Dandan Jiang and Jianguo Sun*

# Group Tests for High-dimensional Failure Time Data with the Additive Hazards Models

DOI 10.1515/ijb-2016-0085

**Abstract:** Statistical analysis of high-dimensional data has been attracting more and more attention due to the abundance of such data in various fields such as genetic studies or genomics and the existence of many interesting topics. Among them, one is the identification of a gene or genes that have significant effects on the occurrence of or are significantly related to a certain disease. In this paper, we will discuss such a problem that can be formulated as a group test or testing a group of variables or coefficients when one faces right-censored failure time response variable. For the problem, we develop a corrected variance reduced partial profiling (CVRPP) linear regression model and a likelihood ratio test procedure when the failure time of interest follows the additive hazards model. The numerical study suggests that the proposed method works well in practical situations and gives better performance than the existing one. An illustrative example is provided.

**Keywords:** additive hazards model, group tests, high-dimensional data

## 1 Introduction

Statistical analysis of high-dimensional data has been attracting more and more attention due to the abundance of such data in various fields such as genetic studies or genomics and the existence of many interesting topics. Among them, one is the identification of a gene or genes that have significant effects on the occurrence of or are significantly related to a certain disease for the purpose of predicting survival rates among others [1–4]. In this paper, we will discuss such a problem that can be formulated as a group test or testing a group of predictor variables or coefficients such as genes or genomic factors when one faces right-censored failure time response variable. By high-dimension, we usually mean that the number of predictor variables denoted by $p$ is much larger than the sample size $n$, and it is well-known that for these situations, traditional statistical methods cannot be applied. In other words, some new procedures that allow $p \geq n$ are required.

Let $T$ denote a failure time variable of interest and $Z$ a $p$-dimensional vector of covariates or predictor variables. For regression analysis of failure time data, one commonly used model is the additive hazards model (AHM) given in the form of the hazard function of $T$ as

$$\lambda(t|Z) = \lambda_0(t) + \sum_{j=1}^{p} \beta_{0j} Z_j(t) \tag{1}$$

given $Z$ [5, 6]. In the above, $\lambda_0(t)$ is an unknown baseline hazard function, $\beta_0 = (\beta_{01}, \cdots, \beta_{0p})^\top$ a $p$-dimensional vector of unknown regression coefficients or parameters, and $Z = (Z_1, \cdots, Z_p)^\top$. Note that in contrast to other models, model (1) describes the additive covariate effects, the type of effects that are often of more interest in many areas such as social sciences [6]. Of course, one may want to consider other models such as linear transformation models if other types of effects are of interest and more comments on this can be found below. Also many authors have discussed the model above for both the traditional situation with a fixed $p$ and the

***Corresponding author: Jianguo Sun,** Center for Applied Statistical Research, School of Mathematics, Jilin University, Changchun, China; Department of Statistics, University of Missouri, Missouri, USA, E-mail: sunj@missouri.edu
**Dandan Jiang,** Center for Applied Statistical Research, School of Mathematics, Jilin University, Changchun, China

high-dimensional situation concerning various topics of interest. However, it seems that there exists little literature on testing the hypothesis

$$H_0 : \beta_{0,d} = 0 \quad \text{v.s.} \quad H_1 : \beta_{0,d} \neq 0\,, \tag{2}$$

or the effects of a set of the predictor variables for the high-dimensional situation. In the above, $d = \{d_1, \cdots, d_q\}$ is a subset of $\{1, \cdots, p\}$ and $\beta_{0,d} = (\beta_{0d_1}, \cdots, \beta_{0d_q})^\top$ denotes a $q$-dimensional sub-vector of $\beta_0$. In genomics among others, it is often the case that a group of genes rather than a single gene may be significantly related to or responsible for a certain disease and thus it is of interest to perform a group test like the hypothesis $H_0$ above.

As mentioned above, many authors have discussed the analysis of high-dimensional data or more specifically the analysis of high-dimensional data under the failure time context or concerning the test of hypotheses similar to $H_0$. For example, [7–10] investigated the parameter estimation problem related to model (1), and [11] and [12] considered a hypothesis test problem similar to that above but with $q = p$ under the context of linear regression models. In addition, [13, 14], [15] and [16] studied similar testing problems under the context of failure time analysis and generalized linear models. Note that all methods above concern the test on the whole set of the predictor variables or coefficients simultaneously and cannot apply to the test of $H_0$ or on a subset of the coefficients as they cannot provide a valid $p$-value. More recently Zhong et al. [4] discussed the test problem regarding a single coefficient ($q = 1$) and developed a variance reduced partial profiling (VRPP) linear regression model for the derivation of their test statistic. Furthermore they briefly considered the testing of the hypothesis $H_0$ and generalized the proposed test statistic for the $q > 1$ case. However, as discussed below, the generalized test procedure may not work properly even for $q = 2$ when the predictor variables or covariates within the set $d$ are highly correlated.

In the following, to test the hypothesis $H_0$, we will develop a corrected variance reduced partial profiling (CVRPP) linear regression model and present a new test statistic. Of course, the proposed test procedure applies to the $q = 1$ situation too. To present the proposed test statistic, we will first define some notation and review the VRPP linear regression model and the test statistic given in Zhong et al. [4] in Section 2. Section 3 discusses the proposed CVRPP model and the resulting likelihood ratio test along with its implementation. In Section 4, we will present some results obtained from a simulation study conducted to evaluate the performance of the test procedure and they suggest that it works well in practical situations. An illustration is given in Section 5 and Section 6 provides some discussion and concluding remarks.

