Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Econom Stat. Author manuscript; available in PMC 2019 May 29.

Published in final edited form as:

Econom Stat. 2017 October ; 4: 105–120. doi:10.1016/j.ecosta.2016.10.004.

Identifying gene-environment interactions for prognosis using a robust approach

Hao Chai[a], **Qingzhao Zhang**[b], **Yu Jiang**[c], **Guohua Wang**[d], **Sanguo Zhang**[d], **Syed Ejaz Ahmed**[e], and **Shuangge Ma**[a,*]

[a]Department of Biostatistics, Yale University, United States

[b]School of Economics and Wang Yanan Institute for Studies in Economics, Xiamen University, China

[c]School of Public Health, University of Memphis, United States

[d]School of Mathematical Sciences, University of Chinese Academy of Sciences, China

[e]Department of Mathematics and Statistics, Brock University, Canada

Abstract

For many complex diseases, prognosis is of essential importance. It has been shown that, beyond the main effects of genetic (G) and environmental (E) risk factors, gene-environment (G × E) interactions also play a critical role. In practical data analysis, part of the prognosis outcome data can have a distribution different from that of the rest of the data because of contamination or a mixture of subtypes. Literature has shown that data contamination as well as a mixture of distributions, if not properly accounted for, can lead to severely biased model estimation. In this study, we describe prognosis using an accelerated failure time (AFT) model. An exponential squared loss is proposed to accommodate data contamination or a mixture of distributions. A penalization approach is adopted for regularized estimation and marker selection. The proposed method is realized using an effective coordinate descent (CD) and minorization maximization (MM) algorithm. The estimation and identification consistency properties are rigorously established. Simulation shows that without contamination or mixture, the proposed method has performance comparable to or better than the nonrobust alternative. However, with contamination or mixture, it outperforms the nonrobust alternative and, under certain scenarios, is superior to the robust method based on quantile regression. The proposed method is applied to the analysis of TCGA (The Cancer Genome Atlas) lung cancer data. It identifies interactions different from those using the alternatives. The identified markers have important implications and satisfactory stability.

Keywords

Gene-environment interaction; Prognosis; Robustness; Exponential squared loss; Marker identification

*Corresponding author. shuangge.ma@yale.edu (S. Ma).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

1. Introduction

For cancer, diabetes, cardiovascular diseases, and many other diseases, prognosis is of essential interest. Profiling studies have been extensively conducted, searching for genetic markers associated with prognosis. It has been suggested that, beyond the main effects of genetic (G) and environmental (E) risk factors, gene-environment (G × E) interactions also have important implications. Multiple statistical methods have been developed for G × E interaction analysis. For reviews, see Caspi and Moffitt (2006), Cordell (2009), Thomas (2010), and others.

Denote T as the prognosis time of interest, $X = (X_1, \ldots, X_q)'$ as the q environmental/clinical variables, and $Z = (Z_1, \ldots, Z_p)'$ as the p genetic variables. Assume n independent subjects. In a typical profiling study, $q << n$, while p can be comparable to or even much larger than n. Regression-based interaction analysis, with its broad applicability, has been extensively adopted and proceeds as follows. (a) For gene j (= 1, …, p), consider the model $T \sim \phi(\alpha_{j,0} + \sum_{k=1}^{q} X_k \alpha_{j,k} + Z_j \beta_j + Z_j \sum_{k=1}^{q} X_k \gamma_{j,k})$, where the form of $\phi(\cdot)$ is known, and $\alpha_{j,0}$, $\alpha_{j,k}$, β_j, $\gamma_{j,k}$ are the unknown regression coefficients. With $q << n$, this is a low-dimensional model and can be fitted using standard, usually likelihood-based, techniques. Denote $p_{j,k}$ as the p-value for $\gamma_{j,k}$. (b) With $\{p_{j,k}: j = 1, \ldots, p; k = 1, \ldots, q\}$, conduct multiple comparison adjustment using the Bonferroini or FDR (false discovery rate) approach, and identify important interactions.

A limitation of the above approach is its lack of robustness. Usually it is assumed that all subjects satisfy the same prognosis models. In practice, most genetic studies cannot afford conducting rigorous subject selection. Seemingly homogeneous subjects can have different disease subtypes (Burgess, 2011; Haibe-Kains et al., 2012), leading to a mixture of survival distributions. Cause of death can be misclassified, leading to contamination in disease-specific survival (Fall et al., 2008). The survival times extracted from medical records are not always reliable (Bowman, 2011; Rampatige et al., 2013). To simplify terminology, we use “contamination” for both data contamination and a mixture of survival distributions. With non-robust – for example likelihood-based – estimation, even a single contaminated observation can lead to severely biased estimates (Huber and Ronchetti, 2009) and so false marker identification. Another limitation is that, significance level-based identification, although asymptotically valid, may generate unreliable results when sample sizes are small to moderate, as in typical profiling studies. Recent studies suggest that regularized estimation can lead to more reliable estimation and hence more accurate marker identification (Shi et al., 2014).

With low-dimensional data, robust methods have demonstrated great power. As suggested in a recent review (Wu and Ma, 2015), with high-dimensional genetic data, the development is limited and unsystematic. In the literature, relevant studies include (Gui et al., 2011), which identifies important interactions using the multifactor dimensionality reduction (MDR) technique. However, this method is limited to categorical data (such as SNPs) and not broadly applicable. Shi et al., 2014 developed a rank-based method, which is robust to model mis-specification but not data contamination. The most relevant study is Wang et al. (2015), which developed a quantile-regression based method. With that method, the quantile

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

needs to be specified, which is not trivial in data analysis. The objective function is not differentiable, causing difficulty in computation. In addition, as to be shown in this article, its numerical performance can be less satisfactory under certain data settings. With low-dimensional data, it has been shown that no robust method dominates the others. It is thus of interest to develop alternative robust methods.

Consider data with a prognosis outcome and both G and E measurements. The goal is to develop a new method for identifying important G × E interactions. Significantly advancing from the existing studies, the proposed method is robust to contamination in the prognosis data. In addition, under certain scenarios, its numerical performance can be better than quantile regression, which is perhaps the most popular robust method for genetic data (Wu and Ma, 2015). A penalization approach is adopted for marker identification, which differs from the significance level-based approach and can have better numerical performance when the sample size is small to moderate. We rigorously establish consistency properties, which provide a solid ground for the proposed method and also have independent value, considering that statistical properties for robust methods are very limitedly studied (Wu and Ma, 2015).

2. Robust identification of G × E interactions

2.1. Data and model settings

For describing prognosis, we adopt the accelerated failure time (AFT) model. Compared to alternatives such as the Cox model, this model has more intuitive interpretations and lower computational cost, both of which are especially desirable with high-dimensional genetic data. With a slight abuse of notation, still use T to denote the logarithm of prognosis time. For gene j, the AFT model postulates that

$$T = \alpha_{j,0} + \sum_{k=1}^{q} X_k \alpha_{j,k} + Z_j \beta_j + Z_j \sum_{k=1}^{q} X_k \gamma_{j,k} + \epsilon,$$

where ϵ is the random error.

Consider the scenario where a small subset of the random errors are contaminated or have a distribution different from that of the rest, leading to contamination in the prognosis times. For subject i (= 1, …, n), denote C_i as the logarithm of censoring time and $\mathbf{x}_i$ and $\mathbf{z}_i$ as the observed X and Z values, respectively. Under right censoring, we observe ($y_i = min(T_i, C_i)$, $\delta_i = I(T_i \leq C_i)$, $\mathbf{x}_i$, $\mathbf{z}_i$). Further denote $\mathbf{u}_{i,j} = (1, \mathbf{x}_i', z_{i,j}, z_{i,j}x_{i,1}, \ldots, z_{i,j}x_{i,q})'$, $\boldsymbol{\zeta}_j = (\alpha_{j,0}, \alpha_{j,1}, \ldots, \alpha_{j,q}, \beta_j, \gamma_{j,1}, \ldots, \gamma_{j,q})'$, and $\mathbf{U}j = (\mathbf{u}_{1,j}, \ldots, \mathbf{u}_{n,j})'$ which is a $n \times (2q + 2)$ matrix. Denote the kth element of $\mathbf{u}_{i,j}$ by $u_{(i,j)_k}$, $k = 1, \ldots, 2q + 2$. For gene j and subject i, the AFT model can now be written as $T_i = \mathbf{u}_{i,j}^{\top}\zeta_j + \epsilon_{i,j}$. Without loss of generality, assume that $\{(y_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i), i = 1, \ldots, n\}$ have been sorted according to y_i's from the smallest to the largest.

2.2. Penalized robust identification

A penalized marker identification method is defined by its loss function and penalty.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Loss function—First consider the scenario without contamination. When the distribution of ϵ is not specified, likelihood-based estimation is not applicable. A popular approach, especially for high-dimensional data, is the weighted least squared estimation developed in Stute (1993) and proceeds as follows. First compute the Kaplan–Meier weights as

$$\omega_1 = \frac{\delta_1}{n},\ \omega_i = \frac{\delta_i}{n-i+1}\prod_{j=1}^{i-1}\left(\frac{n-j}{n-j+1}\right)^{\delta_j},\ i = 2, \ldots, n.$$

For gene j(= 1, …, p), the weighted least squared objective function is defined as

$$\sum_{i=1}^{n}\omega_i(y_i - \mathbf{u}_{i,j}^{\top}\boldsymbol{\zeta}_j)^2. \quad (1)$$

Here to accommodate censoring, a weight function is used to re-weigh different observations according to their observed times and event status. Since the loss function has a quadratic form, it is not robust to data contamination. If subject i is not censored, then $\omega_i \neq 0$, and an arbitrarily large/small y_i results in arbitrarily large estimates. Biased estimation often happens with data contamination, leading to false marker identification.

To accommodate contamination, for gene j(= 1, …, p), consider the exponential squared loss function

$$Q(\boldsymbol{\zeta}_j|\mathrm{y}, \mathrm{U}_j, \boldsymbol{\omega}) = \sum_{i=1}^{n}\omega_i\exp(-(y_i - \mathbf{u}_{i,j}^{\top}\boldsymbol{\zeta}_j)^2/\theta). \quad (2)$$

Here $\mathbf{y}$ and $\boldsymbol{\omega}$ denote the vectors composed of y_i's and ω_i's, respectively, and $\theta > 0$ is a tuning parameter. The rationale of this approach is as follows. For a contaminated subject with y_i deviating from $\mathbf{u}_{i,j}^{\top}\zeta_j$ (the predicted value based on the model), $(y_i - \mathbf{u}_{i,j}^{\top}\boldsymbol{\zeta}_j)^2$ has a large value. The exponential function down-weighs such a contaminated observation. The degree of down-weighing is adjusted by θ: when θ gets smaller, contaminated observations have a smaller influence. To accommodate censoring, ω_i's are imposed in a similar manner as in the original Stute's approach. For low-dimensional linear regression models without censoring, the exponential squared loss has been examined in Wang et al. (2013). Advancing from the existing studies, here we consider the more challenging high-dimensional genetic data with interactions. In addition, the Kaplan–Meier weights are introduced to accommodate censoring. As to be shown in the Appendix, such differences lead to significant differences in statistical development.

