Published in final edited form as:

Stat Med. 2018 February 10; 37(3): 437–456. doi:10.1002/sim.7518.

Dissecting gene-environment interactions: a penalized robust approach accounting for hierarchical structures

Cen Wu^{a,*}, Yu Jiang^b, Jie Ren^a, Yuehua Cui^c, and Shuangge Ma^d

^aDepartment of Statistics, Kansas State University, Manhattan, KS 66506

^bDepartment of Epidemiology, Biostatistics, and Environmental Health, University of Memphis, Memphis, TN 38111

^cDepartment of Statistics and Probability, Michigan State University, 619 Red Cedar Rd, East Lansing, MI 48824

^dDepartment of Biostatistics, Yale University, 60 College Street, New Haven, CT 06520

Abstract

Identification of gene–environment (G×E) interactions associated with disease phenotypes has posed a great challenge in high-throughput cancer studies. The existing marginal identification methods have suffered from not being able to accommodate the joint effects of a large number of genetic variants, while some of the joint-effect methods have been limited by failing to respect the “main effects, interactions” hierarchy, by ignoring data contamination, and by using inefficient selection techniques under complex structural sparsity. In this article, we develop an effective penalization approach to identify important G×E interactions and main effects which can account for the hierarchical structures of the two types of effects. Possible data contamination is accommodated by adopting the least absolute deviation (LAD) loss function. The advantage of the proposed approach over the alternatives is convincingly demonstrated in both simulation and a case study on lung cancer prognosis with gene expression measurements and clinical covariates under the AFT (accelerated failure time) model.

Keywords

Gene–environment interactions; Robust estimation; Penalized selection; Prognosis

1. Introduction

In high-throughput profiling studies, the significance of gene–environment (G×E) interactions in elucidating the genetic basis of complex diseases has been increasingly recognized. The identification of important G×E interactions is challenging due to the high dimensionality of G factors (such as SNPs, gene expressions, methylation and other types of genetic and genomic measurements) generated from profiling studies. Multiple statistical approaches have been proposed to pinpoint important G×E interactions associated with the

*Correspondence to: Cen Wu, Department of Statistics, Kansas State University, Manhattan, KS 66506.

†wucen@ksu.edu; Tel: 785-532-2231; Fax: 785-532-7736

risk and progression of complex diseases such as cancer, type 2 diabetes and cardiovascular diseases (Hutter et al. [1]).

The existing studies on $G \times E$ interactions, although quite successful, still have limitations. A family of methods focus on the marginal effects of interactions, as reviewed in Murcray et al. [2] and Cornelis et al. [3]. Although having simplicity and low computational cost, such methods are in conflict with the fact that the development of complex diseases is attributable to the coordinated effects of a large number of G and E main effects and their interactions. Another limitation is the lack of robust properties which are essential to accommodate data contamination. In, for example, cancer genetic studies, the survival times have often been observed to have heavy-tailed distributions. In Figure 1 (Appendix), we show a plot of the heavy-tailed distribution of the survival times of non-small-cell lung cancer patients from the study of Xie et al. [4]. Multiple factors may lead to such a phenomenon. For instance, misclassification of the cause of death can result in contamination in disease-specific survival (Rampatige et al. [5]). In addition, survival times taken from medical records may not always be reliable (Fall et al. [6] and Bowman [7]). Failing to take robustness into account leads to biased estimation even in the presence of a single contaminated observation (Huber and Ronchetti [8]) and thus false identifications of $G \times E$ interactions.

Despite their popularity in low-dimensional biomedical studies, robust statistical methods are relatively underdeveloped in the joint analysis of high-dimensional omics data (Wu and Ma [9]). Most robust objective functions are non-differentiable and thus cannot benefit from the prevailing gradient-based optimization methods. Furthermore, for $G \times E$ interactions, as the hierarchy between main and interaction effects has resurfaced as a special structure with a significant advantage over the alternatives (Bien et al. [10]), accommodating such a complex structure in robust methods poses extreme difficulties under high-dimensional settings. To identify important $G \times E$ interactions robustly, Shi et al. [11] and Wu et al. [12] recently develop rank-based penalized estimation approaches. Shi et al. [11] focuses on analyzing the marginal effects, while Wu et al. [12] can handle multiple main and interaction effects simultaneously. It is worth noting that both approaches do not respect the hierarchy. Imposing hierarchical structures with robust approaches is a tempting yet challenging task.

Motivated by the aforementioned limitations in current studies, we develop a hierarchical robust penalization approach to dissect $G \times E$ interactions in prognosis studies with high dimensional omics measurements. The proposed approach distinguishes itself from published ones along the following directions. First, the least absolute deviation (LAD) loss function is adopted to accommodate contamination. The LAD loss is a special case of the check loss and can be easily extended to quantile-based losses. Although no robust approach dominantly outperforms the others, quantile regression based approaches have been perhaps the most extensively investigated for high-dimensional data. As reviewed in Wu and Ma [9], most current quantile-based methods, as well as other robust methods, are focused on main effects and have not been extended to interactions yet. This study, with a focus on interactions, is thus warranted beyond the existing studies. Secondly, as our objective is to identify $G \times E$ interactions with prognosis data which are widely available in cancer genetic and other studies, we propose a penalized weighted LAD estimator under the accelerated failure time (AFT) model using Kaplan-Meier weights. The unpenalized weighted LAD

estimator has been examined in Huang et al. [13], however, the penalized version has not been well investigated, especially under the present complicated data and model settings. Furthermore, we develop an effective coordinate descent based algorithm which enjoys computational simplicity and is especially suitable for robust methods under high-dimensional settings. In contrast, for example, the robust identification of G×E interactions proposed in Shi et al. [11] is computationally expensive even though it is a marginal method. In summary, the proposed approach provides a useful addition to the existing studies.

2. Statistical Methods

We adopt the accelerated failure time (AFT) model for prognosis. Compared to popular alternatives such as the Cox model, the AFT model has a low computational cost and is thus more suitable for high dimensional omics data. Denote T as the logarithm of survival time and C as the logarithm of censoring time. Let $X = (X_1, \dots, X_p)$ be the p genetic measurements, and $E = (E_1, \dots, E_q)$ be the q environmental/clinical risk factors. The AFT model assumes that

$$T = \sum_{l=1}^q \alpha_l E_l + \sum_{m=1}^p \gamma_m X_m + \sum_{m=1}^p \sum_{l=1}^q \beta_{ml} X_m E_l + \varepsilon = \sum_{l=1}^q \alpha_l E_l + \sum_{m=1}^p (\gamma_m X_m + \sum_{l=1}^q \beta_{ml} X_m E_l) + \varepsilon = \sum_{l=1}^q \alpha_l E_l + \sum_{m=1}^p U_m^T b_m + \varepsilon, \quad (1)$$

where ε is the random error with an unknown distribution function. Note that we assume data has been properly normalized so that the intercept can be omitted. $b_m = (\gamma_m, \beta_{m1}, \dots, \beta_{mq})^T \equiv (b_{m1}, b_{m2}, \dots, b_{m(q+1)})^T$ and $U_m = (X_m, X_m E_1, \dots, X_m E_q)^T$. Denote $a = (a_1, \dots, a_q)^T$, $b = (b_1^T, \dots, b_p^T)^T$, and $U = (U_1^T, \dots, U_p^T)^T$. b_m and U_m represent all effects – main and interactions of the m th genetic measurement.

We use subscript i to denote the i th subject for the n independent subjects. Under right censoring, we observe $(Y_i, \delta_i, E_{ih}, U_{im})$ where $Y_i = \min(T_i, C_i)$, $\delta_i = 1\{T_i < C_i\}$ and (E_{ih}, U_{im}) are the covariates associated with Y_i . Without loss of generality, assume that $(Y_i, \delta_i, E_{ih}, U_{im})$'s have been ordered according to Y_i 's in an ascending order.

2.1. Robust Hierarchical Penalized identification

We use the Kaplan–Meier weights to accommodate censoring. Following Stute [14], let $\hat{F}_n$ be the Kaplan–Meier estimator of the distribution function F of T , then $\hat{F}_n$ can be expressed as $\hat{F}_n(y) = \sum_{i=1}^n d_{ni} 1\{Y_i \leq y\}$, with the Kaplan–Meier weights d_{ni} defined as

$$d_{n1} = \frac{\delta_1}{n}, d_{ni} = \frac{\delta_i}{n - i + 1} \prod_{j=1}^{i-1} \left(\frac{n - j}{n - j + 1} \right)^{\delta_j}, i = 2, \dots, n.$$

To accommodate contamination in the survival outcome, we propose the weighted LAD loss function

$$L(\alpha, b) = \sum_{i=1}^n d_{ni} |Y_i - \sum_{l=1}^q \alpha_l E_{il} - \sum_{m=1}^p b_m^T U_{im}|. \quad (2)$$

With the LAD loss, we can achieve the much desired robustness property which is not shared by, for example, the least square (LS) loss. It is noted that there may be other ways of robust estimation under the AFT model. Coupling the Kaplan-Meier weights and LAD loss has been highly motivated by the simple form, which leads to a low computational cost. The weights d'_{ni} s are the jumps of the Kaplan-Meier estimator $\hat{F}_n$. Another weighting method

adopts $d_{ni} = \frac{\delta_i}{(1 - \hat{G}_n(Y_i))}$ (Zhou [15] and others), where $\hat{G}$ is the Kaplan-Meier estimator of the censoring distribution G . It can be shown that the two seemingly different weighting schemes are equivalent (Huang et al. [13]).

