Genomics 111 (2019) 1115–1123

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Original Article

Robust identification of gene-environment interactions for prognosis using a quantile partial correlation approach

Yaqing Xu[a,1], Mengyun Wu[b,a,1], Qingzhao Zhang[c], Shuangge Ma[a,*]

[a] Department of Biostatistics, Yale University, United States
[b] School of Statistics and Management, Shanghai University of Finance and Economics, China
[c] School of Economics and Wang Yanan Institute for Studies in Economics, Xiamen University, China

ARTICLE INFO

Keywords:
Robust G-E interaction
Prognosis
Censored quantile partial correlation

ABSTRACT

Gene-environment (G-E) interactions have important implications for the etiology and progression of many complex diseases. Compared to continuous markers and categorical disease status, prognosis has been less investigated, with the additional challenges brought by the unique characteristics of survival outcomes. Most of the existing G-E interaction approaches for prognosis data share the limitation that they cannot accommodate long-tailed or contaminated outcomes. In this study, for prognosis data, we develop a robust G-E interaction identification approach using the censored quantile partial correlation (CQPCorr) technique. The proposed approach is built on the quantile regression technique (and hence has a solid statistical basis), uses weights to easily accommodate censoring, and adopts partial correlation to identify important interactions while properly controlling for the main genetic and environmental effects. In simulation, it outperforms multiple competitors with more accurate identification. In the analysis of TCGA data on lung cancer and melanoma, biologically sensible findings different from using the alternatives are made.

1. Introduction

For many complex diseases, gene-environment (G-E) interactions have important implications for etiology, progression, and response to treatment beyond the main genetic (G) and environmental (E) effects. Many statistical approaches have been developed for detecting important G-E interactions, especially for categorical responses such as disease status. We refer to [1–3] for a survey. Recent studies have also 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 [4]. 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 [5, 6], and a few others.

For the identification of important G-E interactions, there are two generic paradigms. The first paradigm conducts marginal analysis and analyzes one or a small number of genes at a time. The second conducts joint analysis and includes a large number of genes in a single model. Both types of analysis have been extensively conducted, with marginal analysis perhaps being more popular. Comparatively, marginal analysis is computationally simpler, and the results are more stable. In this article, we conduct marginal analysis and briefly discuss the possibility of extending to joint analysis. In the literature, the commonest marginal analysis strategy proceeds as follows. For each gene, fit a model consisting of one E factor (or a few E factors), the gene itself, and its interaction with the E factor. As the model is low-dimensional, standard, especially likelihood-based, estimations are conducted. This model fitting is cycled through all genes and E factors, and the *p*-values for interactions (and main effects) can be obtained. Important interactions can be identified based on the *p*-values. With a prognosis outcome, popular models include the accelerated failure time (AFT) model with weighted least squared estimation [7], Cox model with partial likelihood estimation, and others. A common limitation shared by most of the existing studies is that they adopt non-robust estimations and cannot accommodate long-tailed/contaminated prognosis data.

In practical genetic studies, long-tailed distributions and contamination in prognosis 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

[*] Corresponding author.
E-mail address: shuangge.ma@yale.edu (S. Ma).
[1] Equal contributions.

https://doi.org/10.1016/j.ygeno.2018.07.006
Received 3 April 2018; Received in revised form 23 June 2018; Accepted 5 July 2018
Available online 17 July 2018
0888-7543/