Penalized estimation—For gene j(= 1, …, p), consider the penalized objective function

$$\mathrm{L}_{\lambda,\theta}(\boldsymbol{\zeta}_j\left|\mathbf{y}, \mathbf{U}_j, \boldsymbol{\omega}\right.) = \mathrm{Q}(\boldsymbol{\zeta}_j\left|\mathbf{y}, \mathbf{U}_j, \boldsymbol{\omega}\right.) - \lambda\left\|\boldsymbol{\zeta}_j\right\|_1. \quad (3)$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$\lambda > 0$ is the data-dependent tuning parameter, and $\|\cdot\|_1$ is the ℓ_1 norm. Denote $\tilde{\zeta}_j$ as the maximizer of $L_{\lambda,\theta}(\zeta_j|\mathbf{y}, \mathbf{U}_j, \omega)$. Interactions (and main effects) corresponding to the nonzero components of $\tilde{\zeta}_j$ are identified as important.

Note that we apply the same λ to all p genes, which ensures that they are analyzed on the same ground. This is different from analyzing the p genes separately, which may lead to different levels of regularization.

Multiple penalties can take the place of Lasso. In some recent studies (Bien et al., 2013; Liu et al., 2013), penalties have been developed to respect the "main effects, interactions" hierarchy, which reinforces that the main effects corresponding to the identified interactions must be identified. In studies such as Zimmermann et al. (2011) and Caspi and Moffitt (2006), it has been observed that genes can have important G × E interactions but no main effects. In addition, if the hierarchy has to be reinforced, one can identify important interactions first and then add back corresponding main effects. Computationally, Lasso is much simpler than the existing alternatives. Our limited experience suggests that with the complex robust loss function, more complicated penalties may have a higher probability of running into convergence problems. With the above considerations, the Lasso penalty is adopted.

2.3. Computation

We develop a coordinate-wise updating procedure to compute the solution to (3). For low-dimensional data under simpler settings, an iterative approach is suggested in Wang et al. (2013) to select the robust tuning parameter θ. However, under the present high-dimensional settings and with the coordinate-wise updating procedure, such an approach is computationally infeasible. Alternatively, we propose computing, for each (λ, θ) pair, the solution to each marginal model. This way, we generate a solution surface over the two-dimensional tuning parameter grid. The solution set can be comprehensively examined to identify appropriate tunings. Computation is conducted for each gene separately and can be realized in a highly parallel manner to reduce computer time. Consider gene j. Let $r_i(\zeta_j) = y_i - \mathrm{u}_{i,j}^\top \zeta_j$. The first and second order derivatives of $Q(\zeta_j)$ are

$$\begin{cases} \dot{Q}_k(\zeta_j) = \dfrac{\partial Q(\zeta_j)}{\partial \zeta_{j,k}} = 2\sum_{i=1}^n \omega_i u_{(i,j)_k} r_i(\zeta_j)\exp(-r_i^2(\zeta_j)/\theta)/\theta, \\ \ddot{Q}_{kl}(\zeta_j) = \dfrac{\partial^2 Q(\zeta_j)}{\partial \zeta_{j,k}\partial \zeta_{j,l}} = 2\sum_{i=1}^n \omega_i u_{(i,j)_k} u_{(i,j)_l}\exp(-r_i^2(\zeta_j)/\theta)(2r_i^2(\zeta_j)/\theta - 1)/\theta. \end{cases} \quad (4)$$

For ζ_j^m in a small neighborhood of ζ_j, $Q(\zeta_j)$ can be locally approximated by

$$Q(\zeta_j) \approx Q(\zeta_j^m) + \dot{Q}(\zeta_j^m)^\top(\zeta_j - \zeta_j^m) + \frac{1}{2}(\zeta_j - \zeta_j^m)^\top \ddot{Q}(\zeta_j^m)(\zeta_j - \zeta_j^m). \quad (5)$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Replacing $Q(\boldsymbol{\zeta}_j)$ in (3) with this approximation and taking the first order derivative of $L(\boldsymbol{\zeta}_j)$ with respect to the kth element $\zeta_{j,k}$ give us

$$\zeta_{j,k}^{m+1} = \begin{cases} \zeta_{j,k}^m - \ddot{Q}_{kk}^{-1}\left(\boldsymbol{\zeta}_j^m\right)\dot{Q}_k\left(\boldsymbol{\zeta}_j^m\right) + \ddot{Q}_{kk}^{-1}\left(\boldsymbol{\zeta}_j^m\right)\lambda, & \text{if } \zeta_{j,k}^m - \ddot{Q}_{kk}^{-1}\left(\boldsymbol{\zeta}_j^m\right)\dot{Q}_k\left(\boldsymbol{\zeta}_j^m\right) + \ddot{Q}_{kk}^{-1}\left(\boldsymbol{\zeta}_j^m\right)\lambda > 0; \\ \zeta_{j,k}^m - \ddot{Q}_{kk}^{-1}\left(\boldsymbol{\zeta}_j^m\right)\dot{Q}_k\left(\boldsymbol{\zeta}_j^m\right) - \ddot{Q}_{kk}^{-1}\left(\boldsymbol{\zeta}_j^m\right)\lambda, & \text{if } \zeta_{j,k}^m - \ddot{Q}_{kk}^{-1}\left(\boldsymbol{\zeta}_j^m\right)\dot{Q}_k\left(\boldsymbol{\zeta}_j^m\right) - \ddot{Q}_{kk}^{-1}\left(\boldsymbol{\zeta}_j^m\right)\lambda < 0; \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

Note that when $2r_i^2(\boldsymbol{\zeta})/\theta > 1$, $\ddot{Q}_{kk}(\boldsymbol{\zeta}) \geq 0$. Then (3) can be maximized at infinity, and the algorithm may fail to converge. To tackle this problem, first note that

$$\ddot{Q}_{kk}\left(\boldsymbol{\zeta}_j\right) \geq -2\sum_{i=1}^{n}\omega_i u_{(i,j)_k}^2 \exp\left(-r_i^2\left(\boldsymbol{\zeta}_j\right)/\theta\right)/\theta,$$

for $\omega_i \geq 0$. The right hand side is non-positive. We re-define

$$\ddot{Q}_{kk}\left(\boldsymbol{\zeta}_j\right) \equiv -2\sum_{i=1}^{n}\omega_i u_{(i,j)_k}^2 \exp\left(-r_i^2(\boldsymbol{\zeta})/\theta\right)/\theta$$

and use the minorization–maximization (MM) algorithm to compute the solution. The algorithm that combines coordinate descent and MM is summarized in Algorithm 1.

Algorithm 1: The coordinate descent + MM algorithm for marginal model j.

input : $\mathbf{y}, \boldsymbol{\omega}, \mathbf{U}_j$, and tuning parameters λ and θ.
output: the maximizer $\tilde{\boldsymbol{\zeta}}_j$ defined in (3).
initialization: let $m = 0, \boldsymbol{\zeta}_j^0 = \mathbf{0}$. Normalize $\mathbf{y}$ and $\mathbf{U}_j$ such that $\sum_{i=1}^n \omega_i y_i = 0$, $\sum_{i=1}^n \omega_i u_{(i,j)_k} = 0$, and $\sum_{i=1}^n \omega_i u_{(i,j)_k}^2 = n$ for $k = 1, \ldots, 2q+2$.
repeat
  $m_0 = m$;
  for $k = 1, 2, \ldots, 2q+2$ **do**
    repeat
      update $\zeta_{j,k}^m$ by $\zeta_{j,k}^{m+1}$ in (6) ;
      $\zeta_{j,l}^{m+1} = \zeta_{j,l}^m$ for $l = 1, 2, \ldots, 2q+2, l \neq k$;
      $m = m + 1$;
    until $|\zeta_{j,k}^m - \zeta_{j,k}^{m-1}| \leq$ *a predefined threshold. In our numerical study, we set the threshold* $= 10^{-3}$;
  end
until $||\boldsymbol{\zeta}_j^{m_0} - \boldsymbol{\zeta}_j^m||_2 \leq$ *a predefined threshold*;
return $\boldsymbol{\zeta}_j^m$ as the maximizer.

As mentioned above, we examine a two-dimensional grid of (λ, θ) values. The range of λ is determined as follows. First, its upper bound $\lambda_{\max}$ is selected such that $\tilde{\boldsymbol{\zeta}}_j = 0$ for all j. With the nonrobust weighted least squared loss, $\lambda_{\max} = \max_{j=1}^p \{\|\mathrm{U}_j^\top W \mathrm{y}\|_\infty\}$, where W is the diagonal matrix composed of ω_is. With the robust method, the derivatives in (4) can be viewed as a weighted sum of $u_{(i,j)_k} r_i\left(\boldsymbol{\zeta}_j\right)$s. Because the weight for each subject changes with $\boldsymbol{\zeta}_j$, the previously defined $\lambda_{\max}$ may not guarantee that $\tilde{\boldsymbol{\zeta}}_j = 0$ for all j. After some trials, we find that $\lambda_{\max} = 20\max_{j=1}^p \{\|\mathrm{U}_j^\top W \mathrm{y}\|_\infty\}$ is in general a "safe" upper bound for λ. The lower

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

bound is chosen as $\lambda_{\min} = \lambda_{\max}/1000$. For θ, we examine a relatively wide range to be cautious. Specifically, after centralization, we consider $\theta \in \left(\min_{i=1}^{n} y_i^2/100, \max_{i=1}^{n} y_i^2 \times 100\right)$.

With the specific form of the loss function, the existing techniques are not directly applicable to establish convergence of the proposed algorithm. We conjecture that the convergence properties are similar to those of the "classic" Lasso problems, but postpone rigorous investigation to future research. In all of our numerical studies, convergence is successfully achieved (for most datasets, within twenty iterations). The proposed algorithm is computationally affordable. For a simulated dataset, the analysis takes a few minutes on a regular desktop.

2.4. Consistency properties

Here we rigorously prove that the proposed method can consistently identify important interactions under ultrahigh-dimensional settings. For high-dimensional robust methods, theoretical development has been limited in the literature. With the consistency properties, the proposed method can be preferred over those alternatives whose statistical properties have not been well established. Our theoretical development not only provides a solid ground for the proposed method but also sheds insights upon other robust methods.

For each $j \in \{1, 2, \ldots, p\}$, define the population version of the marginal estimate as

$$\zeta_j^M = \text{argmax}_{\zeta_j} E\{\exp\left(-\left(T - \mathbf{U_j}^\top \zeta_j\right)^2/\theta\right)\},$$

where E denotes expectation under the true model. Denote the kth element of $\boldsymbol{\zeta}_j$ as $\zeta_{j,k}$. The corresponding important covariate effect index set in ζ_j^M is labeled as $S_j = \{k \in \{1, \ldots, 2q+2\} : \zeta_{j,k}^M \neq 0\}$. Denote $\mathcal{A}_X = \bigcup_{j=1}^{p} \left\{t : t \in S_j, 2 \le t \le q+1\right\}$ as the important set with its corresponding environmental variables important in at least one marginal model. If $q + 2 \in S_j$, then the jth gene is associated with prognosis in a marginal sense. The set $\{t : t \in S_j, q+3 \le t \le 2q+2\}$ contains important interactions between the jth gene and environmental variables. Then we have the important gene set $\mathcal{A}_G = \bigcup_{j=1}^{p} \left\{j : q+2 \in S_j\right\}$ and interaction set $\mathcal{A}_I = \bigcup_{j=1}^{p} \left\{(j-1)p + t - q - 2 : t \in S_j, q+3 \le t \le 2q+2\right\}$. Denote S^c and $|S|$ as the complement and cardinality of set S, respectively. If the truly important effects were known, then we would be able to compute the oracle estimator $\hat{\xi}_j$ with $\hat{\xi}_{S_j^c} = 0$ and

$$\hat{\xi}_{S_j} = \text{argmax} \sum_{i=1}^{n} \omega_i \exp(-(y_i - \mathbf{u}_{(i,j)S_j}^\top \xi_{j,S_j})^2/\theta) - \lambda \sum_{k \in S_j} |\zeta_{j,k}|. \quad (7)$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Theorem 1. *Consider the estimator defined in (7). Under conditions C1–C5 (*Appendix) and $6\left(n^{-\kappa}+12\rho_*^{-1}(q+1)\lambda\right)V<\rho_*$, we have

$$\Pr\left(\max_{1\le j\le p}\max_{s\in S_j}\left|\hat{\xi}_{j,s}-\zeta^M_{j,s}\right|\ge n^{-\kappa}+12q\rho_*^{-1}\lambda\right)\le p\exp\left(-\frac{\rho_*^2 n^{1-2\kappa}+144q^2n\lambda^2}{36\rho^*}\right)+4p(q+1)\exp\left(-\frac{n\rho_*^2}{36J(q+1)^2}\right),$$

where $\kappa<1/2$.

