.5(8.8)	1.2(1.5)	63.9(9.9)	58.01(114.22)	0.55(0.07)
LAD-Lasso	5.8(1.6)	4.7(2.4)	1.9(1.3)	12.1(3.4)	4.09(0.33)	0.74(0.06)
RLARS	1.0(1.1)	2.8(2.3)	0.5(0.8)	32.4(8.8)	232.54(898.65)	0.55(0.05)
LTS-MCP	6.2(0.9)	15.5(4.3)	5.6(1.7)	59.7(4.9)	3.57(0.33)	0.83(0.03)
LS-MCP-Hier	5.4(1.5)	198.8(160.5)	2.0(2.7)	70.5(239.0)	59.85(123.92)	0.58(0.08)
			D3: $0.9N(0,1)+0.1N(20,1)$			
LTS-MCP-Hier	8.0(0.2)	1.2(3.0)	12.4(1.3)	0.6(0.8)	1.90(0.34)	0.92(0.01)
LS-MCP	2.5(1.1)	26.5(5.0)	2.5(1.5)	71.4(5.6)	10.67(0.75)	0.63(0.04)
LAD-Lasso	6.5(1.1)	4.3(2.5)	2.8(1.7)	10.8(3.5)	3.79(0.29)	0.78(0.04)
RLARS	1.2(1.0)	1.6(1.7)	0.5(0.7)	25.7(9.8)	5.75(0.75)	0.60(0.06)
LTS-MCP	6.3(0.8)	12.2(4.0)	9.4(1.7)	48.6(9.6)	2.77(0.52)	0.90(0.02)
LS-MCP-Hier	5.9(1.2)	101.6(7.2)	2.2(1.7)	7.6(6.1)	9.60(0.61)	0.66(0.04)
			D4: $N(0,1)$ and with leverage points			
LTS-MCP-Hier	7.2(0.8)	15.4(10.4)	8.7(2.4)	1.2(1.1)	3.67(0.44)	0.85(0.05)
LS-MCP	3.3(1.3)	18.8(4.1)	3.0(1.7)	53.8(5.1)	6.07(0.70)	0.67(0.05)
LAD-Lasso	2.3(1.6)	11.5(4.9)	0.3(0.7)	14.5(6.7)	4.48(0.29)	0.63(0.05)
RLARS	3.9(1.2)	21.4(6.0)	0.3(0.5)	17.4(6.3)	5.38(0.40)	0.64(0.04)
LTS-MCP	5.6(1.1)	19.2(5.3)	4.2(1.6)	59.4(5.9)	4.00(0.35)	0.78(0.04)
LS-MCP-Hier	5.9(1.1)	66.0(8.2)	1.7(1.3)	2.9(2.7)	5.03(0.67)	0.71(0.05)
			D5: $0.9N(0,1)+0.1Cauchy(0,5)$ and with leverage points			
LTS-MCP-Hier	7.5(0.7)	13.6(19.0)	9.1(3.4)	1.0(1.1)	3.28(0.62)	0.84(0.05)
LS-MCP	0.5(0.8)	16.2(7.7)	0.4(0.7)	64.4(9.4)	63.10(119.35)	0.52(0.03)
LAD-Lasso	1.6(1.6)	12.9(5.6)	0.2(0.5)	10.1(5.0)	4.60(0.27)	0.60(0.05)
RLARS	3.4(1.6)	15.4(8.1)	0.4(0.7)	18.8(6.2)	63.57(153.77)	0.61(0.06)
LTS-MCP	6.0(1.0)	16.8(4.2)	5.0(2.0)	59.0(5.1)	3.82(0.37)	0.81(0.03)
LS-MCP-Hier	4.6(2.2)	201.0(155.2)	1.6(2.8)	87.9(263.4)	63.33(132.06)	0.51(0.05)

Table A.22: Summary results under simulation scenarios with categorical G factors and Band structure under linear model. In each cell, mean (sd) based on 200 replicates.

	M:TP	M:FP	I:TP	I:FP	RSSE	PMSE
			D1: $N(0,1)$			
LTS-MCP-Hier	7.9(0.3)	0.4(1.1)	11.6(1.9)	0.8(1.0)	2.18(0.49)	0.98(0.45)
LS-MCP	5.9(1.2)	4.8(5.0)	11.1(1.0)	15.4(13.4)	2.64(0.59)	1.19(0.52)
LAD-Lasso	8.0(0.1)	12.7(5.5)	13.2(1.3)	31.8(11.8)	1.72(0.34)	1.42(0.48)
RLARS	4.1(1.4)	13.4(6.1)	2.3(1.4)	9.0(4.8)	4.72(0.46)	9.01(3.04)
LTS-MCP	6.2(1.1)	6.5(3.9)	11.0(1.1)	25.7(8.3)	2.40(0.59)	1.26(0.26)
LS-MCP-Hier	8.0(0.0)	0.4(1.1)	13.1(1.0)	0.5(0.7)	1.67(0.33)	0.79(0.18)
			D2: $0.9N(0,1)+0.1Cauchy(0,5)$			
LTS-MCP-Hier	7.9(0.3)	0.6(1.8)	11.8(1.5)	0.9(1.0)	2.14(0.40)	1.10(0.43)
LS-MCP	1.8(1.8)	21.8(8.9)	2.2(2.4)	74.2(10.9)	35.32(48.40)	673.99(2547.57)
LAD-Lasso	7.6(0.6)	2.7(2.0)	7.2(2.3)	7.3(3.2)	3.12(0.40)	4.32(1.44)
RLARS	4.0(1.4)	11.5(5.4)	1.8(1.4)	8.2(4.0)	4.81(0.41)	9.70(3.17)
LTS-MCP	6.3(1.0)	8.4(3.7)	10.8(1.3)	31.8(8.1)	2.52(0.51)	1.28(0.29)
LS-MCP-Hier	5.9(1.6)	152.2(118.4)	2.4(3.3)	25.3(74.6)	28.95(42.00)	1126.17(4531.24)
			D3: $0.9N(0,1)+0.1N(20,1)$			
LTS-MCP-Hier	7.9(0.2)	0.6(1.5)	12.0(1.4)	0.7(0.8)	2.07(0.41)	0.99(0.43)
LS-MCP	2.6(1.1)	25.3(5.3)	2.6(1.5)	68.3(5.4)	9.98(0.74)	33.57(6.78)
LAD-Lasso	7.2(0.8)	2.9(1.8)	5.6(2.2)	8.3(2.9)	3.41(0.33)	5.31(1.51)
RLARS	3.7(1.3)	9.6(3.9)	0.9(1.0)	6.4(3.0)	5.06(0.52)	11.42(4.16)
LTS-MCP	6.4(1.1)	7.8(3.3)	11.0(1.1)	27.5(7.3)	2.41(0.58)	1.16(0.23)
LS-MCP-Hier	6.1(1.1)	93.7(6.3)	2.4(1.4)	5.7(5.4)	8.57(0.49)	30.92(5.83)
			D4: $N(0,1)$ and with leverage points			
LTS-MCP-Hier	7.5(0.6)	7.5(8.7)	9.3(2.8)	1.2(1.2)	2.72(0.56)	1.82(1.11)
LS-MCP	4.1(1.2)	21.6(4.9)	6.7(1.7)	55.9(5.9)	5.00(0.69)	7.12(2.55)
LAD-Lasso	6.6(0.9)	7.0(3.9)	4.5(2.2)	9.1(3.5)	3.61(0.36)	6.63(1.89)
RLARS	5.3(1.3)	13.1(7.1)	2.7(1.6)	8.7(5.4)	4.52(0.37)	7.49(1.89)
LTS-MCP	6.0(1.1)	13.3(6.4)	10.0(2.0)	34.8(8.3)	2.85(0.55)	2.54(0.79)
LS-MCP-Hier	7.3(0.6)	49.5(21.6)	6.2(2.1)	2.1(2.2)	3.66(0.67)	4.63(1.97)
			D5: $0.9N(0,1)+0.1Cauchy(0,5)$ and with leverage points			
LTS-MCP-Hier	7.7(0.6)	2.3(6.9)	10.9(2.6)	0.9(1.0)	2.33(0.57)	1.55(0.81)
LS-MCP	1.4(1.5)	22.0(7.8)	1.6(1.8)	77.6(10.0)	39.63(52.58)	743.55(2252.92)
LAD-Lasso	6.8(0.9)	4.4(2.9)	4.5(1.9)	8.9(3.3)	3.60(0.31)	6.35(1.59)
RLARS	5.3(1.3)	10.0(5.2)	2.6(1.6)	9.3(5.7)	4.50(0.37)	7.29(2.12)
LTS-MCP	6.2(1.1)	9.6(3.8)	10.8(1.0)	32.7(7.3)	2.64(0.49)	1.93(0.37)
LS-MCP-Hier	5.3(1.6)	160.8(108.0)	1.6(1.6)	32.8(78.2)	31.77(42.35)	1097.13(4234.09)

Table A.23: Summary results under simulation scenarios with categorical G factors and Band structure under AFT model. In each cell, mean (sd) based on 200 replicates.