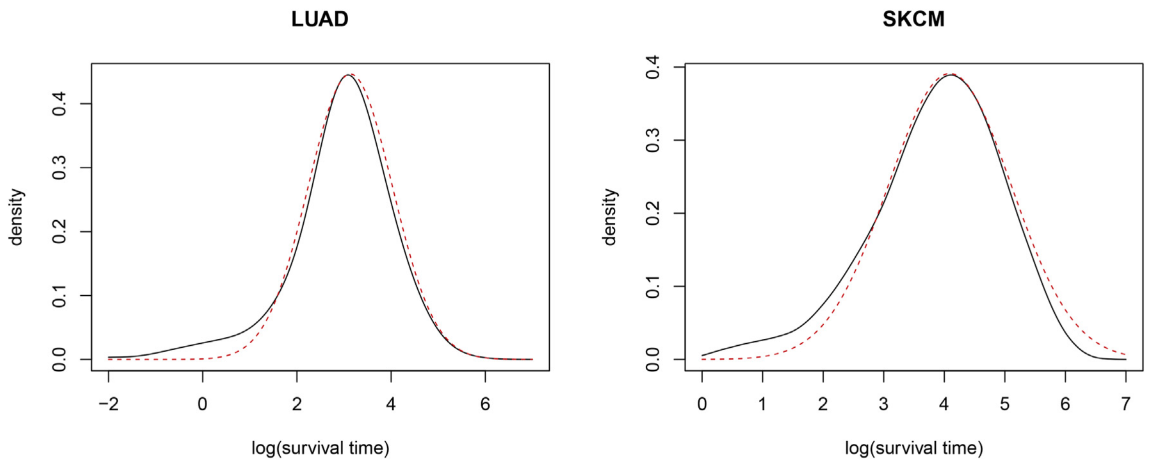

Fig. 1. Analysis of the LUAD and SKCM data: the empirical densities of log(survival time) (solid line) and best-fitted Normal densities (dashed line).

been observed in quite a few studies. In addition, human errors (for example, mistakes in death records) can also cause long-tailed distributions and contamination. For relevant discussions, refer to [8, 9]. As demonstrating examples, consider the LUAD (lung adenocarcinoma) and SKCM (cutaneous melanoma) data collected by TCGA (The Cancer Genome Atlas). 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 study, one has survival time 238.11 months, while the rest 261 have survival times ranging from 0.13 to 129.43 months. For the 225 SKCM subjects, three have survival times 241.20, 268.53, and 339.88 months, while the rest 222 have survival times ranging from 2.04 to 228.42 months. In Fig. 1, we present the empirical densities of the log survival time as well as the best-fitted Normal densities. Compared to Normal, we observe 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, a robust censored quantile regression (CQR) approach has been proposed in [10], which uses a recursive weighting strategy and a generalized Kaplan-Merier (KM) estimator. In [11], a robust least absolute deviation estimation (KMW-LAD) has been developed based on the AFT model and KM weights. Other examples include the rank-based regression [12], S-estimation [13], and others. Overall, development and implementation in G-E interaction analysis with prognosis data are still much limited.

In this study, 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 potentially extended to the analysis of categorical and continuous data. 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 well developed for low-dimensional data [14], and its asymptotic distribution, robustness, and statistical inference have been well established [15]. Compared to non-robust for example 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 adopted for high-dimensional main effect analysis, and shown to have good properties, including consistency, asymptotic normality, and others [16, 17]. 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. Methods

2.1. Modeling

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-dimensional vectors of E and G variables. To study the interaction between the kth E factor and jth gene, consider the model

$$T_i = a_{kj} + \alpha_{kj} X_{ik} + \beta_{kj} Z_{ij} + \theta_{kj} X_{ik} Z_{ij} + \epsilon_i, \tag{1}$$

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, one G factor and their interaction are considered. This strategy has been commonly adopted in the literature. See for example [18, 19]. The proposed approach can straightforwardly accommodate multiple E factors, one G factor and their interactions in a single model. In practice, right censoring is usually present. For subject i, denote C_i as the transformed censoring time, then we observe $Y_i = \ min\,(T_i, C_i)$ and $\delta_i = I(T_i \le C_i)$.