Different from the existing robust studies which mainly focus on main effects (Wu and Ma [9]), here the interest is more on interactions. In $G \times E$ studies, there are two types of main effects, genetic effects and environmental effects, and one type of interaction effect, the $G \times E$ interactions. A critical issue for interaction analysis is to take the relationship, or more specifically, the hierarchical structures between the main and interaction effects into considerations. The hierarchy reinforces that once an interaction is identified, the corresponding main effects must be identified. There are two types of hierarchy. A weak hierarchical structure postulates that an interaction effect can be selected as long as at least one of the two corresponding main effects is selected. In contrast, under the strong hierarchical structure, the interaction can be identified only when both main effects are identified. Bien et al. [10] elaborates why the strong hierarchy is favored in $G \times G$ studies and demonstrates that violation of the strong hierarchy amounts to “postulating a special position for the origin” (McCullagh and Nelder [16]). In $G \times E$ studies, weak hierarchy is less attractive since environmental/clinical factors may be always present due to their importance and low dimensionality (Liu et al. [17]). Thus, we focus on the strong hierarchical structure which can be expressed by imposing:

$$\beta_{ml} \neq 0 \rightarrow \gamma_m \neq 0 \text{ and } \alpha_l \neq 0, \quad \forall m, l, \quad (3)$$

to model (1), and propose a robust hierarchical method for $G \times E$ interactions, termed as LAD-hier.

We propose the penalized robust objective function

$$Q_1(\alpha, b) = L(\alpha, b) + \lambda_1 \sum_{m=1}^p w_m \|b_m\|_2 + \lambda_2 \sum_{m=1}^p \sum_{l=2}^{q+1} w_{m,l} |b_{ml}|, \quad (4)$$

where w_m and $w_{m,l}$ are the adaptive weights with respect to group and individual level penalties. λ_1 and λ_2 are tuning parameters. b_m is the coefficient vector associated with the m th genetic measurement, with the first component corresponding to the main effect and the rest corresponding to the G×E interactions. To determine if the m th G measurement has any effect, we penalize the norm of b_m as a group using an adaptive group LASSO penalty. A nonzero b_m indicates the presence of the main genetic effect, or interaction effects with the environment factors, or both types of effects. The individual level penalty in (4) further determines the presence of specific interaction terms. Therefore, important main and interaction effects can be identified by the sum of the group and individual level penalty terms. We impose penalty on both group and individual levels to select important G×E interaction effects as well as main genetic effects. However, the main genetic effects are only penalized on the group level, while the interaction effects are penalized on both group and individual levels. So once at least one interaction effect is identified and has a nonzero effect, the main effect will automatically enter the model and so the strong hierarchy is respected.

Regularized variable selection has been adopted and investigated widely for high dimensional omics data. Due to the structured sparsity induced by the strong hierarchy between main and interaction effects, simple LASSO and group LASSO type of penalty functions are not adequate. Although the penalty terms of the proposed approach share a similar spirit with the sparse group LASSO [18], they are significantly different in that our approach is robust with the regularized LAD loss function and is designed for prognosis outcomes, which lead to a completely different estimation procedure. Furthermore, the proposed penalty naturally respects the strong hierarchy. We note that there can be alternative penalties that are also able to respect the hierarchy, including for example the hierarchical LASSO type penalties [10]. However, they are computationally very expensive and have not been applied to robust loss functions.

Under the penalization framework, the penalized loss functions of robust methods share a common form of “robust loss function + penalty function”. It is conjectured that the proposed robust hierarchical penalization can be potentially achieved by coupling robust loss functions (Wu and Ma [9]) with the hierarchical sparse group penalty. In addition to LASSO, the SCAD or MCP can also be used as a base penalty function. For example, a possible combination is the exponential squared loss in Wang et al. [19] and the MCP based hierarchical sparse group penalty. Since this combination demands additional tuning to select the robust tuning parameter in the exponential squared loss and smooth approximations for the nonconvex MCP, the computational cost is prohibitively high.

2.2. Computation

We develop a group coordinate descent (GCD) algorithm. In the GCD algorithm, the loss function is minimized with respect to one predictor group at a time, and the minimization is iteratively cycled through all the predictor groups until convergence is achieved.

To improve computational efficiency, at the k th iteration, we consider approximating $\|b_m\|_2$ in (4) by

$$\|b_m\|_2 \approx \|b_m^{(k-1)}\|_2 + \|b_m^{(k-1)}\|_2^{-1} |b_m^{(k-1)}|^T (|b_m| - |b_m^{(k-1)}|) = \|b_m^{(k-1)}\|_2^{-1} \sum_{l=1}^{q+1} |b_{ml}^{(k-1)}| |b_{ml}|,$$

where $b_m^{(k-1)}$ is the estimate at the $(k-1)$ th iteration, and $|b_m| = (|b_{m1}|, |b_{m2}|, \dots, |b_{m(q+1)}|)^T$.

With the above approximation, function (4) becomes

$$\begin{aligned} L(\alpha, b) &+ \lambda_1 \sum_{m=1}^p w_m \|b_m^{(k-1)}\|_2^{-1} \sum_{l=1}^{q+1} |b_{ml}^{(k-1)}| |b_{ml}| + \lambda_2 \sum_{m=1}^p \sum_{l=2}^{q+1} w_{m,l} |b_{ml}| \\ &= L(\alpha, b) + \sum_{m=1}^p \left\{ \lambda_1 w_m \|b_m^{(k-1)}\|_2^{-1} \sum_{l=1}^{q+1} |b_{ml}^{(k-1)}| |b_{ml}| + \sum_{l=2}^{q+1} \lambda_2 w_{m,l} |b_{ml}| \right\} \\ &= L(\alpha, b) + \sum_{m=1}^p \sum_{l=1}^{q+1} c_{ml} |b_{ml}|, \end{aligned}$$

where $w_m = \|\tilde{b}_m\|_2^{-2}$, and $w_{ml} = |\tilde{b}_{m,l}|^{-1}$ with the initial estimate $\tilde{b}_m$, and

$$c_{ml} = \begin{cases} \lambda_1 w_m \|b_m^{(k-1)}\|_2^{-1} |b_{ml}^{(k-1)}|, & l=1, \\ \lambda_1 w_m \|b_m^{(k-1)}\|_2^{-1} |b_{ml}^{(k-1)}| + \lambda_2 w_{ml}, & l=2, \dots, q+1. \end{cases}$$

At the k th iteration, given α and b_m ($m = 1, \dots, j-1, j+1, \dots, p$) fixed at the current estimates, optimization with respect to b_j can subsequently be formulated as a linear programming problem after introducing slack variables:

$$\underset{\xi, b_j}{\text{minimize}} \quad \frac{1}{n} \sum_{i=1}^n (\xi_i^+ + \xi_i^-) + c_{jl} \sum_{l=1}^{q+1} (b_{jl}^+ + b_{jl}^-),$$

$$\text{subject to } \xi_i^+ - \xi_i^- = d_{ni} (Y_i - \sum_{l=1}^q \alpha_l^{(k-1)} E_{il} - \sum_{m=1, m \neq j}^p b_m^{(k-1)T} U_{im} - b_j^T U_{ij}), \quad (5)$$

$$\text{where } \xi_i^+ \geq 0, \xi_i^- \geq 0; i=1, \dots, n,$$

and for an arbitrary variable z , $z^+ = z \mathbb{I}(z \geq 0)$ and $z^- = z \mathbb{I}(z < 0)$.

With fixed tunings, the GCD algorithm proceeds as follows.

- a. Initialize $\tilde{\alpha}$ and $\tilde{b}$ in (4) using LASSO.
- b. At the k th iteration, compute $b_j^{(k)}$ via (5) for $j = 1, \dots, p$.

- c. Update $a^{(k)}$ by minimizing (2) after fixing b as $b^{(k)}$.
- d. Iterate steps (b) and (c) until convergence.

The tuning parameter λ_1 controls group-level sparsity, while λ_2 determines sparsity within groups. We choose the tuning parameters based on independently generated testing datasets. The penalized estimates are obtained from the training data, then prediction performance is assessed on the testing data. A joint search over a two dimensional grid of (λ_1, λ_2) needs to be conducted to find the optimal pair of tuning parameters. In simulation, the tunings based on the independent testing data are very close to those using V-fold cross validation but are computationally less expensive. Such a strategy of using independent testing dataset to selected tuning parameters has been adopted in Wu and Liu [20], and many other studies. Alternatively, tuning parameters can also be selected based on information criterion and its variants. For example, in Tang et al. [21], a modified Schwarz-type information criterion is adopted. As the effective degree of freedom of the penalized regression model is not always clear, especially under complex model settings, the validation (or CV) based criterion can be a “safe” choice. In the case study, since the independent testing datasets are not available, we adopt cross validation to choose the optimal tuning parameters. We have observed that, in both simulation and case study, convergence is achieved in a small to moderate number of iterations. Research code written in R is publicly available at <https://github.com/cenwu>.

3. Simulation

We evaluate performance of the proposed approach, LAD-hier, through simulation study. We consider both continuous and categorical predictors, corresponding to gene expression and Single Nucleotide Polymorphism (SNP) data, respectively. For the gene expression data, we generate X from a multivariate normal distribution under the banded or auto-regressive correlation structure. Under the banded structure, the m th and l th genes have correlation coefficient 0.5 if $|m - l| = 1$, 0.33 if $|m - l| = 2$, and 0 otherwise, while the m th and l th genes have correlation $\rho^{|m-l|}$ under the auto-regressive correlation structure with $\rho=0.5$. In order to simulate SNP data, the values generated above are dichotomized at the first and third quartiles to generate categorical variables with three levels. Similar simulation strategies have been adopted in published studies such as Shi et al. [11] and Liu et al. [17]. We also generated 10 environmental factors from a multivariate normal distribution under the auto-regressive correlation structure. The number of G variables is 200, leading to 210 main effects and 2000 interaction effects in total. The sample size is 200. We consider the following data generating model

$$T = \sum_{l=1}^{10} \alpha_l E_l + \sum_{m=1}^{10} \gamma_m X_m + \sum_{m=1}^5 \sum_{l=1,3,5}^5 \beta_{ml} X_m E_l + \varepsilon \quad (6)$$

10 main effects and 15 interaction effects are simulated as associated with prognosis. We generate coefficients in model (6) from Unif(0.8, 1.2). The censoring rate is around 30%.