## 2 Notation and review

Consider a failure time study and let $T$ and $Z$ be defined as above. Suppose that $T$ follows the model (1) and the main objective is to test the hypothesis $H_0$ defined in eq. (2) for given $d$. Also suppose that there exists a right censoring time denoted by $C$ and the observed data have the form $\{X_i, \Delta_i, \{Z_{ij}(t)\}_{j=1}^p; i = 1, \cdots, n\}$ given by $n$ independent subjects, where $X_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$ with $T_i$, $C_i$ and $Z_i$ defined as $T$, $C$ and $Z$ but with respect to subject $i$. As mentioned above, the focus will be on the situation where $p$ is greater than $n$ and also we will assume that the censoring time $C$ is independent of $T$ given the covariates $Z(t) = \left(Z_{ij}(t)\right)_{n\times p}$, which will be assumed to be centralized such that $E(Z_{ij}(t)) = 0$.

Furthermore let $\Sigma(t) = \left(\sigma_{jk}\right)_{j,k=1}^p$ denote the $p \times p$ covariance matrix of $Z$ and define the counting process $N_i(t) = I(X_i \leq t, \Delta_i = 1)$, the risk indicator $Y_i(t) = I(X_i \geq t)$, and the martingale

$$M_i(t) = N_i(t) - \int_0^t Y_i(u)\{\lambda_0(u) + \sum_{j=1}^p \beta_{0j} Z_{ij}(t)\}\mathrm{d}u\,, \tag{3}$$

$i = 1, ..., n$. Also define $Y(t) = (Y_1(t), \cdots, Y_n(t))^\top$, $\mathcal{Y} = Y(t)(\mathbf{1}^\top Y(t))^{-1}$, which is a $n \times 1$ vector, and $\mathrm{d}A = (\mathrm{d}A_1(t), \cdots, \mathrm{d}A_n(t))^\top$ for a vector of $A = (A_1(t), \cdots, A_n(t))^\top$. Then several authors have shown that we have the equation

$$(\mathbf{I}_n - \mathcal{Y}\mathbf{1}^\top)\mathrm{d}N = \tilde{Z}(t)\beta_0\mathrm{d}t + (\mathbf{I}_n - \mathcal{Y}\mathbf{1}^\top)\mathrm{d}M\,, \tag{4}$$

(see Martinussen and Scheike [9, 10]), where $\tilde{Z}(t) = (\tilde{Z}_1(t), \cdots, \tilde{Z}_p(t)) = (\mathbf{I}_n - \mathcal{Y}\mathbf{1}^\top)\,\mathrm{diag}\,\{Y(t)\}Z(t)$, a $n \times p$ matrix. Note that the equation above has two important features. One is that the first term on the right hand side is linear in regression parameter $\beta_0$ and the other is that the second term on the right hand side involves the martingale difference and thus can be treated as a random error. In other words, one can treat eq. (4) as a linear regression model when making inference.

For a subset $A \subset \{1, \cdots, p\}$, let $\tilde{Z}_A(t)$ denote the $n \times |A|$ submatrix of $\tilde{Z}(t)$ including the columns corresponding to $A$ with $|A|$ denoting the dimension of $A$, and for the time being, we assume that $d = \{1\}$. That is, we are only interested in a single regression parameter $\beta_{0,1}$ for notation convenience. To test $H_0$ in this case, Zhong et al. [4] suggested to estimate $\beta_{0,1}$ by using the eq. (4) first and then to apply the resulting Wald test statistic. For this, note that if $\tilde{Z}_1(t)$ is perpendicular to all other $\tilde{Z}_j(t)$ for $j > 1$, then one can easily estimate $\beta_{0,1}$ by multiplying $\tilde{Z}_1^\top(t)$ on both sides of eq. (4). On the other hand, it is apparent that this is not likely true in practice. To address this, Zhong et al. [4] proposed to obtain a relaxed orthogonalization by dividing all other $\tilde{Z}_j(t)$ or $\{2, ..., p\}$ into two parts $S_1$ and $S_1^c$ based on the measure $Q_{kj} = \mathrm{E}\{\int_0^\tau \tilde{Z}_{ik}(t)\tilde{Z}_{ij}(t)\mathrm{d}t\}$ for any $k, j \in \{1, ..., p\}$, where $S_1 = \{l \in \{2, \cdots, p\} : |Q_{1l}| > \eta^*\}$ for a pre-specified constant $\eta^*$ and $S_1^c$ denotes the complementary set of $\{1, S_1\}$.

Given $S_1$, for estimation of $\beta_{0,1}$, Zhong et al. [4] suggested to partially profile out the effect of $\tilde{Z}_{S_1}$ by multiplying $\mathbf{I}_n - P_{\tilde{Z}_{S_1}}$ on both sides of eq. (4), which gives

$$(\mathbf{I}_n - P_{\tilde{Z}_{S_1}})\,(\mathbf{I}_n - \mathcal{Y}\mathbf{1}^\top)\mathrm{d}N = (\mathbf{I}_n - P_{\tilde{Z}_{S_1}})\tilde{Z}_{\{1\}}(t)\beta_{0,1}\mathrm{d}t$$

$$+\,(\mathbf{I}_n - P_{\tilde{Z}_{S_1}})\,\tilde{Z}_{S_1^c}(t)\,\beta_{0,S_1^c}\,\mathrm{d}t\,+\,(\mathbf{I}_n - P_{\tilde{Z}_{S_1}})\,(\mathbf{I}_n - \mathcal{Y}\mathbf{1}^\top)\,\mathrm{d}M\,, \tag{5}$$