2.2. The CQPCorr approach

Denote X_k, Z_j and T as the random variables corresponding to the kth E factor, jth G factor 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} [5]. Significantly different from the existing studies, we propose quantifying the importance of interaction X_kZ_j using the quantile partial correlation defined as

$$\mathrm{qpcorr}_\tau(kj) = \frac{\mathrm{cov}\{\psi_\tau(T-\eta_0^0-\eta_1^0X_k-\eta_2^0Z_j)X_kZ_j-\gamma_0^0-\gamma_1^0X_k-\gamma_2^0Z_j\}}{\sqrt{\mathrm{var}\{\psi_\tau(T-\eta_0^0-\eta_1^0X_k-\eta_2^0Z_j)\}\,\mathrm{var}(X_kZ_j-\gamma_0^0-\gamma_1^0X_k-\gamma_2^0Z_j)}}. \tag{2}$$

Here for a quantile $0 < \tau < 1$, $\psi_\tau(u) = \tau - \mathbf{1}(u < 0)$ and $\rho_\tau(u) = u\psi_\tau(u)$. $(\eta_0^0, \eta_1^0, \eta_2^0) = \mathrm{argmin}\,\mathbb{E}[\rho_\tau(T - \eta_0 - \eta_1X_k - \eta_2Z_j)]$ and $(\gamma_0^0, \gamma_1^0, \gamma_2^0) = \mathrm{argmin}\,\mathbb{E}[(X_kZ_j - \gamma_0 - \gamma_1X_k - \gamma_2Z_j)^2]$. $\mathbb{E}$ is the expectation function with respect to the random variables X_k, Z_j and T. Note that η_0, η_1, η_2, γ_0, γ_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 (2), the main effects of G and E variables are first removed from T and X_kZ_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 and shown to be competitive [20]. 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. 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 response on the main effects, which corresponds to the first term in the numerator of (2). Specifically, $(\eta_0{}^0, \eta_1{}^0, \eta_2{}^0)$ is estimated as

$$(\hat{\eta}_0, \hat{\eta}_1, \hat{\eta}_2) = \mathrm{argmin} \sum_{i=1}^{n} [w_i\rho_\tau(Y_i - \eta_0 - \eta_1X_{ik} - \eta_2Z_{ij}) + (1 - w_i)\rho_\tau(Y^{+\infty} - \eta_0 - \eta_1X_{ik} - \eta_2Z_{ij})].$$

$Y^{+\infty}$ is a fixed value which 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 [10], w_i is defined for a censored observation as

$$w_i = \frac{\tau - F(C_i \mid X_{ik}, Z_{ij})}{1 - F(C_i \mid X_{ik}, Z_{ij})} \tag{3}$$

if $F(C_i|X_{ik}, Z_{ij}) < \tau$, where $F(t|X_{ik}, Z_{ij})$ is the conditional cumulative distribution function of the survival time given the covariates. For better computational feasibility, we approximate $F(t|X_{ik}, Z_{ij})$ 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 $(\gamma_0{}^0, \gamma_1{}^0, \gamma_2{}^0)$ as

$$(\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2) = \mathrm{argmin} \sum_{i=1}^{n} (X_{ik}Z_{ij} - \gamma_0 - \gamma_1X_{ik} - \gamma_2Z_{ij})^2.$$

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_kZ_j, the censored quantile partial correlation is defined as

$$\mathrm{cqpcorr}_\tau(k,j) = \frac{n^{-1}\sum_{i=1}^{n} [\tau - w_i\mathbf{1}(r_i^{(1)}(k,j) < 0)]r_i^{(2)}(k,j)}{\sqrt{(\overline{w^2}\tau - \overline{w}^2\tau^2)}\sqrt{\frac{1}{n}\sum_{i=1}^{n} (r_i^{(2)}(k,j))^2}}, \tag{4}$$

where

$$r_i^{(1)}(kj) = Y_i - \hat{\eta}_0 - \hat{\eta}_1X_{ik} - \hat{\eta}_2Z_{ij},$$
$$r_i^{(2)}(kj) = X_{ik}Z_{ij} - \hat{\gamma}_0 - \hat{\gamma}_1X_{ik} - \hat{\gamma}_2Z_{ij},$$

$$\overline{w} = n^{-1}\sum_i w_i, \quad \text{and} \quad \overline{w^2} = n^{-1}\sum_i w_i^2.$$