We consider right censored survival data under the model. The log event times are generated from the AFT model with the following error distributions: (1)Error 1: $N(0,1)$, (2)Error 2:

$0.80N(0, 1) + 0.20\text{LogNormal}(0, 1)$, (3) Error 3: $t(2)$, and (4) Error 4: $0.80N(0, 1) + 0.20\text{Cauchy}(0, 1)$.

Besides the proposed robust hierarchical approach (denoted as LAD-hier), we also consider the following alternatives for comparison. LAD-nhier is the robust penalization approach without considering the hierarchical structure. The penalized loss function of LAD-nhier is

$L(\alpha, b) + \lambda_1 \sum_{m=1}^p w_m \|b_m\|_2 + \lambda_2 \sum_{m=1}^p \sum_{l=1}^{q+1} w_{m,l} |b_{ml}|$. Here a penalty is imposed on each main and interaction effect. This approach can select important main and interaction effects, however, the hierarchy is not respected. LAD-group is also a robust penalization approach, with loss function $L(\alpha, b) + \lambda_1 \sum_{m=1}^p w_m \|b_m\|_2$. The group level penalty $\lambda_1 \sum_{m=1}^p w_m \|b_m\|_2$ does not yield sparsity within groups, that is, once a gene (or SNP) is selected, its main effect and all interaction effects are identified as associated with prognosis. In addition to the three robust approaches, we also consider the non-robust alternatives LS-hier, LS-nhier, and LS-group, where the LAD loss $L(\alpha, b)$ in the penalized loss functions of the robust approaches is replaced by the nonrobust least square loss. We acknowledge that multiple approaches can be adopted to analyze the simulated data. The five alternatives are the most relevant for comparison with a similar analysis framework.

The simulation results for gene expression data under the banded correlation structure are shown in Table 1. Without data contamination, performances of the proposed approach LAD-hier and the nonrobust counterpart LS-hier are comparable. With heavy tailed errors, LAD-hier significantly outperforms the alternatives. For example, under error 2, while roughly the same number of total true positives are identified by both approaches, LAD-hier identifies 15.8 false positive effects with standard deviation 9.9, which is much less than the false positives identified by LAD-nhier (22.6 with standard deviation 16.0). It is also worth noting that identification results by the robust methods have smaller standard deviations. Comparison with the robust non-hierarchical method LAD-nhier also shows the advantage of the proposed method. LAD-nhier identifies fewer true positive effects and more false positives. LAD-group and LS-group have the worst performance among all the approaches. Specifically, the lack of sparsity within-group leads to a much higher false positive rate. Under other settings (Table 3, 4 and 5 in Appendix), similar patterns are observed. To better visualize the identification results, we provide Figure 2 (Appendix) to show the ratios of number of true positives over false positives of all the approaches. The upper left panel shows that the bar corresponding to LAD-hier is higher than the bars of the rest approaches for gene expression data with banded structure in Table 1. It is clear to observe the advantage of LAD-hier under different scenarios, as shown in the ratio plot.

To further investigate tuning parameter selection, in Table 6 (Appendix), we compare the results of tuning based on independently generated testing data and CV (5-fold cross-validation) for one simulation scenario. We observe that the results are very close. Results under other settings are also close and omitted.

We have also compared the prediction performance in terms of least absolute error (or L1 error) for robust methods and squared error loss for non-robust methods. Note that it is not sensible to compare the results between robust and non-robust methods directly since the

prediction errors are calculated based on different criterions. Instead, we provide prediction errors for intercept models as a benchmark. The prediction performance on testing datasets for the LAD-hier and alternatives are given in Table 7,8,9 and 10 (Appendix). In general, LAD-hier has better prediction results. For example, in Table 7 (Appendix), under error 4, LAD-hier has a smaller least absolute error (LAE) (1.94, sd 0.29) than LAD-nhier (2.03, sd 0.36) and LAD-group (2.08, sd 0.29). We also note that the standard deviation of non-robust method is much larger than the prediction error in this setting. For example, LS-hier has a squared error loss of 1.11 with standard deviation 5.53.

Simulation indicates that the proposed robust hierarchical method is computationally feasible. Given fixed tunings, the analysis of one simulated replicate (in Table 1, error 3) takes 12.8(LAD-hier), 14.2(LAD-nhier), 27.1(LAD-group), 56.5(LS-hier), 57.4(LS-nhier) and 57.3(LS-hier) seconds on a desktop PC with standard configurations. In the simulation study, we consider $(n, q, p)=(200,10,200)$ for the sample size, dimensions of environmental factors and genetic factors. Here, we note that although $n = p$, the dimension of combined main and interaction effects is 2210, which is much larger than the sample size. Besides, heavy tailed errors and censoring also affect the identification results. With larger sample size, we can observe higher true positives and similar patterns of performances under current settings.

4. Case Study

Lung cancer is the leading cause of death from cancer worldwide for both women and men. As the most common kind of lung cancer, non-small-cell lung cancer (NSCLC) accounts for approximately 85% of lung cancer deaths and has subtypes such as adenocarcinoma and squamous cell carcinoma.

Gene profiling studies have been widely carried out for lung cancer in order to identify genetic markers associated with prognosis. In addition to the genetic factors, the development and progression of lung cancer are also critically dependent on environmental and clinical factors, as well as their interactions with the genetic factors. Here we apply the proposed method, along with the alternatives to the NSCLC data analyzed in Xie et al. [4]. To overcome issues resulting from a small sample size, we follow Xie et al. [4] and analyze data collected from four independent studies, namely the DFCI (Dana-Farber Cancer Institute), HLM (Moffitt Cancer Center), MSKCC (Memorial Sloan-Kettering Cancer Center), and UM (University of Michigan) study. The DFCI study has a total of 78 patients, with 35 deaths during follow-up and 51 months median follow-up. The HLM study has a total of 79 patients, with 60 deaths during follow-up. The median follow-up time is 39 months. The MSKCC study has a total of 102 patients, with 38 deaths during follow-up. The median follow-up time is 43.5 months. The UM study has a total of 92 patients, with 48 deaths during follow-up and 55 months median follow-up. More detailed information can be found from Xie et al. [4].

Gene expression measurements were generated from the Affymetrix U133 plus 2.0 arrays. A total of 22,283 probe sets are available for analysis for all the four datasets. The gene expression normalization is first performed for each dataset individually. We then conduct

across-dataset normalization to improve comparability. We rank the probe sets according to variations as genes with higher variations are often of more interest. The top 500 are screened out for downstream analysis. Each gene expression is subsequently standardized to have mean zero and standard deviation one.

We consider five environmental factors: age, gender, smoke, chemotherapy, and stage. We analyze data using robust approaches LAD-hier, and LAD-nhier and nonrobust LS-hier and LS-nhier. LAD-group and LS-group are not considered due to their inferior performance in simulation. As shown in Table 2, the proposed robust hierarchical method identifies 21 main gene effects and 10 interaction effects, with $-0.042(\text{age})$, $0.001(\text{gender})$, $-0.017(\text{smoke})$, $0.108(\text{chemotherapy})$, and $-0.295(\text{stage})$ as the estimates for the environmental/clinical main effects. These results are sensible. For instance, senior patients and later stage patients tend to have a shorter survival. Smoking and chemotherapy have detrimental and beneficial effects on survival, respectively. Approach LS-nhier identifies 6 main effects and 25 interactions (Table 11 in the Appendix). The identification results with LS-hier and LS-nhier are provided in Table 12 and 13 in Appendix, respectively. The nonrobust methods select significantly different sets of main and interaction effects. In terms of environmental main effects, the estimate of smoking is 0.045 under LAD-nhier, while both LS-hier and LS-nhier generate positive age effects. These do not seem reasonable compared to the estimates generated using the proposed approach. Since independent testing datasets are not available, we use 5-fold cross validation to select tuning parameters. Given fixed tunings, the computational time for the 4 methods are 22.1 (LAD-hier), 26.9 (LAD-nhier), 55.8 (LS-hier) and 71.3 (LS-nhier) seconds, respectively.

To further compare all methods, we evaluate prediction performance in terms of logrank statistics. In survival analysis, the logrank test is perhaps the most commonly-adopted statistical test to compare survival distributions of different groups, for example, the low risk and high risk group of patients in cancer studies. Larger logrank test statistic indicate better predictive performance. The samples are randomly split into two risk groups with an equal sample size according to the predicted risk scores. The logrank statistics, measuring the survival difference of the two groups, are then calculated. The logrank statistics are 18.3 (LAD-hier, $p\text{-value}=1.9\times 10^{-5}$), 11.5 (LAD-nhier, $p\text{-value}=6.9\times 10^{-4}$), 4.2 (LS-hier, $p\text{-value}=0.042$) and 3.9 (LS-nhier, $p\text{-value}=0.047$), respectively. The proposed method has the largest logrank statistic hence the best prediction performance.

The proposed method identifies 21 main gene effects and 10 $G\times E$ interactions, with 7 $G\times \text{age}$ interactions, 1 $G\times \text{smoke}$ interaction, and 2 $G\times \text{stage}$ interactions. Gene PSPH interacts with both age and stage. PSPH encodes protein phosphoserine phosphatase and belongs to the phosphatase super family. Major functions of this protein include being responsible for L-serine formation, catalyzing hydrolysis of L-phosphoserine, and involving in exchange reaction between L-serine and L-phosphoserine. In a multicenter clinical study, an increased level of PSPH is characterized in lung cancer patients with a better response after erlotinib therapy (Tan et al. [23]). HMGA2, high-mobility-group AT-hook protein 2, has been reported to be over-expressed in lung cancer (Meyer et al. [24]). In addition, the over-expression of HMGA2 protein is shown to be positively correlated with tumor grade (Di Cello et al. [25]). We find that increased HMGA2 leads to a shorter survival. The interaction

between HMGA2 and stage shows the higher expression of HMGA2 and the more advanced tumor diagnostic stage, the shorter the survival time. The results are consistent with previous findings [25]. Gene CHL1 has a positive effect on lung cancer survival and an interaction with smoke. CHL1, also known as CALL or L1CAM2, is found to be a putative tumor suppressor gene. Decreasing of CHL1 expressions promotes tumor formation and growth in many types of cancers (Senchenko et al.[26] and He et al. [27]). Our results confirm the biological findings that an over-expression of CHL1 might have a protective effect in cancer development. One of the most interesting genes interacting with age is PON3. Paraoxonase-3 (PON3) is one type of intracellular enzymes and known to have antioxidative effects. Its effects on cancer have been reported to be controversial (Witte et al. [28]), and further investigation of the gene function with aging may better explain its role in cancer progression.