where $P_{\tilde{Z}_{S_1}} = \tilde{Z}_{S_1}(\tilde{Z}_{S_1}^\top\tilde{Z}_{S_1})^{-1}\tilde{Z}_{S_1}^\top$, the projection matrix corresponding to $\tilde{Z}_{S_1}$. Furthermore they proposed to replace $\beta_{0,S_1^c}$ by some reasonable initial estimator $\hat{\beta}_{0,S_1^c}$ in the second term on the right hand side of the equation above and to move the term to the left side, for which they call the resulting equation as the VRPP linear regression model. By applying the linear regression model idea described above to the VRPP model and after taking the integration, Zhong et al. [4] suggested the following variance reduced partial profiling estimator (VRPPE) of $\beta_{0,1}$

$$\hat{\beta}_{0,1} = \left(\int_0^\tau \tilde{Z}_1^\top(t)(\mathbf{I}_n - P_{\tilde{Z}_{S_1}})\tilde{Z}_1(t)\mathrm{d}t\right)^{-1}$$

$$\times\left(\int_0^t \tilde{Z}_1^\top(t)\,(\mathbf{I}_n - P_{\tilde{Z}_{S_1}})\mathrm{d}N - \int_0^\tau \tilde{Z}_1^\top(t)\,(\mathbf{I}_n - P_{\tilde{Z}_{S_1}})\tilde{Z}_{S_1^c}(t)\mathrm{d}t\hat{\beta}_{0,S_1^c}\right). \tag{6}$$

In addition, they suggested to estimate its asymptotic variance by $\hat{\sigma}^2_{\beta_1}/n$ with

$$\hat{\sigma}^2_{\beta_1} = \left(n^{-1}\int_0^\tau \tilde{Z}_1^\top(t)(\mathbf{I}_n - P_{\tilde{Z}_{S_1}})\tilde{Z}_1(t)\mathrm{d}t\right)^{-2} n^{-1}\int_0^t \tilde{Z}_1^\top(t)\,(\mathbf{I}_n - P_{\tilde{Z}_{S_1}})\tilde{Z}_1(t)\mathrm{d}N.$$

It follows that one can test $H_0$ using the Wald statistic based on $\hat{\beta}_{0,1}$, which was shown to follow asymptotically a normal distribution.

For general $d$ with $q << \min(p, n)$, by replacing $S_1$ with $S_d = \cup_{j=1}^q S_{d_j}$, Zhong et al. [4] suggested to construct a Wald test statistic similarly as above, which was shown to asymptotically follow the $\chi^2$ distribution, where $S_{d_j}$ is defined exactly as $S_1$ except for the $j$th component of the set $d$. However, a couple of serious issues can occur with this test procedure. One is that it is easy to see that in general, the sets $d$ and $S_d$ may have or share

some overlapping elements and it is apparent that the number of the overlapping elements will increase with the rising dimensionality $p$ or the highly correlated relationship. This would mean that the parameter estimation throw away a lot of relevant information and yield some serious errors. This can be even more serious if the $\tilde{Z}_j$'s corresponding to the set $d$ are highly correlated and it can be seen below that this can happen even with $q = 2$. Another issue is that the convergence to the $\chi^2$ distribution can be quite slow and thus cannot be relied on for practical situations. In the next section, we will address these issues and propose a new test procedure.

## 3 A likelihood ratio test procedure

Now we will present a new test procedure based on the likelihood ratio principle. For this, we will first present a different partition procedure for $\{d, S_d, S_d^c\}$ given a set $d$ and develop a CVRPP linear regression model. The proposed new test statistic for $H_0$ will then be derived.

To describe the new partition procedure, let $Q_R$ denote the correlation matrix given by the covariance matrix $Q = (Q_{kj})_{k,j=1}^p$. Then define the number matrix $Q_{ord}$ such that its $j$th column is the permutation of $(1, \cdots, p)$ determined by the correlations between $\tilde{Z}_j$ and all the $\tilde{Z}_m$'s, the $j$th column of $Q_R$, in the decreasing order. It is apparent that the first row of $Q_{ord}$ will be $(1, \cdots, p)$. Now for a given subset $d \subset \{1, \cdots, p\}$ and an appropriately chosen integer $l \in \{2, \cdots, p\}$ to be discussed below, define the new subset or partition $S_d^l$ to include all different numbers in the submatrix $Q_{ord}[2 : l, d]$ minus the elements in $d$. For the notation convenience, we will continue to use $S_d$ to denote $S_d^l$ and as before define $S_d^c$ as the complementary set of $\{d, S_d\}$.