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

After applying the above procedure to all G and E factors, important interactions can be identified in at least two ways. First, the CQPCorr values can be directly compared, and the interactions that have larger absolute CQPCorr values are concluded as more important. Second, when more rigorous results are desired, the following inference can be conducted. *P*-values of the CQPCorrs can be obtained using a permutation approach, which has been a popular choice in the literature [21]. Then, the interactions with smaller *p*-values are regarded as more important. In our numerical studies, to be more rigorous, we adopt the second strategy. To more clearly demonstrate the operation of the permutation process, we provide the permutation distributions under the null in Fig. A1 (Appendix). With the null distribution, one possibility is to fit for example a parametric distribution and compute the p-value analytically. In our numerical analysis, to generate more precise *p*-values (especially for limited sample data), we use the empirical p-value directly. It is noted that it may also be possible to apply alternative inference procedures, for example based on bootstrap [22, 23].

Remarks Advancing from the existing quantile partial correlation studies, the proposed approach introduces weights to accommodate censoring. In survival analysis, there are multiple ways to estimate $F(t|X_{ik}, Z_{ij})$ in (3) to accommodate censoring. Both (parametric, semi-parametric) model-based and nonparametric approaches are available. We adopt the KM based approach as it is computationally simpler and has been commonly adopted in the literature. It also has the advantage of making no assumption on the underlying data distribution, 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.

2.3. Computation

A significant advantage of the proposed approach is that it is computationally much feasible. Step I conducts standard quantile regression and can be realized using the R function *rq*. Step II is a linear regression and can be realized using the R function *lm*. The last step includes a straightforward calculation and does not demand any special function/algorithm. As marginal analysis is conducted, to reduce computational time, the proposed procedure can be realized in a highly parallel manner. In addition, the computation of p-values via permutation can also be realized in a parallel manner. To facilitate data analysis and applications beyond this study, we have developed R code and made it publicly available at www.github.com/shuanggema.

2.4. Toy examples

We further consider two toy examples with 100 G factors to investigate the operating characteristics of the proposed approach. Data are simulated under Scenarios C3 and C4 with the AR correlation structure ($\rho = 0.5$) and Error 2, but with a lower dimensionality (see Section 3 for details). Under both scenarios, the relative magnitudes of interactions to main effects are small, making the identification of interactions challenging. We conduct analysis using the proposed approach, censored quantile correlation (CQCorr), censored quantile regression (CQR), and least absolute deviation estimation based on the AFT model and KM weights (KMW-LAD). CQCorr is the one-step counterpart of the proposed approach, where the correlation between the interaction and response is directly computed without conducting Steps I and II of the proposed approach. The comparison with CQCorr can in a relatively direct way establish the merit of the proposed Steps I and II which remove the effects of main E and G factors. CQR and KMW-LAD analyze each interaction as well as its corresponding main effects using the regression model (1), and the analysis framework is different from the proposed correlation-based. All four approaches are robust. In Fig. A2 (Appendix), we present the true values of θ_{kj}'s, together with the average estimated correlations using CQPCorr and CQCorr, and the average estimated regression coefficients using CQR and KMW-LAD over 100 replicates. Compared to CQCorr, the proposed approach is able to identify important interactions more accurately, which provides a strong support to the proposed Steps I and II. With CQR and KMW-LAD, differences across the estimates are also observed, however, not as distinct as the proposed approach. The superior performance of the proposed approach over CQR and KMW-LAD may result from removing main effects in Steps I and II as well as the censored quantile partial correlation framework in Step III. More conclusive results are presented in the next section.

3. Simulation

Simulation is conducted to gauge performance of the proposed approach and compare with competitors. For all simulated data, we set $n = 200$, $p = 1000$, and $q = 5$. There are thus a total of 5000 interactions and 1005 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 auto-regressive (AR) correlation structure, where the jth and lth G variables have correlation coefficient $\rho^{|j-l|}$. Two levels of correlation with $\rho = 0.5$ and 0.3 are examined. (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 AFT model,

$$Y = \sum_{k=1}^{q} \alpha_k X_k + \sum_{j=1}^{p} \beta_j Z_j + \sum_{k=1}^{q} \sum_{j=1}^{p} \theta_{kj} X_k Z_j + \varepsilon, \tag{5}$$