Among the 13 genes that only have main effects, SPINK1, AKR1C2, PLA2G4A, IGHD, and APOD are positively associated with survival. TFF1, CPS1, IFI27, ALDH3A1, HMGA2, PBK, CCL18, and AQP1 are negatively associated. Among the above genes, some have been identified as prognostic markers by other studies. For example, PBK, PDZ-binding kinase, has been shown to have a higher expression level in lung cancer tissues compared to normal tissues. Increasing of PBK links to worse survival (Lei et al. [29]). Overexpression of AQP1 has been found to be associated with a poor survival in lung adenocarcinoma (Yun et al. [30]). Lung cancer patients with a higher serum CCL18 level have a shorter survival (Plönes et al. [31]).

5. Discussion

Substantial efforts have been devoted to detect important G×G and G×E interactions in complex diseases (Clayton and McKeigue [32], Cordell [33],[34]). Statistical methods for interactions can be categorized as marginal and joint analysis (Hutter et al. [1]). Traditionally, G×E interactions have been pursued by assessing the genetic effect under environmental exposures, mostly through marginal testing methods. G×E interactions can be tested marginally by using a variety of models, from the classical linear model with interactions in different studies, such as case only study, case-control study and the two-stage screening study, to more complicated models, including non- and semi-parametric models, and Bayesian models (Murcray et al. [2]; Mukherjee et al. [35]). On the other hand, a joint analysis considers the combined joint effects of multiple genetic and environmental factors, as well as the interaction effects. Recently, penalized variable selections have attracted much attention as a representative joint analysis method for G×G and G×E interactions (Shi et al. [11], Wu et al. [12] and Liu et al. [17]). Furthermore, machine learning and data mining methods have been extensively investigated for interactions under complex diseases (Cordell [34] and McKinney et al. [36]). For example, the popular nonparametric and model free multifactor dimensionality reduction (MDR) method can efficiently identify G×G and G×E interactions (Moore [37]).

In this article, we have developed a robust regularized variable selection method accounting for the hierarchical structures in the identification of important gene-environment interactions for high-throughput biomedical studies. Significantly advancing from the

existing interaction studies, the proposed method is robust by adopting a LAD loss function which is coupled with a weight function to accommodate censoring. We impose both group- and individual-level penalties to respect the hierarchical structure. An approximation to the group-level penalty improves efficiency of the GCD algorithm. The developed method outperforms the robust alternatives that do not respect the hierarchy as well as the nonrobust alternatives in simulation. In the case study of lung cancer with gene expression measurements, the proposed method identifies interesting main and interaction effects and has better prediction performance.

We do realize that the simulation settings may be overly simplified compared to practically observed data. However, in simulation, we have comprehensively considered the relevant robust penalization approaches and the non-robust counterparts. Different approaches are compared on a fair ground, and the comparison clearly demonstrates the superiority of the proposed approach. Data analysis further complements simulations and establishes superiority of the proposed approach. In addition, high dimensional inference is especially challenging. Although there are a few existing inference methods, their validity under the present setting is unclear. In this study, we focus on selection and estimation and will postpone inference study to future research. In our study, environmental factors have “different status” from the genetic factors. They are usually pre-selected based on evidences of association with cancer from past studies. Selection of environmental factors is not of interest and they may always stay in the model. To the best of our knowledge, most of the existing machine learning and data mining methods for $G \times E$ interactions treat G and E main effects equally, and do not respect hierarchy structure between main and interaction effects, especially for survival outcomes. The proposed and alternative approaches considered in our study are the most relevant for a direct comparison within a similar framework for analysis.

This study has been partly motivated by analyzing high dimensional genetics data in cancer prognostic studies. Our robust hierarchical method has been developed to accommodate censoring. With other types of response variables, for example, a continuous phenotype, the Kaplan-Meier weights in the proposed method can simply be replaced by equal weights. Extensions to generalized linear and other models are nontrivial and demands further investigations. We have considered the LAD loss function with the Kaplan-Meier weights to achieve robustness to data contamination in the prognostic outcome. As pointed out by Wu and Ma [9], among the available robust regularized variable selection methods, high or ultrahigh dimensional data can be more readily accommodated by LAD or quantile based methods. Approximations are generally needed for other robust loss functions such as the rank-based loss. Some loss functions, for example, the exponential squared loss proposed in Wang et al. [19], need additional tunings. All these lead to a prohibitively high computational cost in high/ultra-high dimensions. Under such circumstances, robust regularization accommodating complex data structures with penalty functions beyond LASSO or its nonconvex counterparts is not computationally affordable. While the group and individual LASSO type penalties have been imposed in our approach to retain the hierarchy between main and interaction effects, it is conjectured that the nonconvex SCAD or MCP can also be coupled into the sparse group penalty structure after smooth approximations. The computational efficiency of the proposed approach comes from multiple sources. The loss function has an L1 form, which “matches” the Lasso type penalty

and simplifies computation. In addition, the nicely designed penalty “automatically” respect the hierarchical structure, which further simplifies computation. It has been broadly accepted in Fan and Lv [38], Song et al. [39] and studies alike that penalization methods need to be combined with screening strategy to accommodate ultra-high dimensional data such as the genome-wide data. The proposed method can be potentially implemented in a two stage screening-and-penalization framework for GWAS. We will investigate it in the follow-up studies. Overall, the proposed method has a solid statistics basis and achieves satisfactory performance in both simulation and real data analysis. To fully understand the identified main and interaction effects in the case study, more bioinformatics and functional studies are needed in the future.

Acknowledgments

We thank the associate editor and reviewers for their careful review and insightful comments, which have led to a significant improvement of this article. This study has been partly supported by the National Institutes of Health (CA191383, CA204120), the VA Cooperative Studies Program of the Department of VA, Office of Research and Development, an innovative research award from KSU Johnson Cancer Research Center and a KSU Faculty Enhancement Award.

References

1. Hutter CM, Mechanic LE, Chatterjee N, Kraft P, Gillanders EM. Gene-environment interactions in cancer epidemiology: a National Cancer Institute Think Tank report. *Genetic epidemiology*. 2013; 37(7):643–657. [PubMed: 24123198]
2. Murcray CE, Lewinger JP, Gauderman WJ. Gene-environment interaction in genome-wide association studies. *American Journal of Epidemiology*. 2009; 169(2):219–226. [PubMed: 19022827]
3. Cornelis MC, Tchetgen EJ, Liang L, Qi L, Chatterjee N, Hu FB, Kraft P. Gene-environment interactions in genome-wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes. *American Journal of Epidemiology*. 2012; 175(3):191–202. [PubMed: 22199026]
4. Xie Y, Xiao G, Coombes KR, Behrens C, Solis LM, et al. Robust gene expression signature from formalin-fixed paraffin-embedded samples predicts prognosis of non-small-cell lung cancer patients. *Clin Cancer Res*. 2011; 17(17):5705–5714. [PubMed: 21742808]
5. Rampatige R, Gamage S, Peiris S, Lopez AD. Assessing the reliability of causes of death reported by the vital registration system in sri lanka: medical records review in colombo. *HIM J*. 2013; 42(3): 20–28.
6. Fall K, Strömberg F, Rosell J, Andrn O, Varenhorst E. Reliability of death certificates in prostate cancer patients. *Scandinavian Journal of Urology and Nephrology*. 2008; 42(4):352–357. [PubMed: 18609293]
7. Bowman, L. Doctors, researchers worry about accuracy of social security ”death file”. 2011. <http://projects.scrippsnews.com/story/doctors-researchers-worry/>
8. Huber, PJ., Ronchetti, EM. *Robust Statistics*. 2. Hoboken, N.J.: Wiley; 2009. Gene-environment interactions in human diseases. Wiley series in probability and statistics
9. Wu C, Ma S. A selective review of robust variable selection with applications in bioinformatics. *Brief. Bioinform*. 2015; 16(5):873–883. [PubMed: 25479793]
10. Bien J, Taylor J, Tibshirani R. A LASSO for hierarchical interactions. *Ann. Stat*. 2015; 41:1111–1141.
11. Shi XJ, Liu J, Huang J, Zhou Y, Yang X, Ma S. A penalized robust method for identifying gene-environment interactions. *Genetic Epidemiology*. 2014; 38(3):220–230. [PubMed: 24616063]
12. Wu C, Shi XJ, Cui YH, Ma S. A penalized robust semiparametric approach for gene-environment interactions. *Statistics in Medicine*. 2015; 34(30):4016–4040. [PubMed: 26239060]