If $\lambda\to 0$ and $\log p=o(n^{1-2\kappa}+n\lambda^2)$, we have that $n^{-\kappa}+12q\rho_*^{-1}\lambda\to 0$ and the probability bound in Theorem 1 goes to zero. That is, the proposed method enjoys estimation consistency and is able to accommodate ultrahigh-dimensional data.

Recall that $\tilde{\boldsymbol{\xi}}_j=\operatorname{argmax}_{\boldsymbol{\zeta}\in\mathbb{R}^{2q+2}}L(\boldsymbol{\zeta}_j)$, where

$$L(\boldsymbol{\zeta}_j)=\sum_{i=1}^{n}\omega_i\exp(-(y_i-\mathbf{u}_{i,j}^{\top}\boldsymbol{\zeta}_j)^2/\theta)-\lambda\|\boldsymbol{\zeta}_j\|_1. \quad (8)$$

Since $L(\boldsymbol{\zeta}_j)$ is concave, if we can show that the oracle estimator $\hat{\boldsymbol{\xi}}_j$ satisfies the Karush–Kuhn–Tucher (KKT) condition, then $\tilde{\boldsymbol{\xi}}_j=\hat{\boldsymbol{\xi}}_j$. Define $\widetilde{\mathcal{A}_I}=\bigcup_{j=1}^{p}\left\{j:\tilde{\xi}_{j,q+2}\neq 0\right\}$ and $\widetilde{\mathcal{A}_I}=\bigcup_{j=1}^{p}\left\{(j-1)p+t-q-2:\tilde{\xi}_{j,t}\neq 0,q+3\le t\le 2q+2\right\}$. The following theorem establishes that the proposed method has the selection consistency properties.

Theorem 2. *Assume that the conditions in Theorem 1 hold. If* $\min_j\min_{S_j}\left\|\zeta^M_{j,s}\right\|\gg n^{-\kappa}+\lambda$ *and* $\max_j\left\|I_{S_jcS_j}\left(\boldsymbol{\zeta}_j^M\right)I_{S_jS_j}\left(\boldsymbol{\zeta}_j^M\right)^{-1}\right\|_\infty\le K<1$, *then we have*

$$\Pr\left(\mathcal{A}_G=\widetilde{\mathcal{A}_G}\text{ and }\mathcal{A}_I=\widetilde{\mathcal{A}_I}\right)\ge 1-O\left(p\exp\left(-\frac{\rho_*^2n^{1-2\kappa}+144q^2n\lambda^2}{36\rho^*}\right)+p\exp\left(-\frac{n\lambda^2(1-K)^2}{2\rho^*(1+K)^2}\right)\right).$$

With the probability bounds established in Theorems 1 and 2, we can derive the following result, which provides an easier way to comprehend the proposed method.

Corollary 1. *Suppose that conditions C1–C5 hold. If* $\lambda\to 0$, $n\lambda^2\to\infty$, $\log p=o(n^{1-2\kappa}+n\lambda^2)$ *with* $\kappa<1/2$, $\min_j\min_{S_j}\left|\zeta^M_{j,s}\right|\gg n^{-\kappa}+\lambda$ *and* $\max_j\left\|I_{S_jcS_j}\left(\boldsymbol{\zeta}_j^M\right)I_{S_jS_j}\left(\boldsymbol{\zeta}_j^M\right)^{-1}\right\|_\infty\le K<1$, *then with probability approaching one, the proposed method can identify the true sparsity structure and enjoy estimation consistency with an order of* $n^{-\kappa}+\lambda$.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Remark 1. Under simpler settings, it has been shown that the selection consistency of Lasso demands some variants of the irrepresentable condition. The above result shows that, under comparable conditions, the proposed robust method enjoys similar consistency.

3. Simulation

In simulation, we set $n = 300$, $q = 3$, and $p = 500$ and 1000. There are a total of 18 nonzero effects: 3 main E effects, 5 main G effects, and 10 interactions. The positions of nonzero main G effects and interactions are uniformly placed. The nonzero regression coefficients are randomly generated from *uniform*(0.5, 1.5). The E and G factors are generated from multivariate normal distributions with marginal means zero, marginal variances one, and the following variance matrix structures: Independent, AR(0.2), AR(0.8), Band(0.3), and Band(0.6). Under the AR(ρ) correlation structure, for the ith and jth factors, $corr = \rho^{|i-j|}$. Under the Band(ρ) correlation structure, for the ith and jth factors, $corr = \rho I(|i-j| \le 2)$, where $I(\cdot)$ is the indicator function. Under each correlation structure, consider seven different distributions for the random error ϵ: N(0,1), 0.95N(0,1) + 0.05Cauchy, 0.85N(0,1) + 0.15Cauchy, 0.7N(0,1) + 0.3Cauchy, 0.95N(0,1) + 0.05t(3), 0.85N(0,1) + 0.15t(3), and 0.7N(0,1) + 0.3t(3). That is, we consider the scenarios with no contamination and three different levels of contamination. Two contamination distributions are considered with different thickness of tails. The event times are generated from the AFT models. C_is are first generated independently from an exponential distribution and then log transformed. The parameter for the exponential distribution is adjusted so that the censoring rate is about 25%.

Beyond the proposed method (referred to as Robust), we also consider the following three alternatives: (1) the Nonrobust method, which adopts the weighted least squared loss function and applies the Lasso penalization for selecting important effects; (2) the Stute method, which adopts the weighted least squared loss function, does not apply any penalization, and uses significance level (p-value) as the criterion for quantifying the importance of effects; and (3) the Quantile method, which adopts the quantile regression-based robust loss function and applies Lasso penalization. More specifically, this method proceeds as follows. Let $\hat{F}(t) = 1 - \prod_{i=1}^{n} \left\{1 - \frac{1}{\sum_{l=1}^{n} 1(y_l \ge y_i)}\right\}^{\eta_i(t)}$, with $\eta_i(t) = I(y_i \le t, \delta_i = 1)$, be the nonparametric Kaplan–Meier estimate of the cumulative distribution function of survival time T. Define

$$w_i = \begin{cases} 1, & \hat{F}(C_i) > \tau \text{ or } T_i \le C_i; \\ \dfrac{\tau - \hat{F}(C_i)}{1 - \hat{F}(C_i)}, & \hat{F}(C_i) < \tau \text{ and } T_i > C_i. \end{cases}$$

Then for $j (= 1, \ldots, p)$, the penalized objective function is

$$L_n(\zeta_j) = \frac{1}{n}\sum_{i=1}^{n} \{w_i \rho_\tau(y_i - \mathbf{u}_{i,j}^\top \zeta_j) + (1 - w_i)\rho_\tau(Y^{+\infty} - \mathbf{u}_{i,j}^\top \zeta_j)\} + \lambda \|\zeta_j\|_1,$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

where $Y^{+\infty}$ is a value large enough to exceed all $u_{i,j}^{\top}\xi_j$s. The intuition behind the weights has been described in Zou (2006). Both the Nonrobust and Stute methods, along with the proposed, are built on the weighted least squared estimation. Comparing them can establish the advantage of robust loss and penalization-based identification, respectively. The Quantile method is also robust and has been recently developed. Comparing with it can establish the advantage of the weighted exponential squared loss. We acknowledge that some other methods are also applicable to the simulated data. The three alternatives have frameworks closest to that of the proposed and are hence considered.

With the Robust, Nonrobust, and Quantile methods, the number of selected interactions depends on the tuning parameter values. With the Stute method, the number depends on the *p*-value cutoff. To compare different methods on a fair ground, we examine a sequence of tuning parameter values, evaluate identification performance at each value, and use the ROC (receiver operating characteristics) curve to evaluate the interaction identification accuracy of different methods. A representative ROC plot is shown in Fig. 1. In this plot, the proposed method has dominatingly better accuracy.

Summary AUCs based on 100 replicates are shown in Tables 1 ($p = 500$) and 2 ($p = 1000$). When there is no contamination, performance of the proposed method is comparable to or slightly worse than that of the nonrobust method. For example with $p = 500$ and independent G measurements, the Robust and Nonrobust methods have mean AUCs 0.861 and 0.889, respectively. And with the AR(0.2) correlation, the mean AUCs area 0.901 and 0.881, respectively. With contamination, the Robust method outperforms the Nonrobust method. For example, with $p = 500$, the AR(0.2) correlation structure, and 0.7N(0,1) + 0.3Cauchy error, the Robust and Nonrobust methods have mean AUCs 0.886 and 0.751, respectively. Under all simulation scenarios, the Stute method, which adopts the robust loss function but significance level-based identification, has inferior identification accuracy. As has been suggested in the literature, with a moderate sample size, the unregularized estimates can be less reliable, leading to inaccurate identification. When comparing the proposed method with the quantile regression-based, we see that under the majority of the settings, the proposed method has superior performance. For example with $p = 500$, AR(0.2) correlation, and 0.85N(0,1) + 0.15Cauchy error, the proposed and Quantile methods have mean AUCs 0.892 and 0.842, respectively. However, under a small number of settings, the Quantile method excels. For example with $p = 500$, AR(0.8) correlation, and 0.7N(0,1) + 0.3t(3) error, the two methods have mean AUCs 0.863 and 0.933, respectively. Such observations are also reasonable. No robust approach is expected to be able to dominate all others. The proposed method outperforms the quantile regression-based method under most scenarios and provides a useful alternative.

Remark 2. Simulation may also suggest characteristics of the proposed method not fully described by the theoretical properties. It can be observed that the AUCs of the proposed method may have larger variances compared to the Stute and Quantile methods. In addition, the AUCs may not be “monotone” as a function of contamination rate. A closer examination of the consistency properties/proofs, computational algorithm, and computer code does not suggest obvious causes of these observations. It is *suspected* that they may be caused by the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

"interplay" of the parameter θ with data settings. That is, the impact of θ on identification may not be a monotone function of contamination. More investigation will be conducted in the future.

4. Analysis of the lung squamous cell carcinoma data

Lung squamous cell carcinoma is the second most common lung cancer and causes around 400,000 deaths each year worldwide (The Cancer Genome Atlas Research Network, 2012). Profiling studies have been extensively conducted, searching for its prognostic factors. In this section, we analyze the TCGA (The Cancer Genome Atlas) data on the prognosis of lung squamous cell carcinoma. The TCGA data were recently collected and published by NCI and have a high quality. The dataset we analyze was downloaded in April of 2015.