For the development of the CVRPP linear regression model based on the new partition procedure, for each $i = 1, ..., n$ or at time $t = X_1, ..., X_n$, define

$$U_i = (\mathbf{I}_n - P_{\tilde{Z}_{S_d}(X_i)})\,(\mathbf{I}_n - \mathcal{Y}(X_i)\mathbf{1}^\top)\,\Delta N(X_i) - (\mathbf{I}_n - P_{\tilde{Z}_{S_d}(X_i)})\tilde{Z}_{S_d^c}(X_i)\hat{\beta}_{0,S_d^c}\Delta X_i\,,$$

$$V_i = (\mathbf{I}_n - P_{\tilde{Z}_{S_d}(X_i)})\,\tilde{Z}_d(X_i)\,\Delta X_i\,,$$

and

$$\epsilon_i = (\mathbf{I}_n - P_{\tilde{Z}_{S_d}(X_i)})\,(\mathbf{I}_n - \mathcal{Y}(X_i)\mathbf{1}^\top)\Delta M(X_i)\,.$$

Then as the VRPP linear regression model (2.10) in Zhong et al. [4], we have

$$U_i = V_i\,\beta_{0,d} + \epsilon_i\,, \quad \text{for } i = 1, \cdots, n\,. \tag{7}$$

Note that $\epsilon_i$ has mean zero and serves as a random error similar to that in a classical linear regression model. To further deduct the variance or simplify the model above, let $\hat{\beta}_{0,d}$ denote the estimator of $\beta_{0,d}$ as that given by eq. (6) with the new partition, and define $\hat{\epsilon}_i = U_i - V_i\,\hat{\beta}_{0,d}$. It is apparent that we can rewrite eq. (7) as

$$U_i = V_i\beta_{0,d} + \hat{\epsilon}_i + \epsilon_i^*\,, \quad \text{for } i = 1, \cdots, n, \tag{8}$$

where $\epsilon_i^* = \epsilon_i - \hat{\epsilon}_i$.

Note that one can naturally estimate the covariance matrix of $\epsilon_i^*$ by $\hat{\Sigma}_{\epsilon_i} = V_i\,\hat{\Sigma}_{\beta_{0,d}}\,V_i^\top$, where $\hat{\Sigma}_{\beta_{0,d}} = \mathbf{A}^{-1}\,\mathbf{B}\,\mathbf{A}^{-1}$ with $\mathbf{A} = \int_0^\tau \tilde{Z}_d^\top(t)(\mathbf{I} - P_{\tilde{Z}_{S_d}})\tilde{Z}_d(t)\mathrm{d}t$ and $\mathbf{B} = \int_0^\tau \tilde{Z}_d^\top(t)(\mathbf{I} - P_{\tilde{Z}_{S_d}})\tilde{Z}_d(t)\mathrm{d}N$. Define $D_i = \mathrm{diag}(\mathrm{diag}(\hat{\Sigma}_{\epsilon_i}))$, a diagonal matrix given by the variances of the components of $\epsilon_i^*$. By multiplyiing $D_i^{-\frac{1}{2}}$ to both sides of eq. (8), we obtain that

$$D_i^{-\frac{1}{2}}\,U_i = D_i^{-\frac{1}{2}}\,V_i\,\beta_{0,d} + D_i^{-\frac{1}{2}}\,\hat{\epsilon}_i + D_i^{-\frac{1}{2}}\,\epsilon_i^*\,. \tag{9}$$

Note that each component of $D_i^{-\frac{1}{2}}\epsilon_i^*$ has mean 0 and variance 1. This suggests that one can simplify the equation above by replacing $D_i^{-\frac{1}{2}}\epsilon_i^*$ with the random error $\tilde{\epsilon}_i$ satisfying $\mathrm{E}(\tilde{\epsilon}_i) = \mathbf{0}_n$ and $\mathrm{E}(\tilde{\epsilon}_i\tilde{\epsilon}_i^\top) = \mathbf{I}_n$. In other words, instead of model (9), we can consider the CVRPP linear regression model

$$\tilde{U}_i = D_i^{-\frac{1}{2}} V_i \beta_{0,d} + D_i^{-\frac{1}{2}} \hat{\epsilon}_i + \tilde{\epsilon}_i\,, \quad \text{for } i = 1, \cdots, n, \tag{10}$$

for testing the hypothesis $H_0$ in eq. (2), where $\tilde{U}_i = D_i^{-\frac{1}{2}} U_i$ and the $\tilde{\epsilon}_i$'s are generated from the multivariate standard normal distribution $N(\mathbf{0}_n, \mathbf{I}_n)$ for simplicity.

Under model (10) and for each $i$, a natural test statistic for $H_0$ is apparently given by the likelihood ratio statistic $\Lambda_n^{(i)} = |\Sigma_\Omega^{(i)}|/|\Sigma_\omega^{(i)}|$, where

$$\Sigma_\Omega^{(i)} = \frac{1}{n}(\tilde{U}_i - D_i^{-\frac{1}{2}}\hat{\epsilon}_i - D_i^{-\frac{1}{2}} V_i \hat{\beta}_{0,d}^{(i)})^\top (\tilde{U}_i - D_i^{-\frac{1}{2}}\hat{\epsilon}_i - D_i^{-\frac{1}{2}} V_i \hat{\beta}_{0,d}^{(i)})$$

and

$$\Sigma_\omega^{(i)} = \frac{1}{n}(\tilde{U}_i - D_i^{-\frac{1}{2}}\hat{\epsilon}_i)^\top (\tilde{U}_i - D_i^{-\frac{1}{2}}\hat{\epsilon}_i)\,,$$