13. Huang J, Ma S, Xie H. Least absolute deviations estimation for the accelerated failure time model. *Statistica Sinica*. 2007; 17:1533–1548.
14. Stute W. Distributional convergence under random censorship when covariables are present. *Scandinavian Journal of Statistics*. 1996; 23:461–471.
15. Zhou M. M-estimation in censored linear models. *Biometrika*. 1992; 79(4):837–841.
16. McCullagh, P., Nelder, J. Generalized linear models. Chapman and Hall; London: 1983.
17. Liu J, Huang J, Zhang Y, Lan Q, Rothman N, Zheng T, Ma S. Identification of gene-environment interactions in cancer studies using penalization. *Genomics*. 2013; 102(4):189–194. [PubMed: 23994599]
18. Friedman J, Hastie T, Tibshirani R. A note on the group lasso and a sparse group lasso. arXiv preprint. 2010 arXiv:1001.0736.
19. Wang X, Jiang Y, Huang M, Zhang H. Robust variable selection with exponential squared loss. *Journal of the American Statistical Association*. 2013; 108(502):632–643. [PubMed: 23913996]
20. Wu Y, Liu Y. Variable selection in quantile regression. *Statistica Sinica*. 2009; 19(2):801–817.
21. Tang Y, Wang HJ, Zhu Z, Song X. A unified variable selection approach for varying coefficient models. *Statistica Sinica*. 2012; 22(2):601–628.
22. Cui Y, Kang G, Sun K, Qian M, Romero R, Fu W. Gene-centric genomewide association study via entropy. *Genetics*. 2008; 179(1):637–650. [PubMed: 18458106]
23. Tan EH, Ramlau R, Pluzanska A, Kuo HP, Reck M, Milanowski J, Au JK, Felip E, Yang PC, Damsyanov D, Orlov S. A multicentre phase II gene expression profiling study of putative relationships between tumour biomarkers and clinical response with erlotinib in non-small-cell lung cancer. *Annals of Oncology*. 2010; 21(2):217–222. [PubMed: 20110292]
24. Meyer B, Loeschke S, Schultze A, Weigel T, Sandkamp M, Goldmann T, Vollmer E, Bullerdiek J. HMGA2 overexpression in non-small cell lung cancer. *Molecular Carcinogenesis*. 2007; 46(7): 503–511. [PubMed: 17477356]
25. Di Cello F, Hillion J, Hristov A, Wood L, Mukherjee M, Schuldenfrei A, Kowalski J, Bhattacharya R, Ashfaq R, Resar LM. HMGA2 participates in transformation in human lung cancer. *Molecular Cancer Research*. 2008; 6(5):743–750. [PubMed: 18505920]
26. Senchenko V, Krasnov G, Dmitriev A, Kudryavtseva A, Anedchenko E, Braga E, Pronina I, Kondratieva T, Ivanov S, Zabarovskiy E, Lerman M. Differential expression of CHL1 gene during development of major human cancers. *PloS One*. 2011; 6(3):e15612. [PubMed: 21408220]
27. He LH, Ma Q, Shi YH, Ge J, Zhao HM, Li SF, Tong ZS. CHL1 is involved in human breast tumorigenesis and progression. *Biochemical and biophysical research communications*. 2013; 438(2):433–438. [PubMed: 23906755]
28. Witte I, Foerstermann U, Devarajan A, Reddy ST, Horke S. Protectors or traitors: the roles of PON2 and PON3 in atherosclerosis and cancer. *Journal of lipids*. 2012; doi: 10.1155/2012/342806
29. Lei B, Liu S, Qi W, Zhao Y, Li Y, Lin N, Xu X, Zhi C, Mei J, Yan Z, Wan L. PBK/TOPK expression in non-small-cell lung cancer: its correlation and prognostic significance with Ki67 and p53 expression. *Histopathology*. 2013; 63(5):696–703. [PubMed: 24025073]
30. Yun S, Sun PL, Jin Y, Kim H, Park E, Park SY, Lee K, Chung JH. Aquaporin 1 Is an Independent Marker of Poor Prognosis in Lung Adenocarcinoma. *Journal of pathology and translational medicine*. 2016; 50(4):251–257. [PubMed: 27271108]
31. Plönes T, Krohn A, Burger M, Veelken H, Passlick B, Müller-Quernheim J, Zissel G. Serum level of CC-chemokine ligand 18 is increased in patients with non-small-cell lung cancer and correlates with survival time in adenocarcinomas. *PLoS One*. 2012; 7(7):e41746. [PubMed: 22848587]
32. Clayton D, McKeigue PM. Epidemiological methods for studying genes and environmental factors in complex diseases. *The Lancet*. 2001; 358(9290):1356–1360.
33. Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human molecular genetics*. 2002; 11(20):2463–2468. [PubMed: 12351582]
34. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet*. 2009; 10(6):392–404. [PubMed: 19434077]

35. Mukherjee B, Ahn J, Gruber SB, Chatterjee N. Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. *American Journal of Epidemiology*. 2011; 175(3):177–190. [PubMed: 22199027]
36. McKinney BA, Reif DM, Ritchie MD, Moore JH. Machine learning for detecting gene-gene interactions. *Applied bioinformatics*. 2006; 5(2):77–88. [PubMed: 16722772]
37. Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, Barney N, White BC. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of Theoretical Biology*. 2006; 241(2):252–261. [PubMed: 16457852]
38. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*. 2008; 70(5):849–911.
39. Song R, Lu W, Ma S, Jessie Jeng X. Censored rank independence screening for high-dimensional survival data. *Biometrika*. 2014; 101(4):799–814. [PubMed: 25663709]

Appendix

This file contains additional simulation and data analysis results.

Figure 1.
Histogram of survival times from the study of Xie et al. (2011)

Table 3

Identification results for gene expression (GE) **data with AR structure**. In each cell, mean(sd) based on 100 replicates. TP/FP: total number of true/false positives. TP1/FP1: true/false positives for genetic main effects. TP2/FP2: true/false positives for interactions.

GE(AR)		TP	FP	TP1	FP1	TP2	FP2
Error 1	LAD-hier	12.8(3.2)	14.7(10.2)	6.8(1.5)	7.2(5.3)	6.0(1.9)	7.5(5.1)
	LAD-nhier	10.3(2.7)	25.2(7.3)	5.2(1.5)	2.8(1.9)	5.0(1.9)	22.4(6.8)
N(0,1)	LAD-group	17.4(3.7)	177.3(58.3)	6.9(1.5)	12.6(5.6)	10.5(2.7)	164.7(52.9)

GE(AR)		TP	FP	TP1	FP1	TP2	FP2
	LS-hier	11.8(5.1)	18.2(12.2)	5.4(2.1)	6.2(4.3)	6.4(3.5)	11.9(8.3)
	LS-nhier	12.2(5.2)	31.8(23.7)	3.7(2.6)	1.7(1.6)	8.5(3.5)	30.1(22.6)
	LS-group	13.9(6.9)	66.8(36.3)	4.8(2.2)	2.9(2.1)	9.1(4.9)	63.9(34.3)
Error 2	LAD-hier	12.8(3.8)	15.9(9.7)	6.4(1.9)	7.4(5.0)	6.4(2.4)	8.5(4.9)
	LAD-nhier	10.8(2.5)	28.7(7.4)	4.6(1.2)	3.3(1.9)	6.1(1.9)	25.4(6.8)
	LAD-group	16.2(5.6)	149.3(75.6)	6.1(2.3)	10.4(6.4)	10.1(3.6)	138.9(69.4)
80%N(0,1)	LS-hier	13.4(4.8)	26.2(26.7)	6.3(2.1)	7.9(5.6)	7.1(3.1)	18.3(22.4)
	LS-nhier	13.0(3.9)	34.0(30.4)	4.1(1.8)	2.0(3.5)	8.9(2.9)	32.0(27.3)
	LS-group	15.6(5.6)	95.2(50.2)	5.5(1.9)	4.7(3.8)	10.1(3.9)	90.4(46.5)
+20%LogN(0,1)	LAD-hier	11.9(4.4)	14.8(9.5)	6.0(2.0)	7.4(4.9)	5.9(2.8)	7.4(4.8)
	LAD-nhier	9.8(2.9)	28.9(8.9)	4.4(1.5)	3.9(2.3)	5.4(2.1)	25.0(8.1)
	LAD-group	15.3(4.4)	138.4(70.9)	5.5(1.8)	9.8(6.5)	9.8(3.1)	128.6(64.1)
Error 3	LS-hier	10.9(5.7)	17.9(12.3)	4.9(2.4)	5.9(4.1)	5.9(3.6)	12.0(8.7)
	LS-nhier	10.0(5.8)	20.9(16.7)	2.4(1.9)	0.7(1.0)	7.6(4.5)	20.2(16.1)
	LS-group	13.4(6.5)	78.1(46.7)	4.5(2.3)	3.8(2.8)	8.9(4.4)	74.3(44.1)
Error 4	LAD-hier	9.4(6.1)	12.2(9.3)	4.7(2.9)	5.8(4.5)	4.7(3.3)	6.4(5.1)
	LAD-nhier	7.6(4.1)	29.6(9.1)	3.5(2.1)	4.1(3.8)	4.1(2.4)	25.5(7.0)
	LAD-group	11.8(8.0)	110.8(92.8)	4.3(3.1)	7.7(7.4)	7.5(5.1)	103.1(85.6)
80%N(0,1)	LS-hier	8.2(7.4)	21.3(23.3)	3.6(3.2)	5.6(6.8)	4.6(4.3)	15.7(17.0)
	LS-nhier	8.5(7.7)	33.9(37.3)	2.7(2.8)	1.9(3.5)	5.8(5.2)	31.9(34.3)
	LS-group	10.1(8.3)	73.1(68.7)	3.4(2.9)	4.1(4.8)	6.7(5.5)	69.0(64.3)
+20%Cauchy(0,1)	LAD-hier	10.5(4.3)	16.9(10.1)	5.6(2.2)	8.6(5.6)	4.9(2.4)	8.3(4.9)
	LAD-nhier	8.3(2.9)	35.1(7.6)	3.6(1.4)	5.1(2.9)	4.7(2.2)	29.9(6.5)
	LAD-group	12.0(5.9)	90.2(59.7)	4.5(2.6)	6.6(5.3)	7.6(3.7)	83.6(54.8)
N(0,1)	LS-hier	13.1(4.7)	18.8(15.3)	5.6(2.1)	6.1(4.6)	7.5(3.1)	12.7(11.8)
	LS-nhier	11.7(4.5)	21.9(13.3)	2.8(1.8)	1.1(1.3)	8.9(3.4)	20.8(12.8)
	LS-group	15.3(5.4)	88.5(45.0)	5.0(2.1)	4.4(3.3)	10.3(3.7)	84.1(41.9)
Error 2	LAD-hier	12.2(4.3)	19.9(13.1)	6.0(2.1)	10.0(6.9)	6.2(2.4)	9.9(6.2)
	LAD-nhier	10.6(2.6)	33.8(7.7)	4.8(1.4)	4.2(1.8)	5.8(1.9)	29.6(7.1)
	LAD-group	15.7(4.9)	115.9(53.6)	5.6(2.1)	8.1(5.3)	10.1(3.2)	107.8(48.8)

Table 4

Identification results for SNP data with banded structure. In each cell, mean(sd) based on 100 replicates. TP/FP: total number of true/false positives. TP1/FP1: true/false positives for genetic main effects. TP2/FP2: true/false positives for interactions.