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". We have verified that the interactions and main effects important in this joint model are also important in a marginal sense (see Appendix for details). 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 important G-E interactions together with two main E effects and five main G effects. Although the proposed approach focuses on interaction identification, the main effects are assumed to make the simulated dataset closer to practical data. There are two types of important interactions. The first type includes ten interactions (θ_{kj}, $k = 1$, 2 and $j = 1, \cdots, 5$) with both main E (α_1 and α_2) and G ($\beta_j, j = 1, \cdots, 5$) effects. The second type includes six interactions (θ_{kj}, $k = 3{,}4{,}5$ and $j = 6$, 7) without main effects, which violates the "main effects, interactions" hierarchy. Five specific scenarios are considered.

C1 has $\theta_{kj} = 2$, $\alpha_k = 1$, $\beta_j = 1$ for $k = 1$, 2 and $j = 1, \cdots, 5$, and $\theta_{kj} = 1$ for $k = 3{,}4{,}5$ and $j = 6$, 7. All other coefficients are 0. Under this scenario, the first type interactions are stronger than the corresponding main effects.

C2 is the same as C1 except that the first type interactions and the corresponding main effects are at the same level. Specifically, $\theta_{kj} = \alpha_k = \beta_j = 1.5$ for $k = 1$, 2 and $j = 1, \cdots, 5$.

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

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

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

We also examine some other settings with a larger sample size, binary E factors, a banded correlation structure and a higher censoring rate (see Section 3.1 and Appendix for details), covering a wide spectrum of settings.

3.1. Comparison with the alternative approaches

Besides the proposed approach, we also consider four alternatives with the same covariate effects as in (1), including the AFT model, Cox model, CQR, and KMW-LAD. As introduced in Section 1, AFT and Cox models are perhaps the most popular approaches for analyzing prognosis data, but without the capacity of accommodating long-tailed distributions and contamination. Note that our simulation is based on the AFT model, and so the Cox model is mis-specified. Due to its popularity and satisfactory performance, the Cox model has been adopted as an alternative approach in many published studies [24, 25]. Thus, we also include the Cox model for comparison. CQR and KMW-LAD are also robust. Different from the proposed three-step correlation-based approach, they analyze each interaction and its corresponding main effects under the one-step regression framework. For the proposed approach and four alternatives, p-values are computed and used to rank and identify interactions. We note that there are other G-E interaction analysis methods that are potentially applicable to the simulated data. The above four approaches 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 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 [26]; (d) TP.FDR, which is the number of true positives when the number of important interactions is selected using the false discovery rate (FDR) approach with target FDR = 0.1; (e) FP.FDR, which is the corresponding number of false positives; and (f) E.FDR, which is the estimated FDR. All five measures have been adopted in multiple publications.

Under each setting, we simulate 200 replicates. Summary results for

Table 1
Simulation results for Scenario C1 with the AR correlation structure. In each cell, mean (sd) based on 200 replicates.

	Error	Approach	TP20	TP40	pAUC	TP.FDR	FP.FDR	E.FDR
ρ = 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)
ρ = 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
Simulation results for Scenario C2 with the AR correlation structure. In each cell, mean (sd) based on 200 replicates.

	Error	Approach	TP20	TP40	pAUC	TP.FDR	FP.FDR	E.FDR
ρ = 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)
ρ = 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)

Scenarios C1 and C2 are presented in Tables 1 and 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 with $\rho = 0.3$ and Error 1, the proposed approach selects on average 8.1 true nonzero interactions when the model size is 20, while CQR and KMW-LAD select 3.0 and 6.2 on average. When there are strong correlations which are 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 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 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). We also examine an example of the partial ROC curves in Fig. A4 (Appendix) under Scenario C1 with $\rho = 0.3$ and Error 3. It is shown that the solid line representing the proposed approach is superior to the others.