	M:TP	M:FP	I:TP	I:FP	RSSE	Cstat
			D1: $N(0,1)$			
LTS-MCP-Hier	7.9(0.4)	7.0(11.8)	11.8(2.8)	0.9(1.2)	2.52(0.56)	0.89(0.03)
LS-MCP	6.8(1.1)	14.4(5.1)	11.7(1.2)	41.4(7.5)	2.15(0.66)	0.92(0.02)
LAD-Lasso	7.3(1.0)	14.1(7.8)	6.8(4.2)	32.0(15.9)	3.36(0.58)	0.82(0.06)
RLARS	2.3(1.3)	2.9(1.9)	0.7(0.8)	34.7(7.6)	5.51(0.55)	0.61(0.06)
LTS-MCP	6.2(1.0)	14.8(4.0)	7.0(2.0)	56.1(7.2)	3.24(0.36)	0.85(0.04)
LS-MCP-Hier	7.9(0.2)	1.5(3.6)	12.2(1.3)	0.6(0.8)	1.92(0.36)	0.92(0.02)
			D2: $0.9N(0,1)+0.1Cauchy(0,5)$			
LTS-MCP-Hier	7.7(0.5)	8.9(7.4)	9.6(3.1)	1.1(1.1)	2.75(0.58)	0.88(0.04)
LS-MCP	1.0(1.3)	16.5(8.9)	0.8(1.1)	64.9(10.7)	56.73(108.19)	0.54(0.06)
LAD-Lasso	5.8(1.3)	4.9(2.4)	1.7(1.3)	12.9(3.5)	4.11(0.32)	0.74(0.05)
RLARS	0.9(0.9)	2.7(2.4)	0.4(0.6)	32.2(8.6)	198.16(1027.54)	0.54(0.05)
LTS-MCP	6.1(1.1)	15.9(3.9)	5.9(1.9)	59.1(5.8)	3.58(0.39)	0.82(0.03)
LS-MCP-Hier	5.5(1.6)	196.3(157.2)	1.9(2.4)	66.8(224.0)	57.70(118.80)	0.58(0.08)
			D3: $0.9N(0,1)+0.1N(20,1)$			
LTS-MCP-Hier	8.0(0.1)	1.8(3.4)	12.2(1.2)	0.6(0.7)	1.96(0.36)	0.92(0.01)
LS-MCP	2.5(1.0)	26.4(5.9)	2.2(1.2)	72.3(5.5)	10.74(0.70)	0.62(0.04)
LAD-Lasso	6.5(1.0)	4.6(1.8)	2.7(1.8)	11.1(3.0)	3.78(0.31)	0.77(0.04)
RLARS	1.2(0.9)	1.4(1.2)	0.4(0.6)	25.3(9.1)	81.94(762.49)	0.59(0.06)
LTS-MCP	6.2(1.1)	11.4(4.0)	9.5(1.6)	47.3(8.5)	2.81(0.41)	0.90(0.02)
LS-MCP-Hier	5.9(1.1)	101.8(8.3)	2.3(1.5)	7.4(6.8)	9.60(0.58)	0.66(0.04)
			D4: $N(0,1)$ and with leverage points			
LTS-MCP-Hier	7.0(1.0)	18.7(9.7)	8.5(2.3)	0.9(1.0)	3.85(0.42)	0.85(0.05)
LS-MCP	3.0(1.3)	18.7(4.0)	2.6(1.5)	54.4(4.6)	6.12(0.59)	0.65(0.04)
LAD-Lasso	2.2(1.4)	11.9(4.9)	0.3(0.5)	15.4(7.8)	4.52(0.25)	0.62(0.04)
RLARS	3.6(1.2)	21.3(4.4)	0.4(0.6)	18.0(5.5)	5.34(0.43)	0.63(0.05)
LTS-MCP	5.4(1.1)	20.1(4.8)	4.3(1.7)	58.0(4.9)	4.04(0.39)	0.78(0.04)
LS-MCP-Hier	5.5(1.2)	67.6(9.0)	1.7(1.4)	2.5(2.1)	5.06(0.60)	0.70(0.06)
			D5: $0.9N(0,1)+0.1Cauchy(0,5)$ and with leverage points			
LTS-MCP-Hier	7.4(0.7)	12.1(10.4)	9.0(3.0)	1.2(1.1)	3.35(0.59)	0.84(0.05)
LS-MCP	0.5(0.9)	15.3(7.7)	0.5(0.7)	65.6(9.7)	64.24(125.79)	0.52(0.03)
LAD-Lasso	1.6(1.7)	12.9(5.4)	0.3(0.5)	9.7(4.9)	4.62(0.30)	0.60(0.05)
RLARS	3.0(1.6)	15.2(8.2)	0.4(0.7)	19.5(6.6)	104.31(331.36)	0.60(0.06)
LTS-MCP	5.8(1.1)	18.0(4.9)	5.3(1.9)	57.6(5.9)	3.76(0.39)	0.81(0.04)
LS-MCP-Hier	4.3(2.2)	204.0(159.3)	1.4(2.6)	81.3(250.6)	60.77(122.85)	0.51(0.05)

Table A.24: Summary results under simulation scenarios with some weak signals. In each cell, mean (sd) based on 200 replicates.

	M:TP	M:FP	I:TP	I:FP	RSSE	PMSE
			D1: $N(0,1)$			
LTS-MCP-Hier	8.0(0.2)	0.4(1.2)	8.0(1.4)	0.6(0.7)	1.54(0.32)	0.92(0.25)
LS-MCP	6.7(0.9)	2.3(3.0)	7.7(1.3)	9.5(10.2)	2.14(0.45)	1.07(0.39)
LAD-Lasso	8.0(0.0)	5.7(3.2)	8.6(1.3)	16.5(8.3)	1.44(0.26)	1.21(0.32)
RLARS	7.8(0.4)	0.2(0.6)	5.0(1.5)	8.2(5.6)	2.43(0.41)	1.45(0.41)
LTS-MCP	6.8(0.8)	6.9(3.7)	7.2(1.4)	25.3(9.1)	2.03(0.46)	0.97(0.22)
LS-MCP-Hier	8.0(0.0)	0.5(1.0)	9.0(1.2)	0.7(0.7)	1.38(0.21)	0.76(0.14)
			D2: $0.9N(0,1)+0.1Cauchy(0,5)$			
LTS-MCP-Hier	8.0(0.1)	0.6(1.4)	8.3(1.4)	0.8(0.9)	1.56(0.26)	0.89(0.20)
LS-MCP	1.8(1.7)	18.9(8.0)	1.5(1.5)	73.9(10.1)	38.28(56.15)	761.65(2767.61)
LAD-Lasso	8.0(0.2)	2.3(1.5)	5.8(1.6)	7.3(2.4)	2.03(0.30)	1.96(0.58)
RLARS	7.6(0.6)	0.8(1.1)	4.3(1.4)	10.3(6.8)	2.71(0.39)	1.82(0.58)
LTS-MCP	6.7(0.9)	8.2(3.8)	7.1(1.4)	31.1(10.3)	2.12(0.49)	1.00(0.27)
LS-MCP-Hier	5.5(1.5)	166.6(124.1)	1.3(1.6)	31.3(79.7)	32.79(45.78)	1446.11(5735.58)
			D3: $0.9N(0,1)+0.1N(20,1)$			
LTS-MCP-Hier	8.0(0.1)	0.4(1.2)	8.6(1.4)	0.7(0.9)	1.48(0.24)	0.82(0.18)
LS-MCP	2.9(1.1)	25.2(5.4)	2.0(1.0)	67.4(5.0)	9.44(0.63)	31.63(6.25)
LAD-Lasso	7.9(0.3)	2.6(1.7)	5.1(1.7)	7.8(2.4)	2.18(0.36)	2.29(0.77)
RLARS	6.5(1.1)	1.4(1.4)	3.0(1.4)	11.6(5.7)	3.31(0.40)	3.04(1.08)
LTS-MCP	6.8(0.9)	6.5(3.2)	7.4(1.2)	28.2(9.1)	2.01(0.48)	0.93(0.20)
LS-MCP-Hier	5.9(1.0)	93.1(6.0)	1.7(1.4)	5.4(4.6)	8.07(0.52)	28.15(5.19)
			D4: $N(0,1)$ and with leverage points			
LTS-MCP-Hier	7.8(0.6)	1.4(2.7)	7.7(2.0)	1.4(1.8)	1.75(0.54)	1.01(0.93)
LS-MCP	1.5(0.9)	22.1(4.8)	2.3(1.4)	69.4(6.4)	6.41(0.85)	14.50(4.92)
LAD-Lasso	4.5(1.2)	4.0(2.2)	3.2(1.8)	14.2(3.5)	3.27(0.32)	6.70(1.77)
RLARS	7.6(0.6)	0.5(0.7)	5.0(1.4)	10.1(6.3)	2.53(0.38)	1.65(0.59)
LTS-MCP	6.5(1.2)	6.7(4.3)	7.2(1.4)	27.0(12.7)	2.12(0.58)	1.30(0.35)
LS-MCP-Hier	5.5(1.3)	27.1(29.1)	3.7(2.4)	5.1(2.7)	3.99(0.85)	7.47(4.51)
			D5: $0.9N(0,1)+0.1Cauchy(0,5)$ and with leverage points			
LTS-MCP-Hier	7.7(0.9)	1.5(2.8)	8.0(2.1)	1.3(1.3)	1.76(0.53)	1.04(1.04)
LS-MCP	0.5(0.8)	17.1(8.9)	1.1(1.1)	69.2(18.5)	32.48(44.17)	600.83(2389.98)
LAD-Lasso	4.5(1.4)	4.0(2.1)	3.5(1.8)	12.6(3.4)	3.28(0.32)	6.54(1.70)
RLARS	7.3(0.7)	1.0(1.1)	4.3(1.7)	11.3(6.7)	2.81(0.39)	1.98(0.63)
LTS-MCP	6.7(0.9)	8.4(3.9)	7.0(1.4)	32.0(10.5)	2.17(0.51)	1.31(0.29)
LS-MCP-Hier	4.4(1.5)	168.7(117.3)	1.0(1.5)	27.6(71.1)	31.14(43.24)	1186.95(4305.08)

Table A.25: Summary results under simulation scenarios where the hierarchy is violated for some interactions. In each cell, mean (sd) based on 200 replicates.