The prognosis outcome of interest is overall survival. Multiple environmental and clinical variables have been collected. We select the following for analysis: age, gender (female is coded as baseline), smoking level (pack years), and smoking status (non-smoker, reformed smoker for more than 15 years, reformed smoker for less than or equal to 15 years, current smoker; coded as 0, 1, 2 and 3). These variables have been suggested in the literature (Miller et al., 2004; Nakachi et al., 1991). A total of 422 samples have clinical and environmental measurements available.

For G factors, a total of 18,969 gene expression measurements are available for 502 samples. When matching the clinical/environmental data with genetic data, complete records are available for 404 samples. Among them, 129 died during followup, and their median followup was 30 months. The rest 275 were censored, and the median followup was 18 months.

In the literature, multiple approaches (especially cross validation-based) have been applied for selecting tuning parameters. However, most of the commonly used methods are based on the notion of prediction. In this study, we conduct marginal analysis, with the goal of identifying markers top-ranked in a marginal sense, not prediction. Thus, as opposed to applying cross validation and other prediction-based approaches, we follow published studies and vary the values of λ and θ so that a predetermined number of interactions are selected. In particular, we examine a wide range of θ values and focus on those with which the estimates are stable. A closer examination suggests that the estimates are not very sensitive to θ values for fixed λ. In Table 3, we provide results on the 33 top-ranked interactions. Longer or shorter lists of identified interactions are available from the authors. After the interactions are identified, we refit the marginal models without penalization and include the main effects to satisfy the "main effects, interactions" hierarchy. Beyond the proposed method, we also apply the three alternatives considered in simulation. Table 4 in Appendix suggests that different methods identify significantly different genes and interactions. Results under the proposed method are provided in Table 3. Those under the alternative methods are provided in Appendix.

To assess the stability of our findings, we apply the following approach. One sample is first removed from data, and then the proposed method is applied. This step is repeated over all

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

samples. For each identified interaction in Table 3, we compute its probability of being identified in the reduced datasets. It can be seen that three interactions have small stability measures, while all other interactions have stability measures close to 1. We have also examined those interactions not identified and found that their stability measures are all close to 0. This analysis suggests satisfactory stability of the proposed method.

Literature search suggests that the identified interactions and corresponding genes have important implications. Specifically, previous studies have shown that the major G × E interactions occur between genes and smoking status (Shi et al., 2014). Among the gene-smoking interactions identified in our study, the CEBPB (CCAAT/enhancer-binding proteins beta) protein is a transcription factor that works with other CCAAT/enhancer-binding protein family members in the regulation of cell cycle progress, differentiation, and pro-inflammatory gene expression. CEBPB has been found to be upregulated by tobacco smoke in both human lung fibroblast (Miglino et al., 2012) and mice emphysema. Another gene that interacts with smoking is EFNA1 (a.k.a. ephrin-A1). A recent *in vitro* study found that ephrin-A1 is overexpressed in tobacco smoke-treated bronchial airway epithelial cells compared to control cells (Nasreen et al., 2014). In addition, the elevated expressions of ephrin-A1 are positively associated with tumor proliferative capacity in non-small cell lung carcinoma patients. The gene that has a strong main effect as well as an interaction with duration of smoking is TERF1 (telomeric repeat binding factor 1). The TERF1 expression levels have been found to be decreased in lung cancer (Jian et al., 2009) as well as other types of cancer in several studies (Miyachi et al., 2002). It functions as an inhibitor of telomerase and is identified as a prognostic marker for overall survival in non-small cell lung cancer (Jian et al., 2009). The molecular mechanism of why and how TERF1 decreases in the process of cancer is not clear. Our results suggest that smoking can be one of the factors. Other than smoking duration, we also find that several genes interact with smoking intensity as well. The gene that interacts with both gender and smoking intensity is KATNB1. KATNB1 encodes protein katanin p80 subunit B1, which has been found to participate in cytokinesis by interacting with tumor suppressor gene LAPSER1. The disruption of the cytokinesis process may potentially cause genetic instability and cancer (Sudo and Maru, 2007). In addition to smoking, we find that eight genes interact with gender. Among these, STRADB and CA5BP1 draw our attention. STRADB is an important gene in lung cancer progression and metastasis through the activation of LKB1. LKB1 is essential for G1 cell cycle arrest, cell polarity and stress, cell detachment, and adhesion. The STRADB encoded protein also interacts with the X chromosome-linked inhibitor of apoptosis protein by enhancing its anti-apoptotic activity. In addition, gene CA5BP1 is located on the X chromosome and found to be gender-associated.

5. Discussion

G × E interactions have important implications for the prognosis of a large number of complex diseases. In this study, we have developed a new interaction analysis method. It can accommodate contamination and a mixture of distributions of the prognosis outcome, which are not uncommon in practice. In addition, we adopt penalization for identifying important interactions. This strategy differs from the commonly adopted significance level-based identification and may have improved marker identification accuracy for certain datasets.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Significantly advancing from many of the existing interaction studies, the consistency properties have been rigorously established under the ultrahigh-dimensional setting, making the proposed method one of the few with strong theoretical basis. An effective computational algorithm has been developed. In simulation study, the proposed method is observed to have satisfactory performance. It is interesting to note that it outperforms the quantile regression-based method under a variety of settings. In the analysis of lung cancer data, it identifies meaningful genes and interactions, which differ from those using the alternatives and have satisfactory stability.

In the literature, two types of interaction analysis have been conducted: marginal analysis and joint analysis. Both types of analysis have been popular, and neither can replace the other. In this study, we have focused on marginal analysis, where the importance of an interaction (or a main effect) is defined in a marginal sense. It is of interest to extend the proposed technique to joint analysis. However, that is highly nontrivial and will be pursued separately. It is noted that the proposed marginal analysis can be applied in joint analysis as a screening step. We refer to literature on sure independence screening for related discussions.

Acknowledgments

We thank the associate editor and a reviewer for careful review and insightful comments, which have led to a significant improvement of the article. This study has been supported by awards CA191383 and CA204120.

Appendix.

Proof of Theorems 1 and 2

For each $j \in \{1, 2, \ldots, p\}$, define

$$D_n(\zeta_j) = \sum_{i=1}^{n} \omega_i \exp(-(y_i - \mathbf{u}_{i,j}^\top \zeta_j)^2/\theta) \frac{2(y_i - \mathbf{u}_{i,j}^\top \zeta_j)}{\theta} \mathbf{u}_{i,j},$$

$$I_n(\zeta_j) = \frac{2}{\theta} \sum_{i=1}^{n} \omega_i \exp(-(y_i - \mathbf{u}_{i,j}^\top \zeta)^2/\theta) \left(\frac{2(y_i - \mathbf{u}_{i,j}^\top \zeta_j)^2}{\theta} - 1 \right) \mathbf{u}_{i,j} \mathbf{u}_{i,j}^\top.$$

Let τ_Y, τ_T, and τ_C be the end points of the support of Y, T, and C, respectively.

Assume the following regularity conditions. [C1] The observations $\{(y_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i), 1 \le i \le n\}$ are independent; [C2] T and C are independent and $P(T \le C|T, X, Z) = P(T \le C|T)$; [C3] $\tau_T < \tau_C$ or $\tau_T = \tau_C = \infty$; [C4] q is finite. For each $j \in \{1, 2, \ldots, p\}$, $\sqrt{n} D_n(\zeta_j^M) \to_d N(0, \Sigma_j)$, where Σ_j is a positive definite matrix. $I_n(\zeta_j^M)$ converges to a negative-definite matrix $I_n(\zeta_j^M)$ in probability. Moreover, the smallest eigenvalue $\rho_* = \min_j \rho_{\min}(-I(\zeta_j^M))$ and the largest eigenvalue $\rho^* = \max_j \rho_{\max}(\Sigma_j)$ are bounded away from zero and infinity. [C5] (i) Let N_j

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

denote a sufficiently small neighborhood centered around ζ_j^M. For $\zeta_j^1, \zeta_j^2 \in N_j$, there exists a bounded constant V such that $\zeta_j^\top [I(\zeta_j^1) - I(\zeta_j^2)]\zeta_j \le V\|\zeta_j^1 - \zeta_j^2\|_2$ for any $\|\zeta_j\|_2 = 1$. (ii) For any $k_1, k_2 \in \{1, \ldots, 2q+2\}$, $E(\exp(-(y_i - \mathbf{u}_{i,j}^\top \zeta_j^*)^2/\theta)\left(\frac{2(y_i - \mathbf{u}_{i,j}^\top \zeta_j^*)^2}{\theta} - 1\right)u_{(i,j)_{k_1}} u_{(i,j)_{k_2}}\Big)^2 \le J$, where J is finite and $\zeta_j^* \in N_j$.

Remark 3. Conditions C1–C3 have been commonly assumed in studies with random right censoring. Condition C4 is mild and has been proved under weak regularity conditions in Stute (1993) and Huang et al. (2007). Here, for simplicity, it is directly assumed. The constant V in Condition C5 (i) is introduced to ensure the stability of $I(\zeta_j)$ with respect to ζ_j. Since $\max_t \exp(-t)(2t-1) < ½$, a sufficient condition that can lead to Condition C5 (ii) is that $E(u_{(i,j)_{k_1}} u_{(i,j)_{k_2}})^2 \le 4J$. Such a finite second moment condition has been common in the literature.

Partition $I(\zeta_j^M)$ according to S_j as

$$I\left(\zeta_j^M\right) = \begin{pmatrix} I_{S_j S_j}\left(\zeta_j^M\right) & I_{S_j S_j^c}\left(\zeta_j^M\right) \\ I_{S_j^c S_j}\left(\zeta_j^m\right) & I_{S_j^c S_j^c}\left(\zeta_j^M\right) \end{pmatrix}.$$

Below we provide proofs for the two theorems.

Proof of Theorem 1.

Recall that

$$\hat{\zeta}_{j,S_j} = \operatorname{argmax}\left(\sum_{i=1}^n \omega_i \exp(-(y_i - \mathbf{u}_{(i,j)S_j}^\top \zeta_{j,S_j})^2/\theta) - \lambda \sum_{k \in S_j} |\zeta_{j,k}|\right). \quad (9)$$

Denote the above objective function as $R_n(\zeta_{j,S_j})$.

First, let $r_j = n^{-\kappa} + 12\rho_*^{-1}(q+1)\lambda$ and $\eta = \sum_{j=1}^p \exp(-\frac{\rho_*^2 n^{1-2\kappa} + 144(q+1)^2 n^{-1}\lambda^2}{36\rho^*})$, where $\kappa <$ 1/2. To prove

$$\Pr\left\{\left\|\zeta_{j,S_j} - \zeta_{j,S_j}^M\right\|_2 < r_j, j = 1, \ldots, p\right\} \ge 1 - \eta,$$

it suffices to show that

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\Pr\left(\sup_{\zeta_{j,S_j}\in\mathcal{I}} R_n(\zeta_{j,S_j}) < R_n(\zeta^M_{j,S_j}), j=1,\ldots,p\right) \geq 1-\eta, \quad (10)$$

where $\mathcal{I} = \left\{\zeta_{j,S_j} : \left\|\zeta_{j,S_j} - \zeta^M_{j,S_j}\right\|_2 = r_j, j=1,\ldots,p\right\}$. This implies that with probability at least $1-\eta$, $R_n(\zeta_j, S_j)$ has a global maximizer $\hat{\zeta}_{j,S_j}$ that satisfies $\left\|\zeta_{j,S_j} - \zeta^M_{j,S_j}\right\|_2 < r_j$, for $j = 1, \ldots, p$.