With a target FDR of 0.1, it can be seen that the proposed approach performs better in achieving the nominal FDR control and has the smallest estimated FDR under most settings. Except for KMW-LAD, the alternatives do not have a reasonable FDR control. For example, in Table 1 with $\rho = 0.3$ and Error 1, the proposed approach has E.FDR = 0.11, compared to 0.78 (AFT), 0.49 (Cox), 0.93 (CQR), and 0.12 (KMW-LAD). Under the settings with $\rho = 0.3$, the values of TP.FDR with the proposed approach are relatively small which is likely to be caused by the limited sample size. We further examine the results for Scenario C1 with $\rho = 0.3$ and various sample sizes in Tables A1-A3 (Appendix). With a larger sample size, the proposed approach is able to identify the 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 data under Scenarios C3-C5 with $\rho = 0.5$. Summary results are provided in Tables A4-A6 (Appendix). It can be seen that all approaches perform slightly worse under these three scenarios compared to Scenario C1. This may due to that the relative magnitudes of interactions to main effects under Scenarios C3 and C4 are smaller, and the interactions and their corresponding main effects have different directions under Scenario C5. Similar to under the previous simulation scenarios, the proposed approach performs better than or comparable to the alternatives. For example in Table A5 with Error 2 (Scenario C4), the proposed approach has TP20 = 7.6, compared to 1.2 (AFT), 4.2 (Cox), 3.4 (CQR), and 7.2 (KMW-LAD). For Scenarios C1 and C2, we also examine other settings which have G factors with the banded correlation structure, E factors with binary measurements, and a higher censoring rate (35%). Detailed results are provided in Appendix. Similar patterns are observed for the G factors with the banded correlation structure. Performance of all approaches deteriorates when the datasets have 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 data under Scenario C1 with $\rho = 0.5$ using the proposed approach and CQR with various values of τ, and present the summary results in Table A15 (Appendix). The proposed approach can achieve favorable performance with multiple quantiles.

3.2. 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 10,000 permutations, the analysis can be accomplished within 10 min using 100 parallel jobs on a cluster (Intel Xeon CPU E5-2620 v3 at 2.40GHz). A higher degree of parallel computing can further reduce computer time.

4. Data analysis

TCGA is a recent collective effort organized by the NCI. For multiple cancer types, comprehensive data collection has been conducted, generating clinical, environmental, and genetic data. With a high quality, TCGA provides an ideal testbed. We analyze TCGA data on lung adenocarcimona (LUAD) and cutaneous melanoma (SKCM). 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*.

4.1. Analysis of LUAD data

We focus on primary tumor samples of the Whites. The 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 to be potentially associated with lung cancer prognosis [27]. 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 2000 genes with the largest variances across all samples for downstream analyses.

When applying the proposed approach, we compute *p*-values based on 10,000 permutations and use the FDR 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 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 being 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 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 A16 (Appendix). When evaluating the differences in findings, we use both the simple numbers of findings as well as the RV-coefficients [28], 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

Table 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

the authors. Table A16 suggests that although there are overlapping identifications, the proposed approach identifies a different set of interactions. As the numbers of interactions 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 A16 (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) [29], 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 are moderate, which has also been observed in the literature. This may due to the complex correlation structure, low signal-to-noise ratio, high censoring rate, small

Table 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				

(*continued on next page*)

Table 4 (*continued*)

	Breslow thickness	Clark level	Age	Stage	Gender
WDR4	−0.209				
ZFP41	−0.213				
ZNF671		−0.239			
ZUFSP	0.198				

sample size, and other factors. However, the proposed approach still has better stability, which provides support to its superiority.

4.2. Analysis of SKCM data

We focus on metastatic samples of the Whites. Data are available for 225 subjects. The 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, 2000 gene expressions are selected for downstream analysis.

The proposed approach identifies 80 G-E interactions with the FDR control. Details are presented in Table 4. Most of the identified interactions are with Breslow thickness and Clark level, which are the most important prognostic parameters in evaluating primary tumors [30]. Published studies suggest potentially important 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 lower 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 higher 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 A16 (Appendix). Both the FDR control results and (roughly) top forty lists suggest 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).