	M:TP	M:FP	I:TP	I:FP	RSSE	PMSE
			D1: $N(0,1)$			
LTS-MCP-Hier	7.8(0.4)	4.0(4.4)	10.5(2.0)	2.5(1.7)	3.46(0.35)	3.90(0.95)
LS-MCP	5.5(1.0)	5.0(4.4)	16.8(1.0)	17.8(9.8)	2.95(0.44)	1.51(0.60)
LAD-Lasso	7.9(0.4)	15.0(7.4)	17.9(2.9)	36.0(12.1)	2.23(0.59)	2.25(1.17)
RLARS	7.3(0.7)	1.0(1.5)	8.9(2.2)	14.4(7.9)	4.02(0.38)	4.29(1.08)
LTS-MCP	6.2(1.1)	7.0(3.0)	16.8(1.3)	26.8(6.9)	2.48(0.57)	1.11(0.35)
LS-MCP-Hier	7.8(0.4)	6.0(5.9)	11.3(1.9)	2.9(1.6)	3.40(0.45)	3.78(1.28)
			D2: $0.9N(0,1)+0.1Cauchy(0,5)$			
LTS-MCP-Hier	7.7(0.5)	6.0(5.9)	10.1(2.0)	2.9(1.6)	3.58(0.34)	4.19(1.02)
LS-MCP	2.2(1.8)	17.0(7.4)	3.5(3.2)	71.4(10.6)	38.37(84.37)	2057.01(11618.56)
LAD-Lasso	7.2(0.7)	2.0(1.5)	8.2(2.7)	7.7(3.1)	3.88(0.36)	6.93(1.99)
RLARS	7.0(0.9)	1.0(1.5)	7.1(2.3)	12.4(7.2)	4.44(0.43)	5.43(1.59)
LTS-MCP	6.1(1.0)	8.0(3.0)	16.2(1.6)	33.2(6.8)	2.78(0.53)	1.39(0.48)
LS-MCP-Hier	5.8(1.4)	110.0(32.6)	2.4(2.6)	34.2(95.9)	29.73(50.07)	1510.65(6857.39)
			D3: $0.9N(0,1)+0.1N(20,1)$			
LTS-MCP-Hier	7.6(0.6)	5.0(4.4)	10.2(2.2)	2.8(1.6)	3.60(0.47)	4.32(1.49)
LS-MCP	2.6(1.1)	22.0(5.2)	4.0(1.6)	66.9(5.3)	10.47(0.68)	38.04(6.53)
LAD-Lasso	6.9(0.9)	2.5(2.2)	6.4(2.2)	8.7(3.1)	4.13(0.34)	8.07(2.19)
RLARS	5.8(1.1)	1.0(1.5)	4.4(1.9)	11.9(5.8)	5.12(0.49)	8.10(2.43)
LTS-MCP	6.2(1.0)	8.0(3.0)	16.6(1.4)	29.2(6.1)	2.61(0.54)	1.18(0.37)
LS-MCP-Hier	6.2(1.0)	96.5(8.2)	2.7(1.6)	7.3(5.7)	9.76(0.64)	42.01(9.31)
			D4: $N(0,1)$ and with leverage points			
LTS-MCP-Hier	7.3(1.1)	7.5(8.2)	9.4(3.2)	2.8(1.8)	3.78(0.76)	5.35(3.54)
LS-MCP	1.1(0.9)	21.0(4.4)	4.9(2.5)	67.0(6.2)	7.90(0.91)	22.21(7.30)
LAD-Lasso	3.8(1.2)	4.0(3.0)	6.0(2.4)	13.0(4.3)	4.56(0.35)	11.86(2.85)
RLARS	6.8(1.0)	0.0(0.0)	8.1(1.9)	12.8(6.6)	4.27(0.42)	4.78(1.44)
LTS-MCP	6.2(1.1)	9.0(3.0)	16.5(1.5)	32.4(6.6)	2.64(0.57)	1.24(0.43)
LS-MCP-Hier	4.8(1.3)	91.0(7.4)	2.2(2.1)	5.0(4.1)	7.55(1.16)	28.00(10.44)
			D5: $0.9N(0,1)+0.1Cauchy(0,5)$ and with leverage points			
LTS-MCP-Hier	7.5(0.9)	8.0(4.4)	9.5(2.5)	3.0(2.0)	3.77(0.71)	5.32(3.72)
LS-MCP	0.6(0.7)	19.0(5.9)	2.0(1.8)	68.5(17.0)	32.80(78.33)	1370.28(8180.45)
LAD-Lasso	3.9(1.3)	4.0(1.5)	5.5(2.1)	12.2(3.5)	4.57(0.33)	12.16(3.24)
RLARS	6.6(1.0)	1.0(1.5)	6.8(2.2)	12.2(6.5)	4.57(0.40)	5.88(1.67)
LTS-MCP	6.2(1.1)	10.0(3.0)	16.4(1.5)	34.4(7.1)	2.68(0.57)	1.28(0.44)
LS-MCP-Hier	4.4(1.6)	113.0(25.2)	1.4(2.1)	31.8(88.0)	28.25(46.51)	1477.08(7265.55)

Table A.26: Analysis of SKCM data: numbers of overlapping interactions (RV-coefficients) identified by different approaches.

Main: G	LTS-MCP-Hier	LS-MCP	LAD-Lasso	RLARS	LTS-MCP	LS-MCP-Hier
LTS-MCP-Hier	43	0(0.58)	1(0.00)	0(0.00)	12(0.33)	22(0.48)
LS-MCP		13	0(0.00)	0(0.00)	0(0.03)	0(0.03)
LAD-Lasso			1	0(0.00)	0(0.00)	1(0.00)
RLARS				0	0(0.00)	0(0.00)
LTS-MCP					50	15(0.98)
LS-MCP-Hier						47
Interaction	LTS-MCP-Hier	LS-MCP	LAD-Lasso	RLARS	LTS-MCP	LS-MCP-Hier
LTS-MCP-Hier	26	0(0.02)	0(0.73)	0(0.28)	3(0.00)	4(0.58)
LS-MCP		72	0(0.02)	1(0.03)	6(0.00)	1(0.02)
LAD-Lasso			25	0(0.48)	2(0.01)	3(0.41)
RLARS				31	1(0.00)	0(0.20)
LTS-MCP					110	4(0.03)
LS-MCP-Hier						24

Table A.27: Analysis of BRCA data: numbers of overlapping interactions (RV-coefficients) identified by different approaches.

Main: G	LTS-MCP-Hier	LS-MCP	LAD-Lasso	RLARS	LTS-MCP	LS-MCP-Hier
LTS-MCP-Hier	32	1(0.27)	5(0.41)	0(0.22)	2(0.37)	14(0.73)
LS-MCP		6	1(0.27)	0(0.16)	0(0.11)	1(0.23)
LAD-Lasso			27	0(0.21)	0(0.33)	3(0.43)
RLARS				12	1(0.22)	0(0.27)
LTS-MCP					17	2(0.47)
LS-MCP-Hier						51
Interaction	LTS-MCP-Hier	LS-MCP	LAD-Lasso	RLARS	LTS-MCP	LS-MCP-Hier
LTS-MCP-Hier	39	1(0.09)	0(0.20)	0(0.15)	0(0.20)	6(0.33)
LS-MCP		17	2(0.19)	0(0.17)	0(0.12)	1(0.15)
LAD-Lasso			36	3(0.26)	0(0.21)	1(0.32)
RLARS				35	0(0.24)	0(0.09)
LTS-MCP					60	0(0.15)
LS-MCP-Hier						21

Appendix B

Chapter 3

Table B.1: Simulation results of S2 under correlation setting C1 and additional information J2. In each cell, mean(sd) based on 200 replicates.

	M:TP	M:FP	I:TP	I:FP	TP60
			AR(0.3)		
Proposed	19.3(1.1)	1.3(1.4)	36.3(2.7)	6.5(7.5)	58.7(2.2)
HierMCP	15.2(1.7)	2.7(2.4)	18.8(2.0)	29.6(2.6)	34.4(1.2)
MCP-LP	7.5(1.8)	0.1(0.3)	17.3(2.8)	28.3(7)	28.2(3.6)
Lasso	1.2(1.6)	0(0)	7.7(3.4)	2.9(2.7)	8.8(3.5)
MA	0(0)	0(0)	0.5(1.1)	3.3(4.5)	4.1(3.0)
			AR(0.5)		
Proposed	19.7(0.7)	0.4(1.3)	37.7(2.7)	4(5)	58.3(3.7)
HierMCP	15(1.6)	0.1(0.2)	20.7(2.4)	30.8(3.8)	34.7(1.2)
MCP-LP	11.4(4)	0(0)	20.3(0.7)	16.7(5.2)	37.3(3.8)
Lasso	0.9(1.3)	0(0)	8.7(3.7)	1.3(1.5)	9.6(3.7)
MA	0(0)	0(0)	0.9(2)	3.2(6.2)	4.0(3.7)
			Band1		
Proposed	18.7(1.1)	2.4(3.9)	34.3(1.4)	9.1(7.3)	56.2(3.9)
HierMCP	14.7(2.1)	4.6(2.5)	17.3(2.8)	29.6(3.3)	31.9(2.1)
MCP-LP	6.5(1.8)	0.1(0.2)	16.4(3.1)	27.5(5.0)	25(4.4)
Lasso	0.9(1.4)	0(0)	5.9(2.9)	2.6(3.2)	6.8(3.2)
MA	0(0)	0(0)	0.9(1.4)	4.5(5.6)	3.1(2.3)
			Band2		
Proposed	19.6(0.8)	0.4(1.0)	37.2(2.6)	5(6.8)	58.6(2.8)
HierMCP	15.5(1.5)	0.4(0.8)	20.8(1.8)	30.8(2.2)	35.3(1.8)
MCP-LP	11.6(3.9)	0(0)	19.9(1.2)	19.1(5.4)	36(4.0)
Lasso	0.9(1.3)	0(0)	8.9(3.1)	1.0(1.0)	9.7(3.6)
MA	0(0)	0.1(0.2)	0.8(1.3)	4.6(8.2)	4.2(3.6)

Table B.2: Simulation results of S2 under correlation setting C2 and additional information J2. In each cell, mean(sd) based on 200 replicates.