the estimators of the covariance matrix of $\tilde{\epsilon}_i$ under the null and alternative hypotheses, respectively, and $\hat{\beta}_{0,d}^{(i)} = (V_i^\top D_i^{-1} V_i)^-(V_i^\top D_i^{-1/2}(\tilde{U}_i - D_i^{-\frac{1}{2}}\hat{\epsilon}_i))$, the estimator of $\beta_{0,d}$ based on the CVRPP model. The critical region based on $\Lambda_n^{(i)}$ has $\{\Lambda_n^{(i)} < \lambda_0^{(i)}\}$, where $\{\lambda_0^{(i)}\}$ is a suitably chosen number. In reality, for testing $H_0$, it would be natural to combine all $\Lambda_n^{(i)}$'s together for a couple of reasons. One is that it is not easy to find $\lambda_0^{(i)}$ and for it, of course, one could apply some resampling methods. However, this clearly would be time-consuming in computation. Another more important reason is that it is apparent that one has to combine all individual testing results together, which may not be straightforward or could be difficult. On the other hand, if we assume $X_1 = \min(X_1, \cdots, X_n)$ without loss of generality, one can find out through studying the $U_i$'s that one only needs to focus on the eq. (10) with $i = 1$ or the statistic $\Lambda_n^{(1)}$. Furthermore, it follows from Theorem 8.4.5 in Anderson [17] that one can test the hypothesis $H_0$ by using the likelihood-based statistic

$$T_F = \frac{(n-q)\,(1-\Lambda_n^{(1)})}{q\,\Lambda_n^{(1)}}\,, \tag{11}$$

whose distribution can be approximated by the $F$-distribution with the freedom degrees of $q$ and $n-q$. Consequently, the critical region $\{\Lambda_n^{(1)} < \lambda_0^{(1)}\}$ is equivalent to $\{T_F > F_\alpha(q, n-q)\}$, where $F_\alpha(q, n-q)$ is the $\alpha$-upper quantile of the $F$-distribution $F(q, n-q)$.

Note that to implement the likelihood ratio test procedure above, one needs to choose $l$ in the determination of $Q_{ord}$. For this, by following Zhong et al. [4], we suggest first to select a subset $L \subset \{2, \cdots, p\}$ and then to search the optimal $l$ over the range of $L$. More specifically, the subset $L$ should be chosen such that both $S_d$ and $S_d^c$ exist and the dimension of $S_d$ is less than 30% of the whole set. For given $L$, one can choose the smallest $l$ within $L$ such that $D_n(S_{d,l}) \le 5\sqrt{(\log p/n)}$, where

$$D_n(S_{d,l}) = \max_{j \in S_{d,l}^c} || \left( \int_0^\tau \tilde{Z}_d^\top(t)(\mathbf{I} - P_{\tilde{Z}_{S_{d,l}}})\tilde{Z}_d(t)\mathrm{d}t \right)^- \int_0^\tau \tilde{Z}_d^\top(t)(\mathbf{I} - P_{\tilde{Z}_{S_{d,l}}})\tilde{Z}_j(t)||.$$

If no such $l$ exists, one can use $l$ that minimizes $D_n(S_{d,l})$ within $L$.

Two other issues related to the implementation of the proposed test procedure is the determination of $\hat{\beta}_{0,S_1^c}$ in eq. (6) and the generation of the $\tilde{\epsilon}_i$'s in the CVRPP model (10). For the former, Zhong et al. [4] pointed out that any initial estimator satisfying $|\hat{\beta}_{0,S_1^c} - \beta_{0,S_1^c}| = o_p\{\sqrt{(\log p/n)}\}$ can be used and suggested to use a LASSO type estimator for its fast computation. For the latter, it is apparent that the test result may depend on the values of the generated $\tilde{\epsilon}_i$'s. To address this, for a given data set, we suggest to repeat the process many

times as discussed in the example below. For the simulation study in the next section, only one sample will be used.

## 4 A simulation study

In this section, we present some results obtained from a simulation study conducted to assess the performance of the likelihood ratio test procedure proposed in the previous sections. In the study, by following Zhong et al. [4], we generated the true failure times $T_i$'s based on model (1) with $\beta_0 = (2, 1, 0.5, 0, 0, 0, 0, 0, 0, 0, 1, 2, 0.8, 1.2, 1, \mathbf{0}_{p-15}^\top)^\top$ and assumed that the $Z_i$'s follow the multivariate normal distribution with mean 0 and covariance matrix $\Sigma = (\rho^{|i-j|})$ for $(1 \le i, j \le p)$ subject to the constraint $Z_i^\top \beta_0 > -1$, where $\mathbf{0}_{p-15}$ is a vector of $p-15$ zeros. Furthermore the censoring times $C_i$'s were generated from the uniform distribution $(0, c_0)$, where $c_0$ was chosen to give the required percentage of right-censored failure times. In addition to the proposed test statistic $T_F$, for comparison, we also considered the $\chi^2$ test statistic given in Zhong et al. [4], which will be denoted as $T_Z$ below, and obtained the corresponding test results for each situation considered. The results given below are based on 1,000 replications.

Tables 1 and 2 give the estimated size and empirical power of the test procedures based on $T_Z$ and $T_F$ for testing the hypothesis $H_0 : \beta_{0,d} = \mathbf{0}$ with the dimension of $d$ being $q = 1, 2, 5, 10$ or 20, $p = 100, 200$ or 600, and $n = 100$. Here we took $\rho = 0.6$ and $L = \{5, \cdots, 10\}$, and for the power estimation, we set $\beta_{0,d} = \mathbf{1}_q$, the $q$-dimensional vector with all components equal to 1. Table 1 corresponds to the situation with 25% of right-censored failure times ($c_0 = 2$), while Table 2 considered the situation with 35% of right-censored failure times ($c_0 = 1$). One can see from the tables that with $q = 1$, both test procedures seem to give the expected nominal size 5% but the proposed new likelihood ratio test procedure was clearly more powerful than that based on $T_Z$. For the cases with $q > 1$, it is apparent that the test procedure based on $T_Z$ cannot be applied, while the new procedure based on $T_F$ still gave the expected nominal size 5%. Furthermore the new procedure seems still to have good power but the power seems to depend on both $p$ and $n$ as expected. We also investigated other set-ups with different values of $\rho$, different set $L$ and other percentages of right censoring as well as different values of $p$ and $n$ and obtained similar results.