Recall the definitions of $D_n(\zeta_j)$ and $I_n(\zeta_j)$. Partition them according to S_j as

$$D_n(\zeta_j) = \begin{pmatrix} D_{n,S_j}(\zeta_j) \\ D_{n,S_j^c}(\zeta_j) \end{pmatrix}, \quad I_n(\zeta_j) = \begin{pmatrix} I_{n,S_jS_j}(\zeta_j) & I_{n,S_jS_j^c}(\zeta_j) \\ I_{n,S_j^cS_j}(\zeta_j) & I_{n,S_j^cS_j^c}(\zeta_j) \end{pmatrix}.$$

Obviously,

$$D_{n,S_j}(\zeta_j) = \sum_{i=1}^n \omega_i \exp\left(-\left(y_i - \mathbf{u}^\top_{(i,j)S_j}\zeta_{j,S_j}\right)^2/\theta\right) \frac{2\left(y_i - \mathbf{u}^\top_{(i,j)S_j}\zeta_{j,S_j}\right)}{\theta}\mathbf{u}_{(i,j)S_j}$$

and

$$I_{n,S_jS_j}(\zeta_j) = \frac{2}{\theta}\sum_{i=1}^n \omega_i \exp\left(-\left(y_i - \mathbf{u}^\top_{(i,j)S_j}\zeta_{j,S_j}\right)^2/\theta\right)\left(\frac{2\left(y_i - \mathbf{u}^\top_{(i,j)S_j}\zeta_{j,S_j}\right)^2}{\theta} - 1\right)\mathbf{u}_{(i,j)S_j}\mathbf{u}^\top_{(i,j)S_j}.$$

With Taylor's expansion we have

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$
\begin{aligned}
& R_n(\zeta_{j,S_j}) - R_n(\zeta^M_{j,S_j}) \qquad (11)\\
&= \sum_{i=1}^{n} \omega_i \left\{ \exp(-(y_i - \mathbf{u}^\top_{(i,j)_{S_j}} \zeta_{j,S_j})^2/\theta) - \exp(-(y_i - \mathbf{u}^\top_{(i,j)_{S_j}} \zeta^M_{j,S_j})^2/\theta) \right\}\\
&- \lambda \sum_{k \in S_j} (|\zeta_{j,k}| - |\zeta^M_{j,k}|)\\
&= D_{n,S_j}(\zeta^M_j)^\top (\zeta_{j,S_j} - \zeta^M_{j,S_j}) + \frac{1}{2}(\zeta_{j,S_j} - \zeta^M_{j,S_j})^\top I_{n,S_jS_j}(\bar{\zeta}_j)(\zeta_{j,S_j} - \zeta^M_{j,S_j})\\
&+ \lambda \sum_{k \in S_j} (|\zeta^M_{j,k}| - |\zeta_{j,k}|) \Big\},
\end{aligned}
$$

where $\bar{\zeta}_j$ lies between ζ^M_j and ζ_j.

It is easy to see that

$$
\begin{aligned}
& (\zeta_{j,S_j} - \zeta^M_{j,S_j})^\top I_{n,S_jS_j}(\bar{\zeta}_j)(\zeta_{j,S_j} - \zeta^M_{j,S_j}) \qquad (12)\\
&= (\zeta_{j,S_j} - \zeta^M_{j,S_j})^\top I_{S_jS_j}(\zeta^M_j)(\zeta_{j,S_j} - \zeta^M_{j,S_j}) + (\zeta_{j,S_j} - \zeta^M_{j,S_j})^\top \left\{ I_{S_jS_j}(\bar{\zeta}_j) - I_{S_jS_j}(\zeta^M_j) \right\}\\
&(\zeta_{j,S_j} - \zeta^M_{j,S_j}) + (\zeta_{j,S_j} - \zeta^M_{j,S_j})^\top \left\{ I_{n,S_jS_j}(\bar{\zeta}_j) - I_{S_jS_j}(\bar{\zeta}_j) \right\} (\zeta_{j,S_j} - \zeta^M_{j,S_j})\\
&= Q_1 + Q_2 + Q_3 .
\end{aligned}
$$

By C4, $Q_1 \le -\left\|\zeta_{j,S_j} - \zeta^M_{j,S_j}\right\|_2^2 \rho_*$. Moreover, $Q_2 \le V \left\|\zeta_{j,S_j} - \zeta^M_{j,S_j}\right\|_2^3$ under C5. Bernstein inequality and C5 yield

$$
\Pr\left(\left\| I_{n,S_jS_j}(\bar{\zeta}_j) - I_{S_jS_j}(\bar{\zeta}_j) \right\|_F^2 \ge \frac{\rho_*^2}{9} \right) \le 2|S_j| \exp\left(-\frac{n\rho_*^2}{9J|S_j|^2}\right),
$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Since $\lambda_{\max}(I_{n,S_jS_j}(\bar{\zeta}_j) - I_{S_jS_j}(\bar{\zeta})) \le \left\| I_{n,S_jS_j}(\bar{\zeta}_j) - I_{S_jS_j}(\bar{\zeta}_j)_F \right\|$, we have $Q_3 \le \frac{1}{3}\rho_* r_j^2$. Therefore, the second term in (11) is controlled by

$$
\frac{1}{2}(\zeta_{j,S_j} - \zeta^M_{j,S_j})^\top I_{n,S_jS_j}(\zeta^M_j)(\zeta_{j,S_j} - \zeta^M_{j,S_j}) < -\frac{1}{3}\rho_* r_j^2 + \frac{1}{2} V r_j^3, \qquad (13)
$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

with probability at least $1 - 4(q+1)\exp(-\frac{n\rho_*^2}{36J(q+1)^2})$ due to (12) and $|S_j| \le 2q+2$.

Partition Σ_j according to S_j as $\begin{pmatrix} \Sigma_{S_jS_j} & \Sigma_{S_jS_j^c} \\ \Sigma_{S_j^cS_j} & \Sigma_{S_j^cS_j^c} \end{pmatrix}$. For $D_{n,S_j}(\zeta_j^M)$, by the definition of ζ_{j,S_j}^M and C2 and C4, we have

$$\sqrt{n}D_{n,S_j}(\zeta_j^M) \to_d N(0, \Sigma_{S_jS_j}).$$

Then for any given t, an application of Bernstein's inequality yields

$$\Pr(|D_{n,S_j}(\zeta_j^M)^\top(\zeta_{j,S_j} - \zeta_{j,S_j}^M)| > t) \le 2\exp\left(-\frac{nt^2}{2(\zeta_{j,S_j} - \zeta_{j,S_j}^M)^\top \Sigma_{S_jS_j}(\zeta_{j,S_j} - \zeta_{j,S_j}^M)}\right).$$

Recall that $r_j = n^{-\kappa} + 12\rho_*^{-1}(q+1)\lambda$. Let $t = \frac{1}{6}\rho_* r_j^2$, then we have

$$\Pr(D_{n,S_j}(\zeta_j^M)^\top(\zeta_{j,S_j} - \zeta_{j,S_j}^M) > \frac{1}{6}\rho_* r_j^2) \le \exp\left(-\frac{\rho_*^2 n^{1-2\kappa} + 144(q+1)^2 n\lambda^2}{36\rho^*}\right). \quad (14)$$

By the Triangle inequality and $(\sum_{i=1}^d |v_i|)^2 \le d\sum_{i=1}^d v_i^2$ for any sequence v_i, we have

$$\lambda \sum_{k \in S_j} (|\zeta_{j,k}^M| - |\zeta_{j,k}|) \le \lambda \sum_{k \in S_j} |\zeta_{j,k}^M - \zeta_{j,k}| \le \lambda\sqrt{|S_j|}\left\|\zeta_{j,S_j} - \zeta_{j,S_2}^M\right\|_2. \quad (15)$$

Combining (11), (13), (14), (15), and $6(n^{-\kappa} + 12\rho_*^{-1}(q+1)\lambda)V < \rho_*$, we have

$$R_n(\zeta_{j,S_j}) - R_n(\zeta_{j,S_j}^M) < -\frac{1}{6}\rho_* r_j^2 + \lambda\sqrt{|S_j|}r_j + \frac{1}{2}Vr_j^3 < 0 \quad (16)$$

with probability at least $1 - \exp\left(-\frac{\rho_*^2 n^{1-2\kappa} + 144(q+1)^2 n\lambda^2}{36\rho^*}\right) - 4(q+1)\exp\left(-\frac{n\rho_*^2}{36J(q+1)^2}\right)$. With the Bonferroni's inequality, this theorem is proved. □

Proof of Theorem 2.

Recall that $\tilde{\zeta}_j = \operatorname{argmax}_{\zeta \in \mathbb{R}^{2q+2}} L(\zeta_j)$, where

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$L(\boldsymbol{\zeta}_j) = \sum_{i=1}^{n} \omega_i \exp(-(y_i - \mathbf{u}_{i,j}^{\top}\boldsymbol{\zeta}_j)^2/\theta) - \lambda \left\|\boldsymbol{\zeta}_j\right\|_1. \quad (17)$$

Consider the oracle estimator $\hat{\boldsymbol{\zeta}}_j$ with $\hat{\boldsymbol{\zeta}}_{S_j^c} = 0$ and

$$\hat{\boldsymbol{\zeta}}_{j,S_j} = \operatorname{argmax}\left(\sum_{i=1}^{n} \omega_i \exp(-(y_i - \mathbf{u}_{(i,j)S_j}^{\top}\boldsymbol{\zeta}_{j,S_j})^2/\theta) - \lambda \sum_{k \in S_j} |\zeta_{j,k}| \right). \quad (18)$$

Denote the above objective function as $R_{n,S_j}(\boldsymbol{\zeta}_j)$. Since $L(\boldsymbol{\zeta}_j)$ in (8) is concave, if we can show that $\hat{\boldsymbol{\zeta}}_j$ satisfies the Karush–Kuhn–Tucher (KKT) condition, then $\tilde{\boldsymbol{\zeta}}_j = \hat{\boldsymbol{\zeta}}_j$.