SNP(Banded)		TP	FP	TP1	FP1	TP2	FP2
Error 1	LAD-hier	10.5(4.3)	16.9(10.1)	5.6(2.2)	8.6(5.6)	4.9(2.4)	8.3(4.9)
	LAD-nhier	8.3(2.9)	35.1(7.6)	3.6(1.4)	5.1(2.9)	4.7(2.2)	29.9(6.5)
	LAD-group	12.0(5.9)	90.2(59.7)	4.5(2.6)	6.6(5.3)	7.6(3.7)	83.6(54.8)
N(0,1)	LS-hier	13.1(4.7)	18.8(15.3)	5.6(2.1)	6.1(4.6)	7.5(3.1)	12.7(11.8)
	LS-nhier	11.7(4.5)	21.9(13.3)	2.8(1.8)	1.1(1.3)	8.9(3.4)	20.8(12.8)
	LS-group	15.3(5.4)	88.5(45.0)	5.0(2.1)	4.4(3.3)	10.3(3.7)	84.1(41.9)
Error 2	LAD-hier	12.2(4.3)	19.9(13.1)	6.0(2.1)	10.0(6.9)	6.2(2.4)	9.9(6.2)
	LAD-nhier	10.6(2.6)	33.8(7.7)	4.8(1.4)	4.2(1.8)	5.8(1.9)	29.6(7.1)
	LAD-group	15.7(4.9)	115.9(53.6)	5.6(2.1)	8.1(5.3)	10.1(3.2)	107.8(48.8)

SNP(Banded)		TP	FP	TP1	FP1	TP2	FP2
+20%LogN(0,1)	LS-hier	12.4(7.0)	32.5(34.7)	5.2(2.9)	8.4(6.9)	7.4(4.3)	24.0(28.7)
	LS-nhier	13.4(5.4)	60.4(53.1)	3.5(2.1)	4.2(5.2)	9.8(3.9)	56.2(48.5)
	LS-group	14.4(6.4)	98.6(57.8)	4.8(2.4)	5.3(3.9)	9.6(4.3)	93.3(54.1)
Error 3 $\alpha(2)$	LAD-hier	10.9(5.1)	17.6(12.4)	5.5(2.6)	8.9(6.6)	5.4(2.8)	8.7(6.1)
	LAD-nhier	8.9(2.9)	34.6(8.6)	3.8(1.4)	4.4(2.8)	5.1(2.1)	30.2(7.4)
	LAD-group	13.7(5.5)	96.8(50.3)	5.1(2.1)	6.5(4.8)	8.6(3.6)	90.3(46.1)
	LS-hier	11.2(5.9)	28.7(27.8)	4.5(2.5)	7.0(5.8)	6.7(3.9)	21.6(22.7)
	LS-nhier	10.3(5.9)	44.1(43.4)	2.4(2.1)	2.5(4.7)	7.9(4.5)	41.6(39.4)
	LS-group	12.6(6.8)	84.4(64.1)	4.1(2.3)	4.6(4.6)	8.5(4.6)	79.8(59.6)
Error 4	LAD-hier	10.3(5.2)	15.4(9.1)	5.4(2.6)	7.8(4.9)	4.9(2.8)	7.6(4.5)
	LAD-nhier	8.0(3.5)	33.3(8.7)	3.2(1.5)	4.1(2.4)	4.8(2.5)	29.2(7.8)
	LAD-group	12.7(6.2)	93.2(54.2)	4.6(2.4)	6.3(4.4)	8.1(3.9)	86.9(50.2)
+20%Cauchy(0,1)	LS-hier	9.2(6.9)	22.5(30.8)	3.8(2.9)	5.3(5.9)	5.3(4.3)	17.2(25.7)
	LS-nhier	9.4(6.3)	38.8(36.3)	2.2(2.1)	2.5(2.8)	7.1(4.8)	36.3(33.3)
	LS-group	11.2(7.1)	72.3(54.6)	3.8(2.6)	3.6(3.5)	7.4(4.6)	68.7(51.4)

Table 5

Identification results for SNP data with AR structure. In each cell, mean(sd) based on 100 replicates. TP/FP: total number of true/false positives. TP1/FP1: true/false positives for genetic main effects. TP2/FP2: true/false positives for interactions.

SNP(AR)		TP	FP	TP1	FP1	TP2	FP2
Error 1	LAD-hier	10.6(5.1)	14.9(9.4)	5.8(2.7)	7.8(5.1)	4.8(2.7)	7.1(4.5)
	LAD-nhier	9.5(3.0)	29.9(8.6)	4.6(1.7)	3.9(2.1)	4.8(2.1)	26.1(7.6)
N(0,1)	LAD-group	15.1(6.3)	169.8(91.4)	5.9(2.6)	12.6(7.8)	9.3(3.9)	157.2(83.8)
	LS-hier	12.2(5.8)	20.5(13.8)	5.5(2.4)	6.6(4.5)	6.7(3.8)	13.9(9.8)
	LS-nhier	13.4(5.5)	46.7(29.4)	3.7(2.1)	2.2(2.5)	9.7(4.2)	44.5(27.2)
	LS-group	14.3(6.8)	84.5(55.7)	4.9(2.5)	3.9(3.6)	9.4(4.5)	80.5(52.3)
Error 2	LAD-hier	12.5(4.3)	19.2(12.4)	6.3(1.9)	9.5(6.1)	6.1(2.6)	9.7(6.6)
	LAD-nhier	9.6(2.8)	34.6(10.4)	4.1(1.5)	4.6(2.4)	5.5(2.2)	30.0(9.3)
	LAD-group	14.8(6.1)	110.9(57.7)	5.7(2.5)	7.5(5.4)	9.1(3.9)	103.4(52.9)
+20%LogN(0,1)	LS-hier	12.9(6.6)	29.6(21.8)	5.3(2.6)	8.0(5.8)	7.6(4.2)	21.6(16.9)
	LS-nhier	12.3(6.1)	48.4(34.4)	3.1(2.1)	2.2(2.3)	9.3(4.7)	46.2(32.6)
	LS-group	14.5(7.1)	90.2(54.6)	4.9(2.5)	4.6(3.7)	9.6(4.8)	85.2(51.2)
Error 3 $\alpha(2)$	LAD-hier	9.5(5.4)	14.1(11.9)	4.8(2.8)	6.9(5.9)	4.7(2.8)	7.2(6.0)
	LAD-nhier	8.5(3.1)	32.2(8.3)	3.5(1.5)	4.3(2.9)	4.9(2.1)	27.8(6.9)
	LAD-group	15.1(5.1)	151.3(78.9)	5.5(2.2)	11.6(6.9)	9.6(3.4)	139.8(72.2)

SNP(AR)		TP	FP	TP1	FP1	TP2	FP2
	LS-hier	10.2(6.2)	22.6(19.8)	4.2(2.4)	6.4(5.0)	6.0(4.2)	16.3(1.7)
	LS-nhier	10.2(6.1)	35.5(31.4)	2.3(1.8)	2.4(3.2)	7.9(4.9)	33.0(28.7)
	LS-group	10.9(7.5)	63.9(4.1)	3.5(2.5)	3.2(3.7)	7.5(5.1)	60.7(50.7)
Error 4	LAD-hier	8.9(5.4)	14.9(12.8)	4.7(2.8)	7.4(6.5)	4.3(2.9)	7.5(6.5)
	LAD-nhier	6.5(3.4)	39.5(16.4)	2.9(1.6)	5.8(9.1)	3.6(3.5)	33.7(12.8)
80%N(0,1)	LAD-group	12.3(7.7)	138.7(97.4)	4.5(2.9)	10.4(7.9)	7.8(4.9)	128.3(89.6)
+20%Cauchy(0,1)	LS-hier	7.7(6.2)	22.1(22.3)	3.3(2.6)	5.6(5.8)	4.4(3.6)	16.5(17.1)
	LS-nhier	7.7(6.5)	34.6(34.7)	2.0(1.8)	2.1(3.1)	5.7(5.1)	32.5(32.2)
	LS-group	8.4(7.0)	63.1(52.7)	2.9(2.5)	3.4(3.3)	5.4(4.6)	59.7(49.6)

LAD-hier ■ LAD-group ■ LS-nhier ■
 LAD-nhier ■ LS-hier ■ LS-group ■

Figure 2.
Ratio plot of true positives over false positives

Table 6

Identification results for Gene expression data with AR structure and Error 3: $\pi(2)$. In each cell, mean(sd) based on 100 replicates. TP/FP: total number of true/false positives. TP1/FP1: true/false positives for genetic main effects. TP2/FP2: true/false positives for interactions.

GE(AR)		TP	FP	TP1	FP1	TP2	FP2
Validation	LAD-hier	10.9(4.4)	14.7(10.0)	5.6(2.0)	6.6(5.0)	5.3(2.8)	8.0(5.4)
CV(5-fold)	LAD-hier	11.8(4.2)	15.4(11.6)	6.1(2.1)	7.2(5.7)	5.8(2.7)	9.2(6.3)

Table 7

Identification and prediction results for gene expression data with banded structure. In each cell, mean(sd) based on 100 replicates. TP/FP: total number of true/false positives. PE: prediction error for robust (in L1 loss) and non-robust methods (in squared error loss).