	M:TP	M:FP	I:TP	I:FP	TP60
			AR(0.3)		
Proposed	19.1(0.9)	1.4(2.0)	36(2.2)	10.1(12.9)	55.8(5.9)
HierMCP	15.8(1.7)	2(2.2)	19.1(2.7)	30.3(3.3)	34.8(1.9)
MCP-LP	7.3(2.7)	0(0)	18.7(1.6)	26.2(7.1)	30.4(3.1)
Lasso	0.9(1.2)	0(0)	7.9(3.8)	2.6(1.9)	8.8(4.0)
MA	0(0)	0.1(0.3)	0.7(1.4)	3.9(5.1)	4(3.3)
			AR(0.5)		
Proposed	19.8(1.0)	0.3(0.7)	36.5(2.3)	8.4(8.8)	57.1(4.9)
HierMCP	15.1(1.1)	0.2(0.4)	20.5(1.9)	30.4(3.1)	35.5(1.5)
MCP-LP	10(4.6)	0(0)	20.4(0.9)	17.6(4)	37.9(2.5)
Lasso	1.4(1.8)	0(0)	9.7(3.4)	2.4(2.2)	11.1(3.9)
MA	0(0)	0.2(0.7)	1.1(2)	3.5(4.1)	5.3(4.5)
			Band1		
Proposed	18.4(1.1)	1.9(2)	34.6(1.5)	14.4(9.9)	52.5(4.0)
HierMCP	14.7(1.8)	7(4.3)	16.4(3.1)	28.7(3.7)	30.8(3.1)
MCP-LP	5.8(2.4)	0.1(0.3)	14.9(3.4)	32.9(8.3)	25.5(3.8)
Lasso	1.5(1.3)	0(0)	5.5(2.9)	3.7(2.7)	7(3.4)
MA	0(0)	0(0)	0.7(1.1)	3.4(7.1)	3.3(2.6)
			Band2		
Proposed	19.5(1.0)	0.6(1.3)	36.7(2.4)	6.9(6.2)	57.6(3.4)
HierMCP	15.2(1.6)	0.2(0.5)	20.5(1.9)	30.6(2.7)	35.8(1.8)
MCP-LP	8.6(3.2)	0(0)	19.5(1.3)	19.6(4.0)	34.7(2.8)
Lasso	1.3(1.5)	0(0)	6.1(2.5)	1.8(2.4)	7.4(2.7)
MA	0(0)	0(0)	1.4(3.0)	6.8(12.0)	5.5(4.0)

Appendix C

Chapter 4

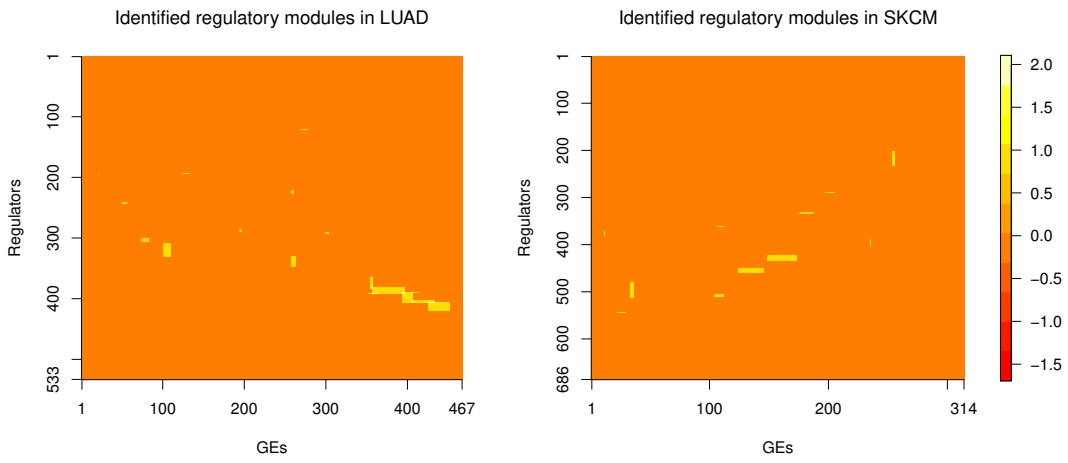


Figure C.1: Data analysis: identified regulatory modules.

Detailed simulation settings

In Step (e) of simulation, the important main molecular effects and M-E interactions are set as follows.

- P1 with a total of 100 important effects under the regulation pattern $\mathbf{\Theta}_1$: The important main molecular effects consist of 15 gene expressions and 20 regulators, among which 30 are involved in one regulatory module and the remaining five are molecular units with individual effects. There are 25 interactions with gene expressions and 40 interactions with regulators, relating to one regulatory module and five individual molecular units.

- P1 with a total of 100 important effects under the regulation pattern $\mathbf{\Theta}_2$: The important main molecular effects consist of 25 gene expressions and 21 regulators, among which 33 are involved in two regulatory modules and the remaining 13 are molecular units with individual effects. There are 36 interactions with gene expressions and 18 interactions with regulators, relating to one regulatory module and nine individual molecular units.

- P2 with a total of 70 important effects under the regulation pattern $\mathbf{\Theta}_1$: The important main molecular effects consist of 15 gene expressions and 20 regulators, among which 30 are involved in one regulatory module and the remaining five are molecular units with individual effects. There are 15 interactions with gene expressions and 20 interactions with regulators, relating to one regulatory module and five individual molecular units.

- P2 with a total of 70 important effects under the regulation pattern $\mathbf{\Theta}_2$: The important main molecular effects consist of 17 gene expressions and 21 regulators, among which 33 are involved in two regulatory modules and the remaining five are molecular units with individual effects. There are 20 interactions with gene expressions and 12 interactions with regulators, relating to one regulatory module and four individual molecular units.

Table C.1: Data analysis: numbers of overlapping main molecular effects and M-E interactions (RV-coefficients) identified by different methods.

LUAD		Proposed	Alt.1	Alt.2	Alt.3	Alt.4
	Proposed	62(1)	7(0.32)	57(0.65)	4(0.24)	9(0.3)
	Alt.1		90(1)	3(0.31)	2(0.08)	33(0.58)
Main effects	Alt.2			140(1)	6(0.24)	11(0.36)
	Alt.3				11(1)	2(0.08)
	Alt.4					66(1)
		Proposed	Alt.1	Alt.2	Alt.3	Alt.4
	Proposed	35(1)	0(0)	4(0.10)	1(0.11)	3(0.08)
	Alt.1		8(1)	1(0.03)	2(0.28)	1(0.12)
Interactions	Alt.2			30(1)	1(0.1)	1(0.07)
	Alt.3				11(1)	6(0.27)
	Alt.4					122(1)
SKCM		Proposed	Alt.1	Alt.2	Alt.3	Alt.4
	Proposed	28(1)	9(0.47)	18(0.62)	9(0.52)	4(0.39)
	Alt.1		22(1)	7(0.39)	12(0.72)	6(0.53)
Main effects	Alt.2			35(1)	7(0.39)	4(0.37)
	Alt.3				13(1)	4(0.51)
	Alt.4					10(1)
		Proposed	Alt.1	Alt.2	Alt.3	Alt.4
	Proposed	12(1)	2(0.36)	0(0.00)	2(0.32)	1(0.11)
	Alt.1		4(1)	0(0.01)	3(0.66)	1(0.19)
Interactions	Alt.2			2(1)	0(0.00)	0(0.01)
	Alt.3				14(1)	1(0.20)
	Alt.4					8(1)

Algorithm 1 Identifying regulatory module

1. Estimate $\hat{\boldsymbol{\Theta}}$ with objective function (4.1) using R package `glmnet`.

2. Initialize $s = 0$ $\boldsymbol{U}^{(s)}$ as the normalized matrix of $\hat{\boldsymbol{\Theta}}$, where $\boldsymbol{U}^{(s)}$ denotes the remaining regulation relationships at iteration s.

3. $s = s + 1$. Apply the sparse 2-means clustering to $\boldsymbol{U}^{(s)}$ based on objective function (4.2), and obtain two clusters $\mathcal{C}_s$ and $\bar{\mathcal{C}}_s$ for regulators as well as the weight vector $\boldsymbol{w}_s$ for gene expressions, using R package `sparcl`.

4. Fix $\mathcal{C}_s$ and $\bar{\mathcal{C}}_s$, and permute the rows of $\boldsymbol{U}^{(s)}$ to calculate weight $w^*_{s,j} = b_j / \sqrt{\sum_{j'} b^2_{j'}}$ with $b_j = \left(\frac{1}{q} \sum_{l=1}^{q} \sum_{l'=1}^{q} d_{l,l',j} - \frac{1}{q_1} \sum_{l,l' \in \mathcal{C}_s} d_{l,l',j} - \frac{1}{q_2} \sum_{l,l' \in \bar{\mathcal{C}}_s} d_{l,l',j} \right)$, under the null hypothesis of no clusters.