**Table 1:** Empirical sizes and powers of the test procedure given in Zhong et al. [4] and the proposed one with 25% right censoring percentage.

| | | Sizes | | Powers | |
|---|---|---|---|---|---|
| $d$ | $(p,n)$ | $T_Z$ | $T_F$ | $T_Z$ | $T_F$ |
| | (100, 100) | 0.051 | 0.040 | 0.176 | 1 |
| {5} | (200, 100) | 0.056 | 0.041 | 0.141 | 1 |
| | (600, 100) | 0.046 | 0.053 | 0.108 | 1 |
| | (100, 100) | 0.128 | 0.043 | 0.124 | 1 |
| {5, 6} | (200, 100) | 0.117 | 0.047 | 0.139 | 1 |
| | (600, 100) | 0.133 | 0.049 | 0.133 | 1 |
| | (100, 100) | 0.321 | 0.041 | 0.243 | 1 |
| {6 : 10} | (200, 100) | 0.309 | 0.053 | 0.207 | 1 |
| | (600, 100) | 0.261 | 0.055 | 0.180 | 1 |
| | (100, 100) | 0.403 | 0.047 | 0.342 | 1 |
| {16 : 25} | (200, 100) | 0.423 | 0.053 | 0.324 | 1 |
| | (600, 100) | 0.375 | 0.049 | 0.233 | 1 |
| | (100, 100) | 0.464 | 0.050 | 0.366 | 0.997 |
| {16 : 35} | (200, 100) | 0.458 | 0.059 | 0.329 | 0.974 |
| | (600, 100) | 0.360 | 0.048 | 0.202 | 0.868 |

**Table 2:** Empirical sizes and powers of the test procedure given in Zhong et al. [4] and the proposed one with 35% right censoring percentage.

| | | Sizes | | Powers | |
|---|---|---|---|---|---|
| $d$ | $(p, n)$ | $T_Z$ | $T_F$ | $T_Z$ | $T_F$ |
| | (100, 100) | 0.058 | 0.052 | 0.142 | 1 |
| {5} | (200, 100) | 0.051 | 0.056 | 0.116 | 1 |
| | (600, 100) | 0.044 | 0.056 | 0.081 | 1 |
| | (100, 100) | 0.155 | 0.051 | 0.126 | 1 |
| {5, 6} | (200, 100) | 0.119 | 0.055 | 0.111 | 1 |
| | (600, 100) | 0.116 | 0.057 | 0.131 | 1 |
| | (100, 100) | 0.303 | 0.062 | 0.293 | 1 |
| {6 : 10} | (200, 100) | 0.286 | 0.058 | 0.272 | 1 |
| | (600, 100) | 0.295 | 0.046 | 0.218 | 1 |
| | (100, 100) | 0.437 | 0.052 | 0.356 | 1 |
| {16 : 25} | (200, 100) | 0.439 | 0.039 | 0.371 | 1 |
| | (600, 100) | 0.423 | 0.050 | 0.289 | 1 |
| | (100, 100) | 0.514 | 0.049 | 0.443 | 0.998 |
| {16 : 35} | (200, 100) | 0.486 | 0.055 | 0.433 | 0.981 |
| | (600, 100) | 0.426 | 0.061 | 0.252 | 0.907 |

To further assess the performance of the two test statistics $T_F$ and $T_Z$, we also obtained the quantile plots of the two test statistics against the $F$-distribution and the $\chi^2$-distribution, respectively, with proper degrees of freedom. Figure 1 presents some representatives of such plots based on the simulated data with $p = 600$, $n = 100$ and 25% right-censored data. In the figure, the plots on the left side correspond to $T_F$ and the $F$-distribution, while the plots on the right side are for $T_Z$ and the $\chi^2$-distribution. The top two plots are for the case of $d = \{5\}$, one dimension, the two plots at the middle for the case of $d = \{5, 6\}$, two dimension, and the two plots at the bottom for the case of 10-dimensional $d = \{16 : 25\}$. They suggest that with the one-dimensional $d$, both $F$-distribution and $\chi^2$-distribution provide reasonable approximations to the distributions of $T_F$ and $T_Z$, respectively. When the dimension of $d$ is greater than one, the $F$-distribution approximation is still appropriate, but the $\chi^2$-distribution approximation is clearly not valid anymore.

**Table 3:** Empirical sizes and powers of the test procedure given in Zhong et al. [4] and the proposed one with the misspecified model.

| | | Sizes | | Powers | |
|---|---|---|---|---|---|
| $d$ | $(p, n)$ | $T_Z$ | $T_F$ | $T_Z$ | $T_F$ |
| | (100, 100) | 0.264 | 0.051 | 0.504 | 1 |
| {5} | (200, 100) | 0.256 | 0.052 | 0.523 | 1 |
| | (600, 100) | 0.242 | 0.049 | 0.523 | 1 |
| | (100, 100) | 0.492 | 0.047 | 0.589 | 1 |
| {5, 6} | (200, 100) | 0.487 | 0.065 | 0.553 | 1 |
| | (600, 100) | 0.503 | 0.053 | 0.586 | 1 |
| | (100, 100) | 0.672 | 0.053 | 0.583 | 1 |
| {6 : 10} | (200, 100) | 0.668 | 0.045 | 0.533 | 1 |
| | (600, 100) | 0.623 | 0.054 | 0.501 | 1 |
| | (100, 100) | 0.665 | 0.045 | 0.489 | 1 |
| {16 : 25} | (200, 100) | 0.616 | 0.040 | 0.433 | 1 |
| | (600, 100) | 0.527 | 0.055 | 0.354 | 0.998 |
| | (100, 100) | 0.524 | 0.044 | 0.460 | 0.949 |
| {16 : 35} | (200, 100) | 0.487 | 0.051 | 0.384 | 0.792 |
| | (600, 100) | 0.377 | 0.064 | 0.225 | 0.386 |