GE(banded)		TP	FP	PE
Error 1	LAD-hier	12.5(4.6)	18.9(10.8)	1.6470(0.0958)
	LAD-nhier	11.6(3.0)	26.1(10.1)	1.6892(0.0739)
N(0,1)	LAD-group	15.3(5.2)	97.2(48.3)	1.8708(0.1283)
	LAD-median	-	-	3.7413(0.1137)
	LS-hier	14.0(5.2)	23.3 (13.0)	0.0092(0.0024)
	LS-nhier	13.7(4.7)	36.0(20.4)	0.0095(0.0024)
	LS-group	16.7(5.6)	85.3(32.9)	0.0081(0.0073)
	LS-mean	-	-	0.0158(0.0012)
Error 2	LAD-hier	12.3(5.2)	18.7(10.9)	1.6013(0.1047)
	LAD-nhier	12.1(3.7)	27.6(7.9)	1.6426(0.0817)
80%N(0,1)	LAD-group	14.7(5.7)	98.0(51.6)	1.8080(0.1125)
	LAD-median	-	-	3.6857(0.1218)
+20%LogN(0,1)	LS-hier	13.0(5.5)	34.4(25.9)	0.0089(0.0028)
	LS-nhier	13.7(4.7)	54.3(33.2)	0.0089(0.0023)
	LS-group	15.3(5.8)	97.9(45.9)	0.0078(0.0025)
	LS-mean	-	-	0.0160(0.0011)
Error 3	LAD-hier	10.3(4.6)	14.3(9.4)	1.7885(0.0961)
	LAD-nhier	9.6(3.3)	27.2(10.6)	1.8344(0.0879)
$\pi(2)$	LAD-group	12.2(5.3)	91.8(58.6)	1.9244(0.1039)
	LAD-median	-	-	3.8979(0.1266)
	LS-hier	10.8(5.7)	20.98(19.1)	0.0135(0.0042)
	LS-nhier	11.1(5.6)	34.7(29.2)	0.0134(0.0043)
	LS-group	13.0(6.9)	81.5(52.8)	0.0123(0.0045)
	LS-mean	-	-	0.0183(0.0030)

GE(banded)		TP	FP	PE
Error 4	LAD-hier	10.0(4.9)	15.3(12.3)	1.9371(0.2916)
	LAD-nhier	9.6(4.3)	26.4(10.3)	2.0287(0.3633)
80%N(0,1)	LAD-group	12.3(6.6)	88.2(63.9)	2.0824(0.2894)
	LAD-median			4.2027(0.5214)
+20%Cauchy(0,1)	LS-hier	8.7(7.3)	21.3(23.5)	1.1149(5.5288)
	LS-nhier	8.6(7.3)	33.6(39.2)	1.1097(5.5276)
	LS-group	11.6(7.1)	68.5(48.9)	1.1000(5.5232)
	LS-mean	-	-	1.1305(5.5664)

Table 8

Identification and prediction results for gene expression data with AR structure. In each cell, mean(sd) based on 100 replicates. TP/FP: total number of true/false positives. PE: prediction error for robust (in L1 loss) and non-robust methods (squared error loss).

GE(AR)		TP	FP	PE
Error 1	LAD-hier	12.9(3.8)	18.6(10.9)	1.6747(0.0929)
	LAD-nhier	12.1(3.5)	25.0(9.96)	1.7154(0.0772)
N(0,1)	LAD-group	15.2(5.3)	95.7(43.6)	1.9207(0.0932)
	LAD-median	-	-	3.7849(0.0976)
	LS-hier	14.4(4.9)	24.9(17.6)	0.0094(0.0024)
	LS-nhier	13.4(5.5)	35.0(27.5)	0.0100(0.0028)
	LS-group	15.3(6.0)	84.9(49.5)	0.0091(0.0027)
	LS-mean	-	-	0.0162(0.0011)
Error 2	LAD-hier	13.0(3.7)	17.8(8.4)	1.6116(0.0879)
	LAD-nhier	12.3(2.8)	24.8(7.5)	1.6514(0.0745)
80%N(0,1)	LAD-group	15.8(4.9)	114.4(63.3)	1.7965(0.1066)
	LAD-median	-	-	3.7069(0.0092)
+20%LogN(0,1)	LS-hier	13.1(4.9)	24.0(14.5)	0.0089(0.0025)
	LS-nhier	13.8(3.9)	44.0(24.0)	0.0085(0.0018)
	LS-group	16.0(5.9)	88.1(40.7)	0.0081(0.0027)
	LS-mean	-	-	0.0162(0.0016)
Error 3	LAD-hier	11.7(4.3)	16.2(11.4)	1.7923(0.1065)
	LAD-nhier	10.1(2.7)	31.6(9.3)	1.8525(0.0809)
t(2)	LAD-group	14.4(5.2)	91.6(52.7)	1.9994(0.1390)
	LAD-median	-	-	3.9087(0.1040)
	LS-hier	12.9(5.0)	26.0(25.4)	0.0150(0.0106)
	LS-nhier	13.2(3.9)	46.7(37.6)	0.0146(0.0110)
	LS-group	14.3(5.5)	82.1(50.8)	0.0139(0.0105)
	LS-mean	-	-	0.0210 (0.0098)

GE(AR)		TP	FP	PE
Error 4	LAD-hier	9.6(5.7)	13.2(10.2)	1.9323(0.2065)
	LAD-nhier	9.2(3.9)	31.5(8.6)	2.1085(0.5985)
80%N(0,1)	LAD-group	12.1(6.4)	70.9(48.85)	2.1016(0.2161)
	LAD-median	-	-	4.1599(0.3834)
+20%Cauchy(0,1)	LS-hier	8.6(7.7)	18.7(22.7)	1.3034(5.4621)
	LS-nhier	8.7(7.1)	31.4(35.3)	1.3078(5.4746)
	LS-group	9.9(8.3)	60.2(58.9)	1.2960(5.4565)
	LS-mean			1.3271(5.5245)

Table 9

Identification and prediction results for SNP data with banded structure. In each cell, mean(sd) based on 100 replicates. TP/FP: total number of true/false positives. PE: prediction error for robust (in L1 loss) and non-robust methods (squared error loss).

SNP(banded)		TP	FP	PE
Error 1	LAD-hier	11.5(4.4)	21.4(12.8)	1.6780(0.0813)
	LAD-nhier	10.2(3.4)	32.1(9.9)	1.7185(0.0679)
N(0,1)	LAD-group	14.4(4.6)	104.7(55.9)	1.9075(0.1272)
	LAD-median	-	-	3.6805(0.0884)
	LS-hier	12.1(5.4)	21.6(16.2)	0.0099(0.0027)
	LS-nhier	11.8(5.5)	36.2(29.1)	0.0101(0.0027)
	LS-group	14.0(5.87)	76.8(42.2)	0.0089(0.00267)
	LS-mean	-	-	0.0144(0.001)
Error 2	LAD-hier	12.0(4.2)	23.5(12.9)	1.6198(0.0898)
	LAD-nhier	10.9(3.6)	33.3(10.4)	1.6564(0.0764)
80%N(0,1)	LAD-group	15.2(4.7)	118.7(55.9)	1.8489(0.1264)
	LAD-median	-	-	3.6026(0.0897)
+20%LogN(0,1)	LS-hier	12.7(5.4)	27.3(25.2)	0.0088(0.0023)
	LS-nhier	12.9(5.0)	46.26(36.0)	0.0087(0.0021)
	LS-group	14.9(5.8)	93.9(52.1)	0.0079(0.0025)
	LS-mean	-	-	0.0141(0.0011)
Error 3	LAD-hier	10.2(5.2)	16.1(11.7)	1.8088(0.1154)
	LAD-nhier	9.4(2.8)	33.2(8.7)	1.8291(0.1132)
$\alpha(2)$	LAD-group	9.8(6.5)	69.9(59.8)	1.9524(0.1412)
	LAD-median	-	-	3.8086(0.1791)
	LS-hier	9.4(5.4)	23.7(24.4)	0.0130(0.0030)
	LS-nhier	9.3(5.9)	37.0(34.4)	0.0130(0.0030)
	LS-group	10.0(6.5)	62.1(52.6)	0.0126(0.0036)

SNP(banded)		TP	FP	PE
Error 4	LS-mean	-	-	0.0160(0.0011)
	LAD-hier	9.8(5.9)	16.5(11.6)	2.0664(0.7856)
	LAD-nhier	8.8(5.5)	32.7(28.0)	3.9013(0.9206)
	LAD-group	9.8(6.7)	66.3(53.9)	2.1978(0.8188)
80%N(0,1)	LAD-median	-	-	4.3523(1.5847)
	LS-hier	9.0(7.0)	22.7(22.2)	0.0634(0.1127)
	LS-nhier	8.9(5.6)	37.9(32.1)	0.0627(0.1108)
	LS-group	10.0(7.2)	71.1(62.1)	0.0614(0.1097)
+20%Cauchy(0,1)	LS-mean	-	-	0.0663(0.1131)

Table 10

Identification and prediction results for SNP data with AR structure. In each cell, mean(sd) based on 100 replicates. TP/FP: total number of true/false positives. PE: prediction error for robust (in L1 loss) and non-robust methods (squared error loss).