Next we want to show that

$$\left|\Omega_n(S_j^c)\right|_{\infty} < \lambda, \quad j = 1, 2, \ldots, p, \quad (19)$$

where $|v|_{\infty} = \max_i |v_i|$ for any vector $v = (v_1, \cdots, v_{|S_j^c|})$ and

$$\Omega_n(S_j^c) = \sum_{i=1}^{n} \omega_i \exp\left(-\frac{(y_i - \mathbf{u}_{(i,j)S_j}^{\top}\hat{\boldsymbol{\zeta}}_{j,S_j})^2}{\theta}\right) \frac{2(y_i - \mathbf{u}_{(i,j)S_j}^{\top}\hat{\boldsymbol{\zeta}}_{j,S_j})}{\theta} \mathbf{u}_{(i,j)S_j^c}.$$

Applying Taylor's expansion, we have

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\Omega_n(S_j^c) = \sum_{i=1}^{n} \omega_i \exp\left\{-\frac{(y_i - \mathbf{u}_{(i,j)S_j}^\top \boldsymbol{\xi}_{j,S_j}^M)^2}{\theta}\right\} \frac{2(y_i - \mathbf{u}_{(i,j)S_j}^\top \boldsymbol{\xi}_{j,S_j}^M)}{\theta} \mathbf{u}_{(i,j)S_j^c} \quad (20)$$

$$+ \frac{2}{\theta} \sum_{i=1}^{n} \omega_i \exp\left\{-\frac{(y_i - \mathbf{u}_{(i,j)S_j}^\top \bar{\boldsymbol{\xi}}_{j,S_j})^2}{\theta}\right\} \left(\frac{2(y_i - \mathbf{u}_{(i,j)S_j}^\top \bar{\boldsymbol{\xi}}_{j,S_j})^2}{\theta} - 1\right)$$

$$\times \mathbf{u}_{(i,j)S_j^c} \mathbf{u}_{(i,j)S_j}^\top (\hat{\boldsymbol{\xi}}_{j,S_j} - \boldsymbol{\xi}_{j,S_j}^M)$$

$$:= \Gamma_n + \Delta_n,$$

where $\bar{\boldsymbol{\xi}}_j$ lies between $\boldsymbol{\xi}_j^M$ and $\hat{\boldsymbol{\xi}}_j$. From the proof of Theorem 1, we have

$$\hat{\boldsymbol{\xi}}_{j,S_j} - \boldsymbol{\xi}_{j,S_j}^M = I_{n,S_jS_j}(\boldsymbol{\xi}_j^M)^{-1}\left\{-D_{n,S_j}(\boldsymbol{\xi}_j^M) + \lambda \mathrm{sgn}(\boldsymbol{\xi}_j^M)\right\}. \quad (21)$$

Substituting (21) into (20), we have

$$\Delta_n = \frac{2}{\theta} \sum_{i=1}^{n} \omega_i \exp\left\{-\frac{(y_i - \mathbf{u}_{(i,j)S_j}^\top \bar{\boldsymbol{\xi}}_{j,S_j})^2}{\theta}\right\} \left(\frac{2(y_i - \mathbf{u}_{(i,j)S_j}^\top \bar{\boldsymbol{\xi}}_{j,S_j})^2}{\theta} - 1\right) \quad (22)$$

$$\times \mathbf{u}_{(i,j)S_j^c} \mathbf{u}_{(i,j)S_j}^\top I_{n,S_j}(\boldsymbol{\xi}_j^M)^{-1}\{-D_{n,S_j}(\boldsymbol{\xi}_j^M) + \lambda \mathrm{sgn}(\boldsymbol{\xi}_j^M)\}$$

$$= -I_{n,S_j{}^cS_j}(\bar{\boldsymbol{\xi}}_j) I_{n,S_jS_j}(\boldsymbol{\xi}_j^M)^{-1} D_{n,S_j}(\boldsymbol{\xi}_j^M) + \lambda I_{n,S_j{}^cS_j}(\bar{\boldsymbol{\xi}}_j) I_{n,S_j}(\boldsymbol{\xi}_j^M)^{-1} \mathrm{sgn}(\boldsymbol{\xi}_j^M).$$

Define

$$\Delta_n^* = -I_{S_j{}^cS_j}\left(\boldsymbol{\xi}_j^M\right) I_{S_jS_j}\left(\boldsymbol{\xi}_j^M\right)^{-1} D_{n,S_j}\left(\boldsymbol{\xi}_j^M\right) + \lambda I_{S_j{}^cS_j}\left(\boldsymbol{\xi}_j^M\right) I_{S_jS_j}\left(\boldsymbol{\xi}_j^M\right)^{-1} \mathrm{sgn}\left(\boldsymbol{\xi}_j^M\right),$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and $\Omega_n^*(S_j^c) = \Gamma_n + \Delta_n^*$. From the proof of Theorem 1, we see that the tail probability for $I_n(\zeta_j)$ is dominated by that for $D_n(\zeta_j)$. Thus

$$\Pr(\left|\Omega_n(S_j^c)\right|_\infty > \lambda) \asymp \Pr(\left|\Omega_n^*(S_j^c)\right|_\infty > \lambda).$$

Therefore, combining the above arguments, we only need to focus on $\Omega_n^*(S_j^c)$. In fact,

$$\begin{aligned}
|\Omega_n^*(S_j^c)|_\infty &\le |\Gamma_n|_\infty + |\Delta_n^*|_\infty \\
&\le |D_{n,S_j^c}(\xi_j^M)|_\infty + |I_{S_j^c S_j}(\xi_j^M) I_{S_j S_j}(\xi_j^M)^{-1} D_{n,S_j}(\xi_j^M)|_\infty \\
&\quad + \lambda |I_{S_j^c S_j}(\xi_j^M) I_{S_j S_j}(\xi_j^M)^{-1} D_{n,S_j}(\xi_j^M) \mathrm{sgn}(\xi_j^M)|_\infty \\
&\le |D_n(\xi_j^M)|_\infty + \|I_{S_j^c S_j}(\xi_j^M) I_{S_j S_j}(\xi_j^M)^{-1}\|_\infty |D_n(\xi_j^M)|_\infty + \lambda \|I_{S_j^c S_j}(\xi_j^M) I_{S_j S_j}(\xi_j^M)^{-1}\|_\infty
\end{aligned} \tag{23}$$

With the condition $\Phi_j = \left\| I_{S_j^c S_j}(\xi_j^M) I_{S_j S_j}(\xi_j^M)^{-1} \right\|_\infty \le K < 1$, if

$$\left|D_n^*(\xi_j^M)\right|_\infty < \lambda \frac{1-\Phi_j}{1+\Phi_j}, \tag{24}$$

then from (23), it follows that

$$|\Omega_n^*(S_j^c)|_\infty \le |D_n(\xi_j^M)|_\infty (1+\Phi_j) + \lambda \Phi_j < \lambda(1-\Phi_j) + \lambda \Phi_j = \lambda,$$

which proves (19). We now derive the probability bound for the event in (24). Similarly to the derivation of (14),

$$\Pr\left\{ |D_n(\xi_j^M)|_\infty \ge \lambda \frac{1-\Phi_j}{1+\Phi_j} \right\} \le 2\exp\left(-\frac{n\lambda^2 (1-\Phi_j)^2}{2\rho^* (1+\Phi_j)^2} \right). \tag{25}$$

By the Bonferroni's inequality, we obtain

$$\Pr\left\{ |D_n(\xi_j^M)|_\infty < \lambda \frac{1-\Phi_j}{1+\Phi_j}, j = 1, 2, \ldots, p \right\} \ge 1 - 2\sum_{j=1}^{p} \exp\left(-\frac{n\lambda^2 (1-\Phi_j)^2}{2\rho^* (1+\Phi_j)^2} \right). \tag{26}$$

This theorem is proved with the above results. □

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

References

Bien J, Taylor J, Tibshirani R, 2013 A lasso for hierarchical interactions. Ann. Stat 41 (3), 1111–1141. [PubMed: 26257447]

Bowman L (2011). Doctors, researchers worry about accuracy of social security "death file" http://projects.scrippsnews.com/story/doctors-researchers-worry/.

Burgess DJ, 2011 Cancer genetics: Initially complex, always heterogeneous. Nat. Rev. Cancer 11 (3), 153–153. [PubMed: 21451547]

Caspi A, Moffitt TE, 2006 Gene-environment interactions in psychiatry: joining forces with neuroscience. Nat. Rev. Neurosci 7 (7), 583–590. [PubMed: 16791147]

Cordell HJ, 2009 Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 10 (6), 392–404. [PubMed: 19434077]

Fall K, Stromberg F, Rosell J, Andren O, Varenhorst E, 2008 Reliability of death certificates in prostate cancer patients. Scand. J. Urol. Nephrol 42 (4), 352–357. [PubMed: 18609293]

Gui J, Moore JH, Kelsey KT, Marsit CJ, Karagas MR, Andrew AS, 2011 A novel survival multifactor dimensionality reduction method for detecting gene–gene interactions with application to bladder cancer prognosis. Hum. Genet 129 (1), 101–110. [PubMed: 20981448]

Haibe-Kains B, Desmedt C, Loi S, Culhane AC, Bontempi G, Quackenbush J, Sotiriou C, 2012 A three-gene model to robustly identify breast cancer molecular subtypes. J. Natl. Cancer Inst 104 (4), 311–325. [PubMed: 22262870]

Huang J, Ma S, Xie H, 2007 Least absolute deviations estimation for the accelerated failure time model. Stat. Sin 17 (4), 1533–1548.

Huber PJ, Ronchetti EM, 2009 Robust Statistics. Wiley series in probability and statistics, 2nd Wiley, Hoboken, N.J.

Jian H, Li S, Lu-Ming W, Su-Jun J, 2009 Analysis of the nuclear localization signal of TRF1 in non-small cell lung cancer. Biol. Res 42, 217–222. [PubMed: 19746267]

Liu J, Huang J, Zhang Y, Lan Q, Rothman N, Zheng T, Ma S, 2013 Identification of gene–environment interactions in cancer studies using penalization. Genomics 102 (4), 189–194. [PubMed: 23994599]

Miglino N, Roth M, Lardinois D, Sadowski C, Tamm M, Borger P, 2012 Cigarette smoke inhibits lung fibroblast proliferation by translational mechanisms. Eur. Respir. J 39 (3), 705–711. [PubMed: 21852335]

Miller VA, Kris MG, Shah N, Patel J, Azzoli C, Gomez J, Krug LM, Pao W, Rizvi N, Pizzo B, Tyson L, Venkatraman E, Ben-Porat L, Memoli N, Zakowski M, Rusch V, Heelan RT, 2004 Bronchioloalveolar pathologic subtype and smoking history predict sensitivity to gefitinib in advanced non-small-cell lung cancer. J. Clin. Oncol 22 (6), 1103–1109. [PubMed: 15020612]

Miyachi K, Fujita M, Tanaka N, Sasaki K, Sunagawa M, 2002 Correlation between telomerase activity and telomeric-repeat binding factors in gastric cancer. J. Exp. Clin. Cancer Res. : CR 21 (2), 269–275. [PubMed: 12148588]

Nakachi K, Imai K, Hayashi S. i., Watanabe J, Kawajiri K, 1991 Genetic susceptibility to squamous cell carcinoma of the lung in relation to cigarette smoking dose. Cancer Res 51 (19), 5177–5180. [PubMed: 1655248]

Nasreen N, Khodayari N, Sriram PS, Patel J, Mohammed KA, 2014 Tobacco smoke induces epithelial barrier dysfunction via receptor epha2 signaling. Am. J. Physiol. - Cell Physiol 306 (12), C1154–C1166. [PubMed: 24717580]

The Cancer Genome Atlas Research Network, 2012 Comprehensive genomic characterization of squamous cell lung cancers. Nature 489 (7417), 519–525. [PubMed: 22960745]

Rampatige R, Gamage S, Peiris S, Lopez AD, 2013 Assessing the reliability of causes of death reported by the vital registration system in sri lanka: medical records review in colombo. HIM J 42 (3), 20–28.

Shi X, Liu J, Huang J, Zhou Y, Xie Y, Ma S, 2014 A penalized robust method for identifying gene-environment interactions. Genet. Epidemiol 38 (3), 220–230. [PubMed: 24616063]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Stute W, 1993 Consistent estimation under random censorship when covariables are present. J. Multivar. Anal 45 (1), 89–103.

Sudo H, Maru Y, 2007 Lapser1 is a putative cytokinetic tumor suppressor that shows the same centrosome and midbody subcellular localization pattern as p80 katanin. FASEB J 21 (9), 2086–2100. [PubMed: 17351128]

Thomas D, 2010 Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. Annu. Rev. Public Health 31, 21–36. [PubMed: 20070199]

Wang G, Zhang Q, Zang Y, Zhang S, Ma S, 2015 Identifying gene-environment interactions associated with prognosis using penalized robust regression. In: Big and Complex Data Analysis: Statistical Methodologies and Applications Springer.