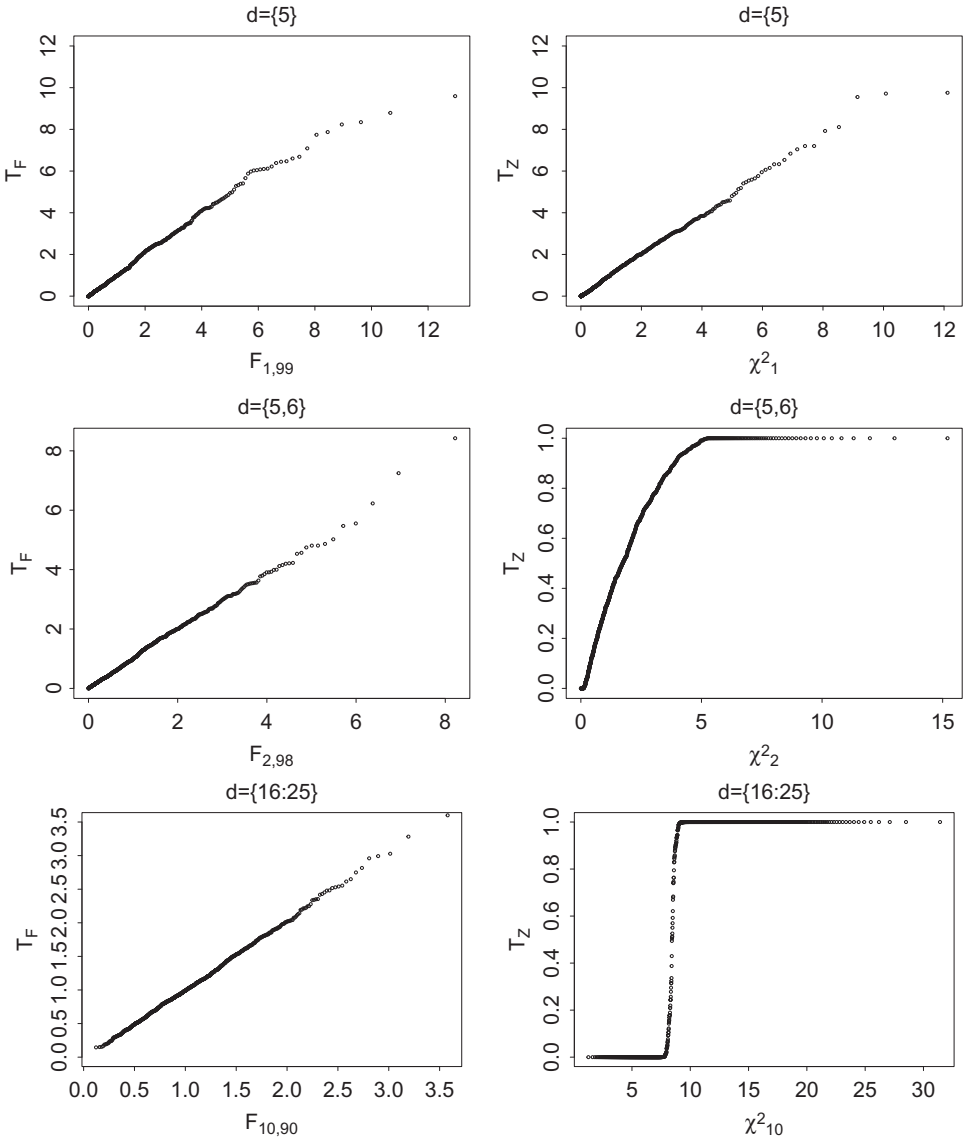


**Figure 1:** Q-Q plots for the proposed test statistic $T_F$ and the test statistic $T_Z$ given in Zhong et al. [4].

In practice, one possible question of interest is the sensitivity of the test procedure proposed in the previous sections with respect to model (1). To investigate this, we performed some simulation studies and Table 3 presents the results obtained exactly in the same way as those given in Table 1 except that the failure times were generated from the following Cox's proportional hazards model

$$\lambda(t|Z) = \lambda_0(t) \exp\left(\sum_{j=1}^{p} \beta_{0j} Z_j(t)\right)$$

instead of model (1), where all the function or parameters are defined as in model (1). One can see from the table that the results gave similar conclusions to those in Table 1 and in particular, they suggest that the proposed test procedure is still valid and does not seem to be sensitive to model (1). In addition, as with Figure 1, we also obtained the corresponding quantile plots of the two test statistics against the F-distribution and the $\chi^2$-distribution, respectively, with proper degrees of freedom, and again they gave similar conclusions.