5. Conclusions

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 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 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. It 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 approach may not respect the “main effects, interactions” hierarchy, which has been stressed in some recent studies [31, 32]. With hierarchy, an interaction can only be identified if the corresponding main effects are also identified. In (4), when the main E and G factors are not associated with the response, the estimated $\hat{\eta}_0$, $\hat{\eta}_1$ and $\hat{\eta}_2$ in $r_i^{(1)}(k,j)$ can be close to zero. Then, no information is removed from the response, and the proposed CQPCorr can still work. Thus, the identified interactions do not necessarily have corresponding main effects. As our main interest is to identify interactions, no specific attention is paid to the identification 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 the least squared regression, as it is computationally simpler and generates satisfactory results in simulation and data analysis. If needed, robust regression, such as quantile-based, can be conducted as in Step I. Besides the KM estimator, it can be of interest to estimate the conditional cumulative distribution function $F(t|X_{ik}, Z_{ij})$ using other approaches. The details will be studied in the future. The proposed approach can also be extended to accommodate non-linear or nonparametric G-E interactions. In Steps I and II, a nonparametric model, such as the varying coefficients model, can be adopted. In Step III, the censored quantile partial correlation can be developed based on a correlation measuring nonlinear dependence, for example the 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. In numerical study, we set quantile $\tau = 0.5$ which is one of the most popular choices in the literature. More numerical analysis with multiple quantiles may be of interest. For example, following the literature [33], we can compare the identified interactions across different quantiles. 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 A16 (Appendix) are at least partly attributable to 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.

Acknowledgments

We thank the associate editor and five reviewers for their careful review and insightful comments, which have led to a significant improvement of this article. This study has been supported by NIH awards R21CA191383 and R01CA204120, awards 61402276 and 91546202 from the National Natural Science Foundation of China, and award 13122402 from Innovative Research Team of Shanghai University of Finance and Economics.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2018.07.006.