Wang X, Jiang Y, Huang M, Zhang H, 2013 Robust variable selection with exponential squared loss. J. Am. Stat.Assoc 108 (502), 632–643. [PubMed: 23913996]

Wu C, Ma S, 2015 A selective review of robust variable selection with applications in bioinformatics. Brief. Bioinform 16 (5), 873–883. [PubMed: 25479793]

Zimmermann P, Bruckl T, Nocon A, Pfister H, Binder E, Uhr M, Lieb R, Moffitt T, Caspi A, Holsboer F, Ising M, 2011 Interaction of fkbp5 gene variants and adverse life events in predicting depression onset: Results from a 10 -year prospective community study. Am. J. Psychiatry 168 (10), 1107–1116. [PubMed: 21865530]

Zou H, 2006 Locally weighted censored quantile regression. J. Am. Stat. Assoc 104 (487), 1117–1128.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

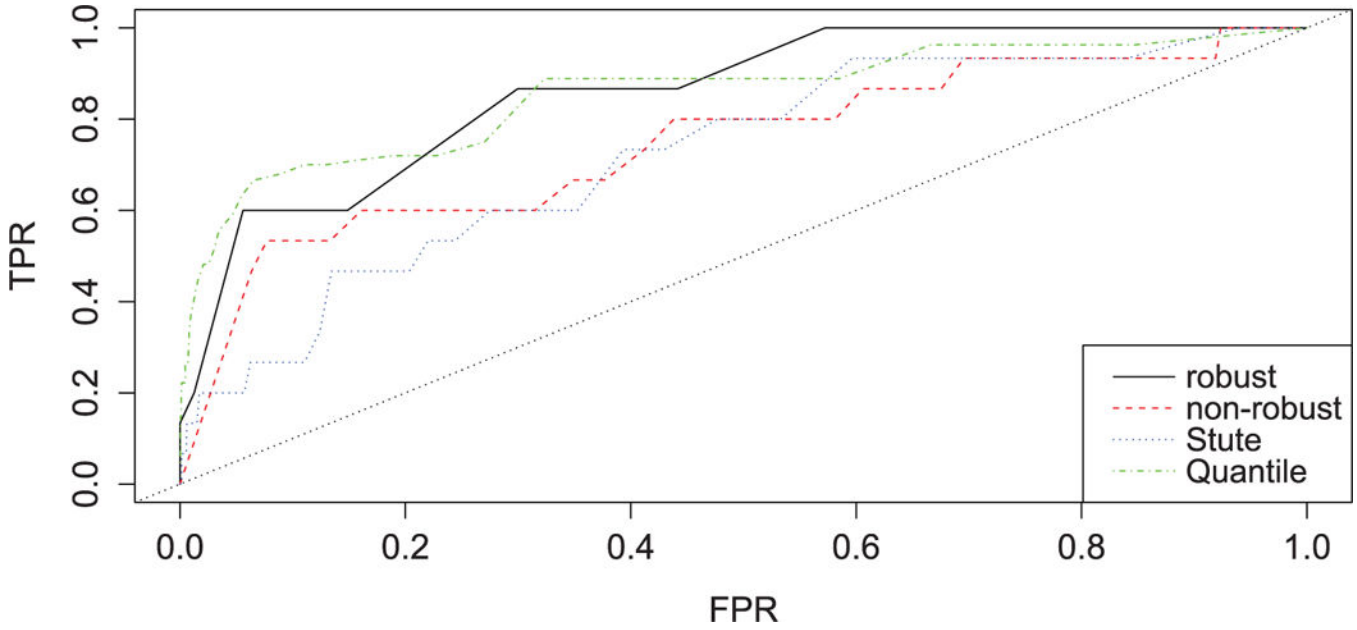

Fig. 1.
An illustration of the ROC curves for the proposed and alternative methods.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

Simulation: AUC × 100 (sd) based on 100 replicates, $p = 500$.

Error	Method	Independent	AR(0.2)	AR(0.8)	Band(0.3)	Band(0.6)
N(0,1)	Robust	86.1(7.7)	90.1(6.6)	86.7(6.9)	89.2(5.7)	94.3(5.8)
	Nonrobust	88.9(6.6)	88.1(5.7)	86.8(6.1)	83.6(6.3)	91.0(8.2)
	Stute	65.7(4.4)	63.3(3.3)	61.6(4.6)	64.7(3.5)	61.0(2.3)
	Quantile	82.9(3.3)	86.3(1.2)	93.5(2.1)	88.9(2.5)	76.5(5.1)
0.95N(0,1)+0.05Cauchy	Robust	91.1(7.8)	87.1(6.9)	88.7(5.1)	87.7(6.6)	88.1(6.5)
	Nonrobust	82.8(8.6)	85.9(10)	88.7(8.5)	81.9(12.6)	78.5(10.2)
	Stute	68.0(4.7)	60.1(4.6)	70.6(4.8)	64.7(3.5)	73.3(4.7)
	Quantile	82.7(3.1)	83.4(3.4)	92.4(2.9)	87.4(2.5)	71.9(3.9)
0.85N(0,1)+0.15Cauchy	Robust	86.6(8.2)	89.2(7.0)	93.9(6.9)	86.8(6.9)	83.5(5.8)
	Nonrobust	77.1(9.4)	85.0(10.2)	87.1(12.7)	72.8(10.3)	76.2(11.6)
	Stute	64.2(6.3)	63.7(5.5)	67.5(6.9)	64.7(3.5)	67.8(6.4)
	Quantile	81.9(3.1)	84.2(1.7)	93.6(2.7)	89.4(2.4)	69.1(5.1)
0.7N(0,1)+0.3Cauchy	Robust	84.8(6.9)	88.6(6.6)	87.9(5.9)	89.2(5.7)	90.1(5.9)
	Nonrobust	71.7(11.8)	75.1(13.3)	77.8(12.8)	80.8(6.5)	86.3(8.2)
	Stute	64.5(8.2)	59.4(5.8)	65.7(8.5)	64.7(3.5)	64.4(6.6)
	Quantile	80.1(2.7)	85.8(2.3)	92.5(2.4)	88.2(2.9)	68.1(8.3)
0.95N(0,1)+0.05t(3)	Robust	80.8(7.3)	89.4(6.1)	93.4(5.0)	69.9(5.4)	88.8(6.7)
	Nonrobust	76.7(9.1)	84.7(8.4)	89.6(9.0)	82.1(11.6)	85.4(10.1)
	Stute	60.8(4.5)	66.7(4.9)	68.6(5.0)	64.7(3.5)	73.8(4.9)
	Quantile	82.6(2.1)	85.4(2.6)	89.7(2.5)	87.8(2.2)	73.4(6.6)
0.85N(0,1)+0.15t(3)	Robust	87.5(7.5)	85.6(6.6)	90.6(6.0)	87.9(6.5)	79.8(5.8)
	Nonrobust	82.3(11.7)	79.4(10.6)	83.5(11.8)	75.5(13.9)	75.5(12.2)
	Stute	67.4(5.4)	61.4(5.1)	68.6(5.0)	64.7(3.5)	68.2(4.4)
	Quantile	83.1(3.4)	85.3(3.1)	92.4(1.6)	87.7(3.6)	72.7(4.1)
0.7N(0,1)+0.3t(3)	Robust	84.4(7.2)	88.6(6.8)	86.3(6.7)	88.4(5.3)	85.4(5.3)
	Nonrobust	71.9(11.7)	71.5(12.3)	76.0(13.1)	80.7(6.4)	80.0(4.8)
	Stute	62.3(8.4)	62.0(7.4)	68.6(5.0)	64.7(3.5)	60.6(6.9)
	Quantile	79.8(4.1)	83.9(2.6)	93.3(1.9)	87.1(3.1)	68.5(6.3)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2

Simulation: AUC × 100 (sd) based on 100 replicates, $p = 1,000$.

Error	Method	Independent	AR(0.2)	AR(0.8)	Band(0.3)	Band(0.6)
N(0,1)	Robust	89.2(6.8)	87.4(6.7)	94.0(5.0)	92.8(7.2)	85.4(5.3)
	Nonrobust	84.4(6.1)	87.9(6.3)	91.7(4.5)	86.8(7.3)	80.0(4.8)
	Stute	64.7(6.0)	66.7(4.2)	68.6(5.0)	76.5(4.9)	73.4(4.0)
	Quantile	82.4(2.2)	85.1(3.1)	92.8(4.2)	88.7(2.7)	71.1(3.9)
0.95N(0,1)+0.05Cauchy	Robust	86.0(7.4)	76.4(6.1)	87.4(5.1)	89.3(6.2)	94.5(5.7)
	Nonrobust	82.0(8.9)	83.9(10.6)	88.4(5.5)	80.0(11.6)	91.2(7.2)
	Stute	62.2(4.6)	64.6(5.9)	68.6(5.0)	65.4(4.1)	71.5(4.6)
	Quantile	82.3(3.4)	85.9(2.3)	90.6(2.1)	88.4(1.9)	70.9(6.6)
0.85N(0,1)+0.15Cauchy	Robust	85.4(6.8)	93.4(7.2)	88.3(7.2)	88.1(5.5)	92.4(6.2)
	Nonrobust	78.4(10.5)	83.1(11.4)	84.8(8.2)	75.5(10.8)	83.0(12.5)
	Stute	65.8(7.0)	72.2(6.2)	67.5(6.7)	61.2(5.1)	70.4(7.3)
	Quantile	82.1(3.1)	85.7(1.9)	92.5(2.3)	87.6(3.1)	69.3(5.8)
0.7N(0,1)+0.3Cauchy	Robust	88.0(7.4)	87.5(7.5)	90.7(5.5)	92.8(6.1)	82.7(6.4)
	Nonrobust	72.7(10.4)	74.4(14)	89.3(7.5)	92.5(8.9)	74.4(10.5)
	Stute	58.2(6.2)	64.9(7.8)	68.6(5.0)	63.0(4.9)	62.4(6.6)
	Quantile	82.3(2.6)	86.3(2.7)	92.4(3.1)	88.6(2.1)	70.6(3.4)
0.95N(0,1)+0.05t(3)	Robust	73.4(5.8)	87.7(6.5)	92.5(5.1)	78.8(5.6)	90.6(6.3)
	Nonrobust	78.9(7.5)	85.8(9.9)	87.6(10.0)	80.7(9.4)	89.0(8.4)
	Stute	66.2(4.8)	68.9(5.5)	68.6(5.0)	68.3(5.1)	68.9(4.2)
	Quantile	80.4(3.1)	83.1(4.1)	89.9(2.4)	86.6(4.1)	69.4(7.0)
0.85N(0,1)+0.15t(3)	Robust	83.0(7.5)	84.7(8.2)	82.6(5.8)	79.4(7.0)	86.4(6.2)
	Nonrobust	74.5(9.8)	76.6(11.2)	74.7(12.5)	76.4(13.6)	77.5(10.0)
	Stute	57.2(5.3)	64.6(6.2)	68.6(5.0)	67.6(5.4)	65.8(6.1)
	Quantile	81.6(3.8)	83.9(3.3)	82.3(2.4)	86.6(2.2)	69.1(4.9)
0.7N(0,1)+0.3t(3)	Robust	85.2(7.3)	90.0(6.7)	85.8(5.6)	94.8(5.3)	81.5(5.9)
	Nonrobust	75.6(11.4)	75.2(10.3)	81.8(6.4)	83.0(5.3)	74.7(11.3)
	Stute	57.2(6.2)	64.8(6.6)	68.6(5.0)	63.3(7.4)	62.0(5.9)
	Quantile	79.7(2.7)	83.6(1.6)	91.9(2.3)	87.9(1.9)	68.3(6.1)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 3

Analysis of the lung cancer data using the proposed method: estimates × 100. For interactions, values in "()" are the stability results.