5. Repeat Step 4 B times, and then compute $w^0_{s,(j)} = \sum_{k=1}^{B} w^*_{s,(j),k} / B$ with $\boldsymbol{w}^*_{s,(j),k}$ being the jth order statistic of the weights at iteration k of Step 4.

6. Conduct a two-sample Kolmogorov-Smirnov test to compare $\boldsymbol{w}_s$ and $\boldsymbol{w}^0_s$.

7. If the test at Step 6 rejects the null hypothesis at significance level 0.05, then j^* gene expressions with the largest weights are selected, where $j^* = \arg\max_j \left(w_{(p-j+1)} - w^0_{(p-j+1)} \right) - \left(w_{(p-j)} - w^0_{(p-j)} \right)$. Denote the corresponding index set as $\mathcal{D}_s$. Update $\boldsymbol{U}^{(s+1)}$ by excluding the information of the identified module $\{\mathcal{C}_s, \mathcal{D}_s\}$ as

$$U^{(s+1)}_{lj} = \begin{cases} U^{(s)}_{lj} - (\bar{U}^{(s)}_{\mathcal{C}_s,j} - \bar{U}^{(s)}_{\bar{\mathcal{C}}_s,j}), & \text{if } l \in \mathcal{C}_s \text{ and } j \in \mathcal{D}_s \\ U^{(s)}_{lj}, & \text{otherwise,} \end{cases}$$

   where $\bar{U}^{(s)}_{\mathcal{C}_s,j} = \frac{1}{q_1} \sum_{i \in \mathcal{C}_s} U^{(s)}_{ij}$ and $\bar{U}^{(s)}_{\bar{\mathcal{C}}_s,j} = \frac{1}{q_2} \sum_{i \in \bar{\mathcal{C}}_s} U^{(s)}_{ij}$.

8. Repeat Steps 3-7 until the test at Step 6 fails to reject the null hypothesis, and return the final regulatory modules $\{\mathcal{C}_1, \mathcal{D}_1\}, \cdots, \{\mathcal{C}_S, \mathcal{D}_S\}$ with $S+1$ being the termination iteration.

Algorithm 2 M-E interaction analysis with integrated molecular data

1. Initialize $t = 0$, $\boldsymbol{\Phi}^{(0)} = \mathbf{0}$, and $\boldsymbol{res}^{(0)} = \boldsymbol{Y}$, where $\boldsymbol{\Phi}^{(t)}$ and $\boldsymbol{res}^{(t)}$ denote the estimates of $\boldsymbol{\Phi}$ and residual $\boldsymbol{res}$ at iteration t.

2. Update $t = t + 1$. Optimize $Q(\boldsymbol{\Phi})$ by cycling through $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_s$, $\boldsymbol{\gamma}$, $\boldsymbol{\eta}_{sm}$ and $\boldsymbol{\tau}_m$.

   (a) Update $\boldsymbol{\alpha}$ with the least squared solution. Let $\tilde{\boldsymbol{Y}} = \boldsymbol{res}^{(t-1)} + \boldsymbol{E}\boldsymbol{\alpha}^{(t-1)}$, then $\boldsymbol{\alpha}^{(t)} = (\boldsymbol{E}'\boldsymbol{E})^{-1}\boldsymbol{E}'\tilde{\boldsymbol{Y}}$. Update $\boldsymbol{res}^{(t-1)} = \tilde{\boldsymbol{Y}} - \boldsymbol{E}\boldsymbol{\alpha}^{(t)}$.

   (b) For $s = 1, \ldots, S$, update $\boldsymbol{\beta}_s$ sequentially. Let $\tilde{\boldsymbol{Y}} = \boldsymbol{res}^{(t-1)} + \boldsymbol{X}_s\boldsymbol{\beta}_s^{(t-1)} + \sum_{m=1}^{M}(\boldsymbol{E}'_m \odot \boldsymbol{X}'_s)' \left(\boldsymbol{\beta}_s^{(t-1)} * \boldsymbol{\eta}_{sm}^{(t-1)}\right)$ and $\tilde{\boldsymbol{W}}_s = (\tilde{W}_{s1}, \ldots, \tilde{W}_{s,p_s}) = \boldsymbol{X}_s + \sum_m (\boldsymbol{E}'_m \odot \boldsymbol{X}'_s)' \odot \left(\boldsymbol{\eta}_{sm}^{(t-1)}\right)'$. Then, if $||\tilde{\boldsymbol{W}}'_s\tilde{\boldsymbol{Y}}||_2 < \lambda_1\sqrt{p_s}$, update $\boldsymbol{\beta}_s^{(t)} = \mathbf{0}$; Otherwise, update $\beta_{sj}^{(t)} = \underset{\beta_{sj}}{\arg\min}\frac{1}{2}||\tilde{\boldsymbol{Y}} - \sum_{j'\neq j}\tilde{W}_{sj'}\hat{\beta}_{sj'}^{(t)} - \tilde{W}_{sj}\beta_{sj}||_2^2 + \lambda_1\sqrt{p_s}||\beta_{sj}||_2$, for $j = 1, \ldots, p_s$, using the R function `optimize`. Update $\boldsymbol{res}^{(t-1)} = \tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{W}}_s\boldsymbol{\beta}_s^{(t)}$.

   (c) For $d = 1, \ldots, p_z$, update γ_d sequentially. Let $\tilde{\boldsymbol{Y}} = \boldsymbol{res}^{(t-1)} + \boldsymbol{Z}_d\gamma_d^{(t-1)} + \sum_{m=1}^{M}(\boldsymbol{E}_m * \boldsymbol{Z}_d)\left(\gamma_d^{(t-1)}\tau_{md}^{(t-1)}\right)$ and $\tilde{\boldsymbol{W}}_d = \boldsymbol{Z}_d + \sum_m(\boldsymbol{E}_m * \boldsymbol{Z}_d)\tau_{md}^{(t-1)}$, update $\gamma_d^{(t)} = ST\left((\tilde{\boldsymbol{W}}'_d\tilde{\boldsymbol{W}}_d)^{-1}\tilde{\boldsymbol{W}}'_d\tilde{\boldsymbol{Y}}, (\tilde{\boldsymbol{W}}'_d\tilde{\boldsymbol{W}})^{-1}\lambda_2\right)$, where $ST(a, b) = \text{sign}(a)(|a| - b)_+$ is the soft-thresholding operator. Update $\boldsymbol{res}^{(t-1)} = \tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{W}}_d\gamma_d^{(t)}$.

   (d) For $m = 1, \ldots, M$ and $s \in \left\{s : \boldsymbol{\beta}_s^{(t)} \neq 0, s = 1, \cdots, S\right\}$, update $\boldsymbol{\eta}_{sm}$ sequentially. Let $\tilde{\boldsymbol{Y}} = \boldsymbol{res}^{(t-1)} + (\boldsymbol{E}'_m \odot \boldsymbol{X}'_s)' \left(\boldsymbol{\beta}_s^{(t-1)} * \boldsymbol{\eta}_{sm}^{(t-1)}\right)$ and $\tilde{\boldsymbol{W}}_{sm} = (\tilde{W}_{sm,1}, \ldots, \tilde{W}_{sm,p_s}) = (\boldsymbol{E}'_m \odot \boldsymbol{X}'_s)' \odot \left(\boldsymbol{\beta}_s^{(t-1)}\right)'$. Then, if $||\tilde{\boldsymbol{W}}'_{ms}\tilde{\boldsymbol{Y}}|| < \lambda_1\sqrt{p_s}$, update $\boldsymbol{\eta}_{sm}^{(t)} = \mathbf{0}$; otherwise, update $\eta_{sm,j}^{(t)} = \underset{\eta_{sm,j}}{\arg\min}\frac{1}{2}||\tilde{\boldsymbol{Y}} - \sum_{j'\neq j}\tilde{\boldsymbol{W}}_{sm,j'}\eta_{sm,j'}^{(t)} - \tilde{\boldsymbol{W}}_{sm,j}\eta_{sm,j}||_2^2 + \lambda_1\sqrt{p_s}||\eta_{msj}||_2$ for $j = 1, \ldots, p_s$. Update $\boldsymbol{res}^{(t-1)} = \tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{W}}_{sm}\boldsymbol{\eta}_{sm}^{(t)}$.

   (e) For $m = 1, \ldots, M$ and $d \in \left\{d : \gamma_d^{(t)} \neq 0, d = 1, \ldots, p_z\right\}$, update $\hat{\tau}_{md}$ sequentially. Let $\tilde{\boldsymbol{Y}} = \boldsymbol{res}^{(t-1)} + (\boldsymbol{E}_m * \boldsymbol{Z}_d)\left(\gamma_d^{(t-1)}\tau_{md}^{(t-1)}\right)$ and $\tilde{\boldsymbol{W}}_{md} = (\boldsymbol{E}'_m \odot \boldsymbol{Z}'_d)'\gamma_d^{(t-1)}$, then $\tau_{md}^{(t)} = ST\left((\tilde{\boldsymbol{W}}'_{md}\tilde{\boldsymbol{W}}_{md})^{-1}\tilde{\boldsymbol{W}}'_{md}\tilde{\boldsymbol{Y}}, (\tilde{\boldsymbol{W}}'_{md}\tilde{\boldsymbol{W}}_{md})^{-1}\lambda_2\right)$. Update $\boldsymbol{res}^{(t)} = \tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{W}}_{md}\tau_{md}^{(t)}$.