References

[1] D.J. Hunter, Gene–environment interactions in human diseases, Nat. Rev. Genet. 6 (2005) 287–298.
[2] D. Thomas, Gene–environment–wide association studies: emerging approaches, Nat. Rev. Genet. 11 (2010) 259–272.
[3] N.I. Simonds, A.A. Ghazarian, C.B. Pimentel, S.D. Schully, G.L. Ellison, E.M. Gillanders, L.E. Mechanic, Review of the gene-environment interaction literature in cancer: what do we know? Genet. Epidemiol. 40 (2016) 356–365.
[4] T.T. Batchelor, R.A. Betensky, J.M. Esposito, L.D.D. Pham, M.V. Dorfman, N. Piscatelli, S. Jhung, D. Rhee, D.N. Louis, Age-dependent prognostic effects of genetic alterations in glioblastoma, Clin. Cancer Res. 10 (2004) 228–233.
[5] X. Shi, J. Liu, J. Huang, Y. Zhou, Y. Xie, S. Ma, A penalized robust method for identifying gene–environment interactions, Genet. Epidemiol. 38 (2014) 220–230.
[6] N. Sharafeldin, M.L. Slattery, Q. Liu, C. Franco-Villalobos, B.J. Caan, J.D. Potter, Y. Yasui, A candidate-pathway approach to identify gene-environment interactions: analyses of colon cancer risk and survival, J. Natl. Cancer Inst. 107 (2015).
[7] W. Stute, Distributional convergence under random censorship when covariables are present, Scand. J. Stat. (1996) 461–471.
[8] J.W. Osborne, A. Overbay, The power of outliers (and why researchers should always check for them), Pract. Assessment. Res. Eval. 9 (2004) 1–12.
[9] A.D. Shieh, Y.S. Hung, Detecting outlier samples in microarray data, Stat. Appl. Genet. Mol. Biol. 8 (2009) 1–24.
[10] H.J. Wang, L. Wang, Locally weighted censored quantile regression, J. Am. Stat. Assoc. 104 (2009) 1117–1128.
[11] J. Huang, S. Ma, H. Xie, Least absolute deviations estimation for the accelerated failure time model, Stat. Sin. (2007) 1533–1548.
[12] Y.G. Wang, M. Zhu, Rank-based regression for analysis of repeated measures, Biometrika 93 (2006) 459–464.
[13] K. Tharmaratnam, G. Claeskens, C. Croux, M. Salibian-Barrera, S-estimation for penalized regression splines, J. Comput. Graph. Stat. 19 (2010) 609–625.
[14] R. Koenker, G. Bassett Jr., Regression quantiles, Econom. J. Econom. Soc. (1978) 33–50.
[15] R. Koenker, J.A.F. Machado, Goodness of fit and related inference processes for quantile regression, J. Am. Stat. Assoc. 94 (1999) 1296–1310.
[16] H.J. Wang, L.A. Stefanski, Z. Zhu, Corrected-loss estimation for quantile regression with covariate measurement errors, Biometrika 99 (2012) 405–421.
[17] S. Lee, Y. Liao, M.H. Seo, Y. Shin, Oracle estimation of a change point in high dimensional quantile regression, J. Am. Stat. Assoc. (2017), https://doi.org/10.1080/01621459.2017.1319840.
[18] H.R. Frost, L. Shen, A.J. Saykin, S.M. Williams, J.H. Moore, A.D.N. Initiative, Identifying significant gene-environment interactions using a combination of screening testing and hierarchical false discovery rate control, Genet. Epidemiol. 40 (2016) 544–557.
[19] P. Zhang, J.P. Lewinger, D. Conti, J.L. Morrison, W.J. Gauderman, Detecting gene-environment interactions for a quantitative trait in a genome-wide association study, Genet. Epidemiol. 40 (2016) 394–403.
[20] S. Ma, R. Li, C.L. Tsai, Variable screening via quantile partial correlation, J. Am. Stat. Assoc. 112 (2017) 650–663.
[21] D. Lee, T. Neocleous, Bayesian quantile regression for count data with application to environmental epidemiology, J. R. Stat. Soc.: Ser. C: Appl. Stat. 59 (2010) 905–920.
[22] B. Efron, R. Tibshirani, Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy, Stat. Sci. (1986) 54–75.
[23] A. Hagemann, Cluster-robust bootstrap inference in quantile regression models, J. Am. Stat. Assoc. 112 (2017) 446–456.
[24] R. Song, W. Lu, S. Ma, X. Jessie Jeng, Censored rank independence screening for high-dimensional survival data, Biometrika 101 (2014) 799–814.
[25] Y. Liang, H. Chai, X.Y. Liu, Z.B. Xu, H. Zhang, K.S. Leung, Cancer survival analysis using semi-supervised learning method based on Cox and AFT models with $L_{1/2}$ regularization, BMC Med. Genet. 9 (2016) 11.
[26] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.C. Sanchez, M. Muller, pROC: an open-source package for R and S+ to analyze and compare ROC curves, BMC Bioinforma. 12 (2011) 77.
[27] P.M. Westcott, K.D. Halliwill, M.D. To, M. Rust Rashid, et al., The mutational landscapes of genetic and chemical models of Kras-driven lung cancer, Nature 517 (2015) 489–492.
[28] A.K. Smilde, H.A.L. Kiers, S. Bijlsma, C.M. Rubingh, M.J. Van Erk, Matrix correlations for high-dimensional data: the modified RV-coefficient, Bioinformatics 25 (2008) 401–405.
[29] J. Huang, S. Ma, Variable selection in the accelerated failure time model via the bridge method, Lifetime Data Anal. 16 (2010) 176–195.
[30] P.V. Dickson, J.E. Gershenwald, Staging and prognosis of cutaneous melanoma, Surg. Oncol. Clin. N. Am. 20 (2011) 1–17.
[31] J. Liu, J. Huang, Y. Zhang, Q. Lan, N. Rothman, T. Zheng, S. Ma, Identification of gene–environment interactions in cancer studies using penalization, Genomics 102 (2013) 189–194.
[32] C. Wu, Y. Jiang, J. Ren, Y. Cui, S. Ma, Dissecting gene-environment interactions: A penalized robust approach accounting for hierarchical structures, Stat. Med. 37 (2018) 437–456.
[33] A. Wey, L. Wang, K. Rudser, Censored quantile regression with recursive partitioning-based weights, Biostatistics 15 (2013) 170–181.