	Main effects					Interactions			
	Age	**Gender**	**Intensity**	**Status**	**Gene**	**Age**	**Gender**	**Intensity**	**Status**
AP1S2	−1.0	−22.6	−0.1	37.4	20.8			−0.2(0.998)	
BTD	0.8	−8.6	0.2	5.1	−56.7				15.9(1.000)
C10ORF54	0.7	2.7	0.2	6.2	3.2			−0.5(0.998)	
CA5BP1	−1.2	−6.3	0.1	29.0	7.5		14.8(1.000)		
CAPN1	−3.5	−8.9	0.3	−8.6	−50.0		−24.7(0.000)		
CEBPB	−1.0	−15.5	0.0	29.4	−41.6				8.1(0.998)
EFNA1	−0.7	5.9	1.0	4.0	−65.3				12.7(1.000)
FAM107B	−1.2	−19.3	−0.1	16.2	29.2			−0.8(1.000)	
FLRT3	−1.2	24.6	0.8	6.1	33.4			−1.5(0.998)	
KATNB1	−0.0	0.5	0.2	26.6	−76.2		39.8(0.995)	30.6(0.059)	
LRRC1	−3.6	−13.6	−0.3	64.5	25.5		−24.3(0.995)		
LYRM5	0.8	−10.2	0.1	4.4	−12.0		−6.8(1.000)		
AP1S2.1	−1.0	−22.6	−0.1	37.4	20.8			−0.2(0.998)	
MYO18A	1.5	−15.9	0.1	−7.8	16.3			0.2(1.000)	
NOD1	−1.3	16.0	−0.1	4.6	−28.9			−0.3(0.998)	
NPLOC4	−3.3	−34.8	0.7	58.4	−76.3				14.4(0.998)
PLEKHO2	0.2	11.2	0.1	1.1	−15.3			−0.5(0.998)	
POLR3GL	−4.3	−1.7	0.0	63.0	28.2			0.1(0.998)	
RAB27A	0.6	−12.9	0.0	10.1	14.4			−0.5(0.998)	
SECISBP2L	−3.7	−34.7	−0.5	55.3	30.4			−0.8(0.998)	
SGTB	2.0	−17.5	0.2	3.3	7.4			0.2(0.995)	
STRADB	−0.7	−3.4	0.1	4.1	1.2		−48.5(0.995)		
SWSAP1	1.9	−14.6	0.1	−2.9	−12.1				0.8(0.995)
TEP1	1.7	−33.3	0.3	−0.9	5.4			0.3(1.000)	
TERF1	0.2	−22.0	−0.0	43.2	46.4				−18.9(1.000)
THOC1	0.9	30.9	1.2	−0.8	31.4		−67.4(0.002)		
TIGD5	1.5	−11.6	1.1	−32.5	3.9				−10.4(1.000)
TK2	−1.0	−13.8	−0.4	25.0	24.7			−0.8(1.000)	
TMEM54	3.7	−40.0	0.5	−33.2	−6.1		−69.2(1.000)		
TMEM106A	−0.8	−11.2	0.2	20.4	4.1			−0.3(0.998)	
TOMM7	1.8	0.1	0.1	−7.7	−21.4				2.3(0.998)
TRIM34	0.4	5.5	0.2	2.1	−13.4			−0.3(0.998)	
YARS2	0.5	8.1	0.9	−9.7	−5.8				−2.5(0.998)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 4

Summary analysis results for the lung cancer data using different methods. Diagonals are the numbers of identified genes using different methods. Off-diagonals are the numbers of overlap-ping genes. In "()" are the numbers of overlapping interactions.

	Robust	Nonrobust	Stute	Quantile
Robust	33	9(5)	0(0)	0(0)
Nonrobust	-	31	3(0)	1(0)
Stute	-	-	30	2(1)
Quantile	-	-	-	28

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 5

Analysis of the lung cancer data using the Nonrobust method: estimates × 100.

	Main effect					Interaction			
	Age	**Gender**	**Intensity**	**Status**	**Gene**	**Age**	**Gender**	**Intensity**	**Status**
BCL10	5.5	79.8	0.3	105.6	10.3				−29.2
BTD	6.1	38.4	0.9	91.5	−187.4				94.8
CAPN1	6.1	77.3	−0.0	97.9	22.1				−36.2
CASK	6.4	59.1	0.4	81.9	90.5				−63.1
CSNK1G2	5.9	31.0	0.5	98.0	−20.5		−70.4		
DOCK6	5.7	100.0	0.4	85.6	6.5				−35.4
ECI2	7.3	42.0	1.1	39.6	−87.0				57.3
ELMO3	5.1	88.5	0.1	122.5	−34.5				−12.1
FAM83H	5.6	57.3	0.5	116.1	113.1				−72.0
FASN	6.0	53.2	0.7	93.0	117.7				−78.4
KATNB1	6.2	59.6	0.4	85.0	55.6				−49.1
LYRM5	7.4	35.4	0.9	40.4	−43.4		12.2	32.8	
MACROD1	5.8	73.2	0.3	99.1	73.1				−58.4
NACC2	6.1	65.9	0.3	92.6	100.7				−70.9
PKP3	5.8	73.9	0.1	99.0	43.3				−54.8
RBFA	5.8	62.4	0.1	104.4	10.9		−66.1		
RNH1	6.0	70.6	0.5	79.8	14.9				−32.7
SCYL1	5.5	84.1	0.4	105.6	−36.2				−6.3
STRADB	6.1	59.7	0.9	76.1	−108.6				62.6
SWSAP1	5.8	69.7	0.4	92.9	18.0				−30.6
TEN1	5.6	66.3	0.6	104.1	21.9				−34.3
TIGD5	5.5	46.3	0.8	113.7	125.3				−73.9
TMEM54	6.1	54.6	−0.0	101.4	−1.3		−81.0		
TNIP2	5.1	85.9	0.6	100.9	−50.6				−4.2
TOLLIP	6.3	68.2	0.5	73.2	39.8				−48.4
TTC22	5.6	72.3	0.1	114.0	−20.3		−51.8		
WASH2P	5.7	25.2	0.3	109.1	19.9		−129.7		
YARS2	8.1	18.8	1.1	13.9	−41.6				26.5
YIPF2	5.7	75.2	0.4	99.8	7.8				−27.4
ZNF512B	6.3	54.7	0.4	89.8	94.9		−45.3	−49.6	
ZNF699	5.8	93.7	0.5	86.0	59.1				−56.4

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 6

Analysis of the lung cancer data using the Stute method: estimates × 100.

	Main effects					Interactions			
	Age	**Gender**	**intensity**	**Status**	**Gene**	**Age**	**Gender**	**Intensity**	**Status**
KLHL9	8.1	17.9	1.0	29.2	−358.5	5.4			
TPD52L2	7.5	56.8	1.0	31.2	313.5	−4.7			
TMEM129	7.7	21.9	1.1	28.5	−522.3	7.4			
PCGF3	8.2	20.2	1.1	15.9	−527.4	7.6			
XPNPEP1	7.8	30.0	0.9	26.1	−544.6	7.9			
FDPS	7.6	24.8	1.1	37.0	254.2	−3.8			
GFOD2	7.3	24.7	1.0	40.8	−447.1	6.1			
MAGED4B	7.3	23.3	1.0	44.8	−654.5	9.2			
PIGG	8.0	11.6	1.0	27.6	−354.2	5.1			
UVSSA	8.3	21.3	1.0	17.1	−527.7	7.7			
LAMTOR2	7.4	37.0	1.0	36.8	263.9	−3.7			
CSNK1G2	7.2	26.4	0.9	44.5	−390.9	5.2			
POLR3C	7.7	24.1	1.2	26.7	230.6	−3.3			
CTBP1	7.9	17.1	1.1	24.4	−361.4	5.2			
PRUNE	7.6	26.2	1.1	33.6	258.6	−3.7			
NELFA	7.6	28.8	1.0	32.4	−464.8	6.6			
MAEA	7.8	14.9	1.0	33.5	−418.3	6.0			
RPS27A	7.7	33.9	1.0	27.1	401.0	−5.8			
TBC1D14	7.6	21.4	0.9	34.0	−510.9	7.2			
MAN2B2	7.7	27.7	1.1	21.9	−481.9	6.8			
DGKQ	7.3	31.0	1.1	36.0	−552.6	7.7			
RBFA	7.6	−0.7	0.9	39.7	−565.2	7.8			
ACOX3	7.7	19.0	1.2	20.3	−595.1	8.2			
TCF25	7.6	27.0	1.1	33.5	−578.8	8.5			
PIGC	7.6	27.7	1.1	32.6	190.4	−2.7			
TOLLIP	7.2	25.6	1.0	38.7	−393.2	4.9			
CREG1	7.4	28.9	1.0	44.0	198.6	−2.9			
RAB11B	7.1	30.2	1.1	46.4	−342.1	4.7			
CDKN2AIP	7.8	40.2	0.9	32.3	−576.4	8.5			
GAK	7.9	31.3	1.1	19.4	−452.0	6.5			

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 7

Analysis of the lung cancer data using the Quantile method: estimates × 100.

	Main effects					Interactions			
	Age	**Gender**	**Intensity**	**Status**	**Gene**	**Age**	**Gender**	**Intensity**	**Status**
ARHGAP1	8.0	−18.2	0.8	16.1	−338.5	5.1			
B3GALT4	7.9	−8.1	1.2	14.1	−58.9			1.5	
BCL11A	8.1	−12.8	0.8	16.3	9.8				
CDKN1A	8.1	−5.3	0.7	16.9	−193.6	2.5			
CHIC2	8.3	−0.3	0.9	−1.0	−530.8	7.8	127.8		
CHST10	8.1	−7.3	0.8	16.6	18.9				
DNM2	8.0	−5.0	0.9	13.5	−245.4	3.4			
FAM65A	7.8	5.0	0.8	18.5	−220.7	2.7			
GATAD1	8.1	−4.2	0.9	11.8	−42.0				19.6
GGT5	8.2	−14.7	0.8	14.2	−152.4	1.7			
GRB7	8.0	−5.1	0.6	24.0	−1.1	66.4			
HIST1H4D	8.0	−13.5	0.9	16.7	13.0				
LAS1L	7.8	−2.1	0.8	22.9	−12.8				
LETM1	7.8	−3.4	1.1	15.6	−54.2			1.1	
MAEA	7.8	−3.5	1.2	13.2	−36.1			0.6	
PAPOLG	8.0	−16.6	0.8	12.5	16.0				
PDGFA	7.9	2.5	0.8	23.1	−240.6	3.5			
PDGFRA	8.0	5.7	0.6	21.1	−384.8	4.9	83.9		
PPARD	8.1	−5.6	0.8	13.7	−270.3	3.8			
PTAFR	8.2	−7.0	0.8	11.6	−258.1	3.7			
RXRB	7.8	−5.2	1.4	7.1	−71.6			1.3	
SCARNA9	8.1	−9.5	0.7	16.7	23.7				
SLC12A4	8.0	−11.2	0.8	18.3	−224.0	2.7			
TMEM204	8.1	−2.0	0.8	11.5	−317.9	4.2			
TOLLIP	8.0	−6.3	0.9	11.7	−336.9	4.8			
TOMM5	8.0	−6.9	0.8	17.8	25.2				
ZNF141	7.5	24.8	1.4	8.0	−105.4	65.8		1.3	
ZNF761	8.0	−19.2	1.0	17.0	−37.4	83.0			