3. Repeat Step 2 until convergence. In our numerical study, convergence is concluded if $\frac{|Q(\hat{\boldsymbol{\Phi}}^{(t-1)}) - Q(\hat{\boldsymbol{\Phi}}^{(t)})|}{|Q(\hat{\boldsymbol{\Phi}}^{(t-1)})|} < 10^{-4}$.

Bibliography

Aalen, O. O. (1988). Heterogeneity in survival analysis. *Statistics in medicine*, 7(11):1121–1137.

Albert, P. S., Ratnasinghe, D., Tangrea, J., and Wacholder, S. (2001). Limitations of the case-only design for identifying gene-environment interactions. *American journal of epidemiology*, 154(8):687–693.

Alfons, A., Croux, C., Gelper, S., et al. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics*, 7(1):226–248.

Ardlie, K. G., Kruglyak, L., and Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 3(4):299–309.

Batchelor, T. T., Betensky, R. A., Esposito, J. M., Pham, L.-D. D., Dorfman, M. V., Piscatelli, N., Jhung, S., Rhee, D., and Louis, D. N. (2004). Age-dependent prognostic effects of genetic alterations in glioblastoma. *Clinical Cancer Research*, 10(1):228–233.

Becker, K. G., Hosack, D. A., Dennis, G., Lempicki, R. A., Bright, T. J., Cheadle, C., and Engel, J. (2003). Pubmatrix: a tool for multiplex literature mining. *BMC bioinformatics*, 4(1):61.

Benzaquen, J., Boutros, J., Marquette, C., Delingette, H., and Hofman, P. (2019). Lung cancer screening, towards a multidimensional approach: why and how? *Cancers*, 11(2):212.

Bien, J., Simon, N., and Tibshirani, R. (2015). Convex hierarchical testing of interactions. *The Annals of Applied Statistics*, pages 27–42.

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

Boolell, V., Alamgeer, M., Watkins, D. N., and Ganju, V. (2015). The evolution of therapies in non-small cell lung cancer. *Cancers*, 7(3):1815–1846.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The annals of applied statistics*, 5(1):232.

Breitling, R., Amtmann, A., and Herzyk, P. (2004). Graph-based iterative group analysis enhances microarray interpretation. *BMC bioinformatics*, 5(1):100.

Breslow, A. (1970). Thickness, cross-sectional areas and depth of invasion in the prognosis of cutaneous melanoma. *Annals of surgery*, 172(5):902.

Bryant, A. and Cerfolio, R. J. (2007). Differences in epidemiology, histology, and survival between cigarette smokers and never-smokers who develop non-small cell lung cancer. *Chest*, 132(1):185–192.

Bychkov, D., Linder, N., Turkki, R., Nordling, S., Kovanen, P. E., Verrill, C., Walliander, M., Lundin, M., Haglund, C., and Lundin, J. (2018). Deep learning based tissue analysis predicts outcome in colorectal cancer. *Scientific Reports*, 8(1):3395.

Chai, H., Zhang, Q., Jiang, Y., Wang, G., Zhang, S., Ahmed, S. E., and Ma, S. (2017). Identifying gene-environment interactions for prognosis using a robust approach. *Econometrics and statistics*, 4:105–120.

Chang, X., Wang, Y., Li, R., and Xu, Z. (2018). Sparse k-means with l_∞/l_1 penalty for high-dimensional data clustering. *Statistica Sinica*, 28(3):1265–1284.

Chatterjee, N. and Wacholder, S. (2009). Invited commentary: efficient testing of gene-environment interaction. *American journal of epidemiology*, 169(2):231–233.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Chen, J., Stampfer, M. J., Hough, H. L., Garcia-Closas, M., Willett, W. C., Hennekens, C. H., Kelsey, K. T., and Hunter, D. J. (1998). A prospective study of n-acetyltransferase genotype, red meat intake, and risk of colorectal cancer. *Cancer research*, 58(15):3307–3311.

Choi, N. H., Li, W., and Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105(489):354–364.

Cordell, H. J. (2009). Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392–404.

Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A. L., Razavian, N., and Tsirigos, A. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24(10):1559.

Dempfle, A., Scherag, A., Hein, R., Beckmann, L., Chang-Claude, J., and Schäfer, H. (2008). Gene–environment interactions for complex traits: definitions, methodological requirements and challenges. *European Journal of Human Genetics*, 16(10):1164–1172.

Dickson, P. V. and Gershenwald, J. E. (2011). Staging and prognosis of cutaneous melanoma. *Surgical Oncology Clinics*, 20(1):1–17.

Fass, L. (2008). Imaging and cancer: a review. *Molecular Oncology*, 2(2):115–152.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Frost, H. R., Shen, L., Saykin, A. J., Williams, S. M., Moore, J. H., and Initiative, A. D. N. (2016). Identifying significant gene-environment interactions using a combination of screening testing and hierarchical false discovery rate control. *Genetic epidemiology*, 40(7):544–557.

Gao, X., Ahmed, S., and Feng, Y. (2017). Post selection shrinkage estimation for high-dimensional data analysis. *Applied Stochastic Models in Business and Industry*, 33(2):97–120.

García-Closas, M. and Lubin, J. H. (1999). Power and sample size calculations in case-control studies of gene-environment interactions: comments on different approaches. *American journal of epidemiology*, 149(8):689–692.

García-Closas, M., Malats, N., Silverman, D., Dosemeci, M., Kogevinas, M., Hein, D. W., Tardón, A., Serra, C., Carrato, A., García-Closas, R., et al. (2005). Nat2 slow acetylation, gstm1 null genotype, and risk of bladder cancer: results from the spanish bladder cancer study and meta-analyses. *The Lancet*, 366(9486):649–659.

Gauderman, W. J. (2002). Sample size requirements for matched case-control studies of gene–environment interaction. *Statistics in medicine*, 21(1):35–50.

Graber, M. L. (2013). The incidence of diagnostic error in medicine. *BMJ quality & safety*, 22(Suppl 2):ii21–ii27.

Green, J., Banks, E., Berrington, A., Darby, S., Deo, H., and Newton, R. (2000). N-acetyltransferase 2 and bladder cancer: an overview and consideration of the evidence for gene–environment interaction. *British journal of cancer*, 83(3):412–417.

Gross, S. M. and Tibshirani, R. (2015). Collaborative regression. *Biostatistics*, 16(2):326–338.

Guerra, R. and Goldstein, D. R. (2009). *Meta-analysis and combining information in genetics and genomics*. CRC Press.

Guo, J., Hu, J., Jing, B.-Y., and Zhang, Z. (2016). Spline-lasso in high-dimensional linear regression. *Journal of the American Statistical Association*, 111(513):288–297.

Gurcan, M. N., Boucheron, L., Can, A., Madabhushi, A., Rajpoot, N., and Yener, B. (2009). Histopathological image analysis: A review. *IEEE Reviews in Biomedical Engineering*, 2:147–171.

Hao, N., Feng, Y., and Zhang, H. H. (2018). Model selection for high-dimensional quadratic regression via regularization. *Journal of the American Statistical Association*, 113(522):615–625.

Hao, N. and Zhang, H. H. (2017). A note on high-dimensional linear regression with interactions. *The American Statistician*, 71(4):291–297.

Hebiri, M., Van De Geer, S., et al. (2011). The smooth-lasso and other $l1+ l2$-penalized methods. *Electronic Journal of Statistics*, 5:1184–1226.

Hein, D. W. (2002). Molecular genetics and function of nat1 and nat2: role in aromatic amine metabolism and carcinogenesis. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 506:65–77.

Helgeson, E. S., Liu, Q., Chen, G., Kosorok, M. R., and Bair, E. (2019). Biclustering via sparse clustering. *Biometrics*.

Higashi, M. K., Veenstra, D. L., Kondo, L. M., Wittkowsky, A. K., Srinouanprachanh, S. L., Farin, F. M., and Rettie, A. E. (2002). Association between cyp2c9 genetic variants and anticoagulation-related outcomes during warfarin therapy. *Jama*, 287(13):1690–1698.

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., and Aerts, H. J. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8):500–510.

Hsu, L., Jiao, S., Dai, J. Y., Hutter, C., Peters, U., and Kooperberg, C. (2012). Powerful cocktail methods for detecting genome-wide gene-environment interaction. *Genetic epidemiology*, 36(3):183–194.

Huang, J. and Ma, S. (2010). Variable selection in the accelerated failure time model via the bridge method. *Lifetime data analysis*, 16(2):176–195.

Huang, J., Ma, S., Li, H., and Zhang, C.-H. (2011a). The sparse laplacian shrinkage estimator for high-dimensional regression. *Annals of statistics*, 39(4):2021.

Huang, J., Ma, S., and Xie, H. (2006). Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics*, 62(3):813–820.

Huang, J., Ma, S., and Xie, H. (2007). Least absolute deviations estimation for the accelerated failure time model. *Statistica Sinica*, pages 1533–1548.

Huang, Y.-T., Lin, X., Liu, Y., Chirieac, L. R., McGovern, R., Wain, J., Heist, R., Skaug, V., Zienolddiny, S., Haugen, A., et al. (2011b). Cigarette smoking increases copy number alterations in nonsmall-cell lung cancer. *Proceedings of the National Academy of Sciences*, 108(39):16345–16350.

Hunter, D. J. (2005). Gene–environment interactions in human diseases. *Nature Reviews Genetics*, 6(4):287–298.