SNP(AR)		TP	FP	PE
Error 1	LAD-hier	11.9(4.8)	20.5(12.3)	1.6704(0.0873)
	LAD-nhier	10.6(3.2)	30.6(9.2)	1.7119(0.0701)
N(0,1)	LAD-group	14.9(5.3)	104.2(53.6)	1.8828(0.1107)
	LAD-median	-	-	3.6869(0.0958)
	LS-hier	12.9(5.5)	24.8(16.0)	0.0097(0.0025)
	LS-nhier	12.9(4.5)	43.9(26.3)	0.0098(0.0020)
	LS-group	14.3(6.3)	87.5(48.4)	0.0092(0.0030)
	LS-mean	-	-	0.0143(0.0010)
Error 2	LAD-hier	13.3(3.4)	23.7(12.9)	1.6119(0.0892)
	LAD-nhier	11.8(2.7)	30.9(9.3)	1.6773(0.0765)
80%N(0,1)	LAD-group	14.5(5.9)	109.2(60.3)	1.8404(0.1152)
	LAD-median	-	-	3.6466(0.1236)
+20%LogN(0,1)	LS-hier	12.8(5.3)	28.6(20.8)	0.0087(0.0024)
	LS-nhier	12.7(5.7)	48.3(38.4)	0.0094(0.0027)
	LS-group	15.1(6.8)	88.1(54.2)	0.0083(0.0031)
	LS-mean	-	-	0.0145(0.0009)
Error 3	LAD-hier	9.7(5.2)	16.6(11.7)	1.8040(1.106)
	LAD-nhier	8.9(3.3)	34.4(8.8)	1.8409(0.0831)
$\kappa(2)$	LAD-group	11.1(7.4)	81.1(61.7)	1.9894(0.1075)
	LAD-median	-	-	3.8398(0.0987)
	LS-hier	9.8(6.2)	21.5(16.9)	0.0146(0.0066)
	LS-nhier	10.7(5.2)	40.3(26.2)	0.0140(0.0064)

SNP(AR)		TP	FP	PE
	LS-group	11.8(6.5)	78.4(52.7)	0.0135(0.0066)
	LS-mean	-	-	0.0176(0.0055)
Error 4	LAD-hier	9.1(4.8)	16.7(12.8)	1.9630(0.2251)
	LAD-nhier	8.4(4.1)	37.7(14.7)	2.1778(0.6067)
80%N(0,1)	LAD-group	11.0(7.2)	89.8(67.9)	2.0996(0.2430)
	LAD-median	-	-	4.1039(0.4442)
+20%Cauchy(0,1)	LS-hier	10.5(6.8)	26.2(27.6)	1.4838(8.8279)
	LS-nhier	9.4(6.5)	40.6(40.7)	1.4767(8.7760)
	LS-group	11.6(7.4)	75.2(58.2)	1.4492(8.6095)
	LS-mean	-	-	1.5181(9.0188)

Table 11

Analysis of Lung Cancer data(LAD-nhier): identified main and interaction effects

Gene	Main Effects	Interactions				
		Age	Gender	Smoke	Chemo	Stage
SPINK1		0.011				0.055
TFF1	-0.104					
HLA-DQA1	0.010					
FGB		0.223				
AKR1C2		0.054				
WIF1		-0.103				-0.028
CPS1	-0.343					
PLA2G4A			0.046			0.007
PSPH		-0.182				-0.082
ADH1B						-0.038
IGHD						0.054
IFI27						-0.093
APOD						0.093
AHSG		-0.431				
ALDH3A1	-0.082					
SPRR1B						-0.067
APOB			-0.099			
HBG2		0.099				
HLA-DQB2		-0.047				
HMGA2						-0.126
CADM1				0.054		
PON3		-0.103				
PBK	-0.224					
CCL18	-0.063					

Gene	Main Effects	Interactions				
		Age	Gender	Smoke	Chemo	Stage
TFP1		0.083				
CHL1				0.230		
AQP1			-0.080			

Table 12

Analysis of Lung Cancer data(LS-hier): identified main and interaction effects

Gene	Main Effects	Interactions				
		Age	Gender	Smoke	Chemo	Stage
AKR1C2	0.001	0.001				
CPS1	-0.006					0.001
FGA	0.001	0.005	-0.001			
PLA2G4A	0.002		0.005			-0.001
PSPH	-0.004	-0.004				0.001
IGHD	0.002	-0.001	0.003			0.002
IFI27	-0.011	0.011	-0.015	0.001	0.001	0.001
FMO5	0.001					0.004
UBD	-0.001					
SERPINA3	0.006	0.008	-0.001			-0.001
PBK	-0.005					

Table 13

Analysis of Lung Cancer data(LS-nhier): identified main and interaction effects

Gene	Main Effects	Interactions				
		Age	Gender	Smoke	Chemo	Stage
CPS1	-0.006					0.001
FGA		0.002				
PLA2G4A	-0.001		0.005			-0.001
PSPH	-0.004	-0.005				0.001
IFI27	0.001	0.001	-0.016	0.001	0.001	0.001
SERPINA3	-0.001	0.005	-0.001			
PBK	-0.006					0.001
MAGEA2	-0.001	-0.001	0.0123		-0.001	-0.001

Table 1

Identification results for gene expression (GE) data **with banded structure**. In each cell, mean(sd) based on 100 replicates. TP/FP: total number of true/false positives. TP1/FP1 : true/false positives for genetic main effects. TP2/FP2: true/false positives for interaction effects.

GE(Banded)		TP	FP	TP1	FP1	TP2	FP2
Error 1	LAD-hier	13.1(3.9)	13.0(9.0)	6.6(2.0)	6.7(4.3)	6.5(2.3)	6.2(3.8)
	LAD-nhier	10.3(2.9)	23.5(7.2)	4.6(1.5)	2.9(1.9)	5.7(1.8)	20.6(6.7)
	LAD-group	16.0(4.7)	150.5(61.0)	6.1(2.2)	11.0(5.7)	9.9(2.9)	139.4(55.8)
	LS-hier	12.2(5.0)	17.0(10.5)	5.5(2.3)	5.7(3.8)	6.7(3.1)	11.2(7.1)
	LS-nhier	11.3(4.9)	20.7(14.9)	3.3(2.0)	0.9(1.3)	8.0(3.4)	19.8(13.9)
	LS-group	14.7(5.4)	76.1(41.0)	5.0(2.1)	3.2(2.5)	9.7(3.6)	72.9(38.6)
Error 2	LAD-hier	13.4(3.9)	15.8(9.9)	6.6(1.8)	7.9(5.8)	6.8(2.4)	7.9(5.4)
	LAD-nhier	11.0(2.7)	27.3(6.9)	4.7(1.2)	3.6(1.8)	6.3(2.2)	23.6(6.2)
	LAD-group	17.0(3.4)	154.7(63.2)	6.6(1.6)	11.5(5.6)	10.4(2.1)	143.2(57.9)
80%N(0,1)	LS-hier	13.6(5.3)	22.6(16.0)	5.8(2.2)	6.8(4.2)	7.8(3.4)	15.8(12.7)
	LS-nhier	13.5(5.1)	37.4(24.1)	3.8(2.2)	1.8(1.8)	9.7(3.5)	35.6(22.9)
	LS-group	15.5(5.5)	81.4(36.1)	5.2(1.9)	3.8(2.7)	10.2(3.8)	77.6(33.7)
Error 3	LAD-hier	11.0(4.3)	16.3(11.8)	5.7(2.3)	8.5(6.4)	5.3(2.3)	7.8(5.7)
	LAD-nhier	8.6(2.7)	34.2(8.5)	3.9(1.5)	4.8(2.0)	4.7(1.8)	29.4(8.0)
	LAD-group	14.9(5.4)	138.2(79.7)	5.4(2.1)	10.1(6.9)	9.5(3.7)	128.1(73.02)
t(2)	LS-hier	11.9(5.7)	22.6(18.1)	5.0(2.3)	7.1(5.3)	6.8(3.6)	15.5(13.5)
	LS-nhier	11.9(5.4)	37.9(28.3)	3.4(2.2)	2.1(2.4)	8.4(3.8)	35.8(26.3)
	LS-group	13.6(6.9)	83.3(50.7)	4.5(2.3)	4.2(3.4)	9.1(4.8)	79.1(47.5)
Error 4	LAD-hier	9.6(5.5)	13.4(11.7)	5.0(2.8)	6.6(5.6)	4.5(3.2)	6.8(6.2)
	LAD-nhier	8.0(3.6)	31.0(9.5)	3.5(1.8)	3.9(2.6)	4.5(2.3)	27.1(8.5)
	LAD-group	12.9(7.4)	112.0(82.9)	4.8(3.0)	7.8(6.5)	8.0(4.5)	104.2(76.7)
80%N(0,1)	LS-hier	8.5(6.0)	19.9(21.4)	3.7(2.5)	5.2(4.8)	4.8(3.6)	14.6(17.0)
	LS-nhier	8.8(6.1)	34.9(37.2)	2.2(1.8)	1.9(2.8)	6.5(4.6)	32.9(34.7)
	LS-group	10.6(7.0)	74.6(60.7)	3.6(2.5)	4.1(3.9)	7.0(4.6)	70.5(56.8)
+20%Cauchy(0,1)	LS-hier	8.5(6.0)	19.9(21.4)	3.7(2.5)	5.2(4.8)	4.8(3.6)	14.6(17.0)
	LS-nhier	8.8(6.1)	34.9(37.2)	2.2(1.8)	1.9(2.8)	6.5(4.6)	32.9(34.7)
	LS-group	10.6(7.0)	74.6(60.7)	3.6(2.5)	4.1(3.9)	7.0(4.6)	70.5(56.8)

Table 2
Analysis of Lung Cancer data(LAD-HIER): identified main and interaction effects

Gene	Main Effects	Interactions			
		Age	Gender	Smoke	Chemo
SPINK1	0.049				
TFF1	-0.104				
FGB	0.008	0.218			
AKR1C2	0.046				
WIF1	-0.005	-0.111			
CPS1	-0.343				
PLA2G4A	0.071				
PSPH	0.081	-0.171			-0.223
IGHD	0.049				
IFI27	-0.047				
APOD	0.085				
AHSG	-0.124	-0.341			
ALDH3A1	-0.081				
HBG2	-0.009	0.171			
HMG2	-0.078				-0.074
PON3	0.097	-0.167			
PBK	-0.234				
CCL18	-0.056				
TFP1	0.028	0.106			
CHL1	0.001			0.229	
AQP1	-0.109				