Jacobsen, A. (2017). *cgdsr: R-Based API for Accessing the MSKCC Cancer Genomics Data Server (CGDS)*. R package version 1.2.6.

Jung, S., Marron, J. S., et al. (2009). Pca consistency in high dimension, low sample size context. *The Annals of Statistics*, 37(6B):4104–4130.

Karlsson, A., Ringnér, M., Lauss, M., Botling, J., Micke, P., Planck, M., and Staaf, J. (2014). Genomic and transcriptional alterations in lung adenocarcinoma in relation to smoking history. *Clinical Cancer Research*, 20(18):4912–4924.

Khan, J. A., Van Aelst, S., and Zamar, R. H. (2007). Robust linear model selection based on least angle regression. *Journal of the American Statistical Association*, 102(480):1289–1299.

Khoury, M. J. and Wacholder, S. (2009). Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies: challenges and opportunities. *American journal of epidemiology*, 169(2):227–230.

Kim, G., Lai, C.-Q., Arnett, D. K., Parnell, L. D., Ordovas, J. M., Kim, Y., and Kim, J. (2017). Detection of gene–environment interactions in a family-based population using scad. *Statistics in medicine*, 36(22):3547–3559.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

Koenker, R. and Machado, J. A. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the american statistical association*, 94(448):1296–1310.

Kooperberg, C. and LeBlanc, M. (2008). Increasing the power of identifying gene× gene interactions in genome-wide association studies. *Genetic epidemiology*, 32(3):255–263.

Kristensen, V. N., Lingjærde, O. C., Russnes, H. G., Vollan, H. K. M., Frigessi, A., and Børresen-Dale, A.-L. (2014). Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer*, 14(5):299–313.

Kurnaz, F. S., Hoffmann, I., and Filzmoser, P. (2018). Robust and sparse estimation methods for high-dimensional linear and logistic regression. *Chemometrics and Intelligent Laboratory Systems*, 172:211–222.

Kwak, E. L., Bang, Y.-J., Camidge, D. R., Shaw, A. T., Solomon, B., Maki, R. G., Ou, S.-H. I., Dezube, B. J., Jänne, P. A., Costa, D. B., et al. (2010). Anaplastic lymphoma kinase inhibition in non–small-cell lung cancer. *New England Journal of Medicine*, 363(18):1693–1703.

Lake, S. and Laird, N. (2004). Tests of gene-environment interaction for case-parent triads with general environmental exposures. *Annals of Human Genetics*, 68(1):55–64.

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

Langfelder, P. and Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559.

Lee, S., Liao, Y., Seo, M. H., and Shin, Y. (2018). Oracle estimation of a change point in high-dimensional quantile regression. *Journal of the American Statistical Association*, 113(523):1184–1194.

Li, C. and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences*, 98(1):31–36.

Li, X., Shi, Y., Yin, Z., Xue, X., and Zhou, B. (2014). An eight-mirna signature as a potential biomarker for predicting survival in lung adenocarcinoma. *Journal of translational medicine*, 12(1):159.

Liang, Y., Chai, H., Liu, X.-Y., Xu, Z.-B., Zhang, H., and Leung, K.-S. (2016). Cancer survival analysis using semi-supervised learning method based on cox and aft models with l 1/2 regularization. *BMC medical genomics*, 9(1):11.

Liu, J., Huang, J., Zhang, Y., Lan, Q., Rothman, N., Zheng, T., and Ma, S. (2013). Identification of gene–environment interactions in cancer studies using penalization. *Genomics*, 102(4):189–194.

Luo, X., Zang, X., Yang, L., Huang, J., Liang, F., Rodriguez-Canales, J., Wistuba, I. I., Gazdar, A., Xie, Y., and Xiao, G. (2017). Comprehensive computational pathological image analysis predicts lung cancer prognosis. *Journal of Thoracic Oncology*, 12(3):501–509.

Ma, S., Huang, J., and Song, X. (2011). Integrative analysis and variable selection with multiple high-dimensional data sets. *Biostatistics*, 12(4):763–775.

Ma, S., Li, R., and Tsai, C.-L. (2017). Variable screening via quantile partial correlation. *Journal of the American Statistical Association*, 112(518):650–663.

McAllister, K., Mechanic, L. E., Amos, C., Aschard, H., Blair, I. A., Chatterjee, N., Conti, D., Gauderman, W. J., Hsu, L., Hutter, C. M., et al. (2017). Current challenges and new opportunities for gene-environment interaction studies of complex diseases. *American journal of epidemiology*, 186(7):753–761.

McKinney, B. A., Reif, D. M., Ritchie, M. D., and Moore, J. H. (2006). Machine learning for detecting gene-gene interactions. *Applied bioinformatics*, 5(2):77–88.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Meng, C., Zeleznik, O. A., Thallinger, G. G., Kuster, B., Gholami, A. M., and Culhane, A. C. (2016). Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in bioinformatics*, 17(4):628–641.

Minafra, L., Bravata, V., Cammarata, F. P., Russo, G., Gilardi, M. C., and FORTE, G. I. (2018). Radiation gene-expression signatures in primary breast cancer cells. *Anticancer research*, 38(5):2707–2715.

Mukherjee, B., Ahn, J., Gruber, S. B., Ghosh, M., and Chatterjee, N. (2010). Case–control studies of gene–environment interaction: Bayesian design and analysis. *Biometrics*, 66(3):934–948.

Mukherjee, B. and Chatterjee, N. (2008). Exploiting gene-environment independence for analysis of case–control studies: an empirical bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics*, 64(3):685–694.

Murcray, C. E., Lewinger, J. P., Conti, D. V., Thomas, D. C., and Gauderman, W. J. (2011). Sample size requirements to detect gene-environment interactions in genome-wide association studies. *Genetic epidemiology*, 35(3):201–210.

Murcray, C. E., Lewinger, J. P., and Gauderman, W. J. (2009). Gene-environment interaction in genome-wide association studies. *American journal of epidemiology*, 169(2):219–226.

Network, C. G. A. R. et al. (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543–550.

Nordquist, L. T., Simon, G. R., Cantor, A., Alberts, W. M., and Bepler, G. (2004). Improved survival in never-smokers vs current smokers with primary adenocarcinoma of the lung. *Chest*, 126(2):347–351.

Nöthlings, U., Yamamoto, J. F., Wilkens, L. R., Murphy, S. P., Park, S.-Y., Henderson, B. E., Kolonel, L. N., and Le Marchand, L. (2009). Meat and heterocyclic amine intake,

smoking, nat1 and nat2 polymorphisms, and colorectal cancer risk in the multiethnic cohort study. *Cancer Epidemiology and Prevention Biomarkers*, 18(7):2098–2106.

Osborne, J. W. and Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research, and Evaluation*, 9(1):6.

Paez, J. G., Jänne, P. A., Lee, J. C., Tracy, S., Greulich, H., Gabriel, S., Herman, P., Kaye, F. J., Lindeman, N., Boggon, T. J., et al. (2004). Egfr mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science*, 304(5676):1497–1500.

Pollard, D. (1981). Strong consistency of k-means clustering. *The Annals of Statistics*, pages 135–140.

Ren, J., Zhou, F., Li, X., Chen, Q., Zhang, H., Ma, S., Jiang, Y., and Wu, C. (2019). Semiparametric bayesian variable selection for gene-environment interactions. *Statistics in Medicine*.

Risch, A. and Plass, C. (2008). Lung cancer epigenetics and genetics. *International journal of cancer*, 123(1):1–7.

Ritchie, M. D., Hahn, L. W., and Moore, J. H. (2003). Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genetic epidemiology*, 24(2):150–157.

Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F., and Moore, J. H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics*, 69(1):138–147.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, 12(1):77.

Sharafeldin, N., Slattery, M. L., Liu, Q., Franco-Villalobos, C., Caan, B. J., Potter, J. D., and Yasui, Y. (2015). A candidate-pathway approach to identify gene-environment inter-

actions: analyses of colon cancer risk and survival. *JNCI: Journal of the National Cancer Institute*, 107(9).

Sherman, B. T., Lempicki, R. A., et al. (2009). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1):44.

Shi, X., Liu, J., Huang, J., Zhou, Y., Xie, Y., and Ma, S. (2014). A penalized robust method for identifying gene–environment interactions. *Genetic epidemiology*, 38(3):220–230.

Shi, X., Zhao, Q., Huang, J., Xie, Y., and Ma, S. (2015). Deciphering the associations between gene expression and copy number alteration using a sparse double laplacian shrinkage approach. *Bioinformatics*, 31(24):3977–3983.

Shieh, A. D. and Hung, Y. S. (2009). Detecting outlier samples in microarray data. *Statistical applications in genetics and molecular biology*, 8(1):1–24.

Shipitsin, M., Campbell, L. L., Argani, P., Weremowicz, S., Bloushtain-Qimron, N., Yao, J., Nikolskaya, T., Serebryiskaya, T., Beroukhim, R., Hu, M., et al. (2007). Molecular definition of breast tumor heterogeneity. *Cancer cell*, 11(3):259–273.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245.

Simonds, N. I., Ghazarian, A. A., Pimentel, C. B., Schully, S. D., Ellison, G. L., Gillanders, E. M., and Mechanic, L. E. (2016). Review of the gene-environment interaction literature in cancer: What do we know? *Genetic epidemiology*, 40(5):356–365.

Skipper, P. L., Tannenbaum, S. R., Ross, R. K., and Mimi, C. Y. (2003). Nonsmoking-related arylamine exposure and bladder cancer risk. *Cancer Epidemiology and Prevention Biomarkers*, 12(6):503–507.

Smilde, A. K., Kiers, H. A., Bijlsma, S., Rubingh, C., and Van Erk, M. (2009). Matrix correlations for high-dimensional data: the modified rv-coefficient. *Bioinformatics*, 25(3):401–405.

Soliman, K. (2015). Cellprofiler: novel automated image segmentation procedure for super-resolution microscopy. *Biological Procedures Online*, 17(1):11.

Song, R., Lu, W., Ma, S., and Jessie Jeng, X. (2014). Censored rank independence screening for high-dimensional survival data. *Biometrika*, 101(4):799–814.

Sordillo, J. E., Kelly, R., Bunyavanich, S., McGeachie, M., Qiu, W., Croteau-Chonka, D. C., Soto-Quiros, M., Avila, L., Celedón, J. C., Brehm, J. M., et al. (2015). Genome-wide expression profiles identify potential targets for gene-environment interactions in asthma severity. *Journal of Allergy and Clinical Immunology*, 136(4):885–892.

Sun, D., Li, A., Tang, B., and Wang, M. (2018a). Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome. *Computer methods and programs in biomedicine*, 161:45–53.

Sun, R., Carroll, R. J., Christiani, D. C., and Lin, X. (2018b). Testing for gene–environment interaction under exposure misspecification. *Biometrics*, 74(2):653–662.

Tabesh, A., Teverovskiy, M., Pang, H.-Y., Kumar, V. P., Verbel, D., Kotsianti, A., and Saidi, O. (2007). Multifeature prostate cancer diagnosis and gleason grading of histological images. *IEEE Transactions on Medical Imaging*, 26(10):1366–1378.

Teschendorff, A. E., Yang, Z., Wong, A., Pipinikas, C. P., Jiao, Y., Jones, A., Anjum, S., Hardy, R., Salvesen, H. B., Thirlwell, C., et al. (2015). Correlation of smoking-associated dna methylation changes in buccal cells with dna methylation changes in epithelial cancer. *JAMA oncology*, 1(4):476–485.

Tharmaratnam, K., Claeskens, G., Croux, C., and Salibián-Barrera, M. (2010). S-estimation for penalized regression splines. *Journal of Computational and Graphical Statistics*, 19(3):609–625.

Thomas, D. (2010). Gene–environment-wide association studies: emerging approaches. *Nature Reviews Genetics*, 11(4):259–272.

Thrall, J. H., Li, X., Li, Q., Cruz, C., Do, S., Dreyer, K., and Brink, J. (2018). Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. *Journal of the American College of Radiology*, 15(3):504–508.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

Travis, R. C., Reeves, G. K., Green, J., Bull, D., Tipper, S. J., Baker, K., Beral, V., Peto, R., Bell, J., Zelenika, D., et al. (2010). Gene–environment interactions in 7610 women with breast cancer: prospective evidence from the million women study. *The Lancet*, 375(9732):2143–2151.

Tzeng, J.-Y., Zhang, D., Pongpanich, M., Smith, C., McCarthy, M. I., Sale, M. M., Worrall, B. B., Hsu, F.-C., Thomas, D. C., and Sullivan, P. F. (2011). Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. *The American Journal of Human Genetics*, 89(2):277–288.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L. (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117.

Wang, G., Zhao, Y., Zhang, Q., Zang, Y., Zang, S., and Ma, S. (2017). Identifying gene–environment interactions associated with prognosis using penalized quantile regression. In *Big and Complex Data Analysis*, pages 347–367. Springer.

Wang, H., Xing, F., Su, H., Stromberg, A., and Yang, L. (2014). Novel image markers for non-small cell lung cancer classification and survival prediction. *BMC Bioinformatics*, 15(1):310.

Wang, H. J., Stefanski, L. A., and Zhu, Z. (2012a). Corrected-loss estimation for quantile regression with covariate measurement errors. *Biometrika*, 99(2):405–421.

Wang, H. J. and Wang, L. (2009). Locally weighted censored quantile regression. *Journal of the American Statistical Association*, 104(487):1117–1128.

Wang, W., Baladandayuthapani, V., Morris, J. S., Broom, B. M., Manyam, G., and Do, K.-A. (2012b). ibag: integrative bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*, 29(2):149–159.

Wang, X., Xu, Y., and Ma, S. (2019). Identifying gene-environment interactions incorporating prior information. *Statistics in medicine*, 38(9):1620–1633.

Wang, Y.-G. and Zhu, M. (2006). Rank-based regression for analysis of repeated measures. *Biometrika*, 93(2):459–464.

Westcott, P. M., Halliwill, K. D., To, M. D., Rashid, M., Rust, A. G., Keane, T. M., Delrosario, R., Jen, K.-Y., Gurley, K. E., Kemp, C. J., et al. (2015). The mutational landscapes of genetic and chemical models of kras-driven lung cancer. *Nature*, 517(7535):489–492.

Wilcox, R. R. (2011). *Introduction to robust estimation and hypothesis testing*. Academic Press.

Witten, D. M. and Tibshirani, R. (2010). Survival analysis with high-dimensional covariates. *Statistical Methods in Medical Research*, 19(1):29–51.

Wu, C., Jiang, Y., Ren, J., Cui, Y., and Ma, S. (2018). Dissecting gene-environment interactions: A penalized robust approach accounting for hierarchical structures. *Statistics in medicine*, 37(3):437–456.

Wu, C. and Ma, S. (2015). A selective review of robust variable selection with applications in bioinformatics. *Briefings in bioinformatics*, 16(5):873–883.

Wu, C., Shi, X., Cui, Y., and Ma, S. (2015). A penalized robust semiparametric approach for gene–environment interactions. *Statistics in medicine*, 34(30):4016–4030.

Wu, C., Zhou, F., Ren, J., Li, X., Jiang, Y., and Ma, S. (2019a). A selective review of multi-level omics data integration using variable selection. *High-throughput*, 8(1):4.

Wu, M. and Ma, S. (2019). Robust genetic interaction analysis. *Briefings in bioinformatics*, 20(2):624–637.

Wu, M., Zang, Y., Zhang, S., Huang, J., and Ma, S. (2017). Accommodating missingness in environmental measurements in gene-environment interaction analysis. *Genetic epidemiology*, 41(6):523–554.

Wu, M., Zhang, Q., and Ma, S. (2019b). Structured gene-environment interaction analysis. *Biometrics*.

Xu, Y., Wu, M., Zhang, Q., and Ma, S. (2019). Robust identification of gene-environment interactions for prognosis using a quantile partial correlation approach. *Genomics*, 111(5):1115–1123.

Young, R., Hopkins, R., and Eaton, T. (2007). Forced expiratory volume in one second: not just a lung function test but a marker of premature death from all causes. *European Respiratory Journal*, 30(4):616–622.

Yu, K., Wacholder, S., Wheeler, W., Wang, Z., Caporaso, N., Landi, M. T., and Liang, F. (2012). A flexible bayesian model for studying gene–environment interaction. *PLoS genetics*, 8(1).

Yu, K.-H., Berry, G. J., Rubin, D. L., Re, C., Altman, R. B., and Snyder, M. (2017). Association of omics features with histopathology patterns in lung adenocarcinoma. *Cell systems*, 5(6):620–627.

Yu, K.-H., Zhang, C., Berry, G. J., Altman, R. B., Ré, C., Rubin, D. L., and Snyder, M. (2016). Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature Communications*, 7:12474.

Yuan, Y., Failmezger, H., Rueda, O. M., Ali, H. R., Gräf, S., Chin, S.-F., Schwarz, R. F., Curtis, C., Dunning, M. J., Bardwell, H., et al. (2012). Quantitative image analysis of

cellular heterogeneity in breast tumors complements genomic profiling. *Science Translational Medicine*, 4(157):157ra143.

Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T., de Bakker, P. I., Abecasis, G. R., Almgren, P., Andersen, G., et al. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature genetics*, 40(5):638.

Zhang, C.-H. et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942.

Zhang, L. and Zhang, S. (2019). Learning common and specific patterns from data of multiple interrelated biological scenarios with matrix factorization. *Nucleic acids research*, 47(13):6606–6617.

Zhang, P., Lewinger, J. P., Conti, D., Morrison, J. L., and Gauderman, W. J. (2016). Detecting gene–environment interactions for a quantitative trait in a genome-wide association study. *Genetic epidemiology*, 40(5):394–403.

Zhang, S., Xue, Y., Zhang, Q., Ma, C., Wu, M., and Ma, S. (2019). Identification of gene–environment interactions with marginal penalization. *Genetic epidemiology*.

Zhang, Y., Dai, Y., Zheng, T., and Ma, S. (2011). Risk factors of non-hodgkin's lymphoma. *Expert Opinion on Medical Diagnostics*, 5(6):539–550.

Zhong, T., Wu, M., and Ma, S. (2019). Examination of independent prognostic power of gene expressions and histopathological imaging features in cancer. *Cancers*, 11(3):361.

Zhu, R., Zhao, H., and Ma, S. (2014). Identifying gene–environment and gene–gene interactions using a progressive penalization approach. *Genetic epidemiology*, 38(4):353–368.

Zhu, R., Zhao, Q., Zhao, H., and Ma, S. (2016a). Integrating multidimensional omics data for cancer outcome. *Biostatistics*, 17(4):605–618.

Zhu, X., Yao, J., Luo, X., Xiao, G., Xie, Y., Gazdar, A., and Huang, J. (2016b). Lung cancer survival prediction from pathological images and genetic data: An integration study. In

2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), pages 1173–1176. IEEE.

Zhu, X., Yao, J., Zhu, F., and Huang, J. (2017). Wsisa: Making survival prediction from whole slide histopathological images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7234–7242.