# 5 An application

Now we apply the likelihood ratio test procedure proposed in the previous sections to a set of the data on the kidney cancer discussed in Sultmann et al. [18] and Zhong et al. [4]. The data set consists of 74 patients and in addition to the censored survival times (in months), some genomic and non-genomic information was also observed. Specifically, for each patient, 4,224 microarray gene expression was measured. For non-genomic factors, in addition to gender and age, the patients were classified into three groups based on their renal cell carcinoma (RCC), clear cell (ccRCC) type, papillary (pRCC) type and chromophobe cell (chRCC) type. One objective of interest is to evaluate the effects of the non-genomic factors on the survival rate given the genomic factors. In the following, by following others, we will focus on the 60 patients whose survival times are either observed or censored.

To perform the analysis, for each patient, we will define $X_1 = 1$ if the RCC type is ccRCC and $X_1 = 0$ otherwise, $X_2 = 1$ if the RCC type is pRCC and $X_2 = 0$ otherwise, and $G = 1$ if the patient is male and $G = 0$ otherwise. In addition, we will use $Ag$ to denote the age of the patient and $\mathbf{Z}_2$, a 4224-dimensional vector, to denote the gene expression. By using the notation above, model (1) becomes

$$\lambda(t|Z) = \lambda_0(t) + \beta_1 X_1 + \beta_2 X_2 + \beta_3 G + \beta_4 Ag + \beta_5 \mathbf{Z}_2,$$

and the main goal is to test the hypothesis $H_0$ with $d = \{1, 2, 3, 4\}$. In the above, $\mathbf{Z} = (X_1, X_2, G, Ag, \mathbf{Z}_2^\top)^\top$ and $\beta_5$ is a 4224-dimensional vector of unknown parameters.

**Table 4:** The $p$-values obtained for kidney cancer data.

| $L$ = | Procedure in Zhong et al. [4] | The proposed procedure |
|---|---|---|
| c(5 : 10) | 0.8732 | 0.5024 |
| c(5 : 20) | 0.8732 | 0.5057 |
| c(5 : 30) | 0.8732 | 0.5014 |
| c(5 : 50) | 0.8732 | 0.4990 |
| c(5 : 100) | 0.8732 | 0.4939 |

Table 4 gives the average $p$-values obtained by the proposed likelihood ratio test for testing $H_0$ based on 10,000 sets of the generated $\tilde{\epsilon}_i$'s. To see the possible effect of the selection $L$ on the results, we considered several choices for $L$ including $L = \{5 : 10\}$, $L = \{5 : 20\}$, $L = \{5 : 30\}$, $L = \{5 : 50\}$, and $L = \{5 : 100\}$. For comparison, we also included the $p$-values given by the test procedure proposed in Zhong et al. [4]. One can see from the table that both test procedures suggest that conditional on the genomic factors, the non-genomic factors did not seem to have any significant effects on the patient's survival rate. In addition, the results seem to indicate that the selection of $L$ did not seem to have much effect on the test results.

# 6 Discussion and conclusion remarks

In the previous sections, we have considered a group test problem in the high-dimensional situations with the response variable of interest being a failure time arising from the additive hazards model. As discussed above, the problem occurs quite often and in many areas, especially in genetic studies or genomics where the determination or identification of a group of genes significantly related to a certain disease is often of interest. Also due to the dimensionality, traditional methods cannot be apply. For the problem, corresponding to the approach given in Zhong et al. [4], a new partition algorithm was presented. Furthermore, a likelihood ratio test procedure was developed and the numerical studies suggested that the proposed approach seems to perform well in practical situations and gives better performance than that given in Zhong et al. [4].

More research is clearly needed for the investigation of some issued related to the problem discussed here and the generalization of the proposed test procedure to other situations. One is the goodness-of-fit test for model (1). In standard failure time data or the data with low dimensions, some procedures have been developed for checking the appropriateness of the additive hazards model (1) and one typical way is to apply some residual-based statistics [19]. For the high-dimensional situation considered here, however, it does not seem to exist an established procedure for checking model (1) or other commonly used regression models for failure time data. It is apparent that one can define residuals similarly but the similar statistics may not have similar properties in the high-dimensional situations.

Note that in the preceding sections, the focus has been on the group test and it is apparent that one may be also interested in estimation of regression parameters. For this, it is clear that one can employ some existing penalized methods or develop some new methods based on the new partition procedure. In the proposed approach, it has been assumed that the failure time of interest follows the AHM and in practice, this may not be true. In other words, it may be useful to develop similar approaches for the situations where the failure time follows other models such as the proportional hazards model or linear transformation model. However, such generalizations may not be straightforward since under these situations, the relationship (4) may not exist. Finally, of course, the development of some theoretical justification for the proposed test procedure would be helpful too.

**Acknowledgment:** The authors wish to thank the Editor and two reviewers for their helpful comments and suggestions, which greatly improved the paper.

**Funding:** Jiang's work was partly supported by the National Natural Science Foundation of China, Project No. 11471140.

## References

1. Foster JC, Liu D, Albert PS, Liu A. Identifying subgroups of enhanced predictive accuracy from longitudinal biomarker data using tree-based approaches: applications to monitoring fetal growth. J R Stat Soc Ser A 2016. DOI:10.1111/rssa.12182.
2. Lin HZ, Li Y, Tan M. Estimating a unitary effect summary based on combined survival and quantitative outcomes in clinical trials. Comput Stat Data Anal 2013;66:129–39.
3. Liu Z, Bensmail H, Tan M. Efficient feature selection and multiclass classification with integrated instance and model based learning. Evol Bioinf 2012;8:1-10.
4. Zhong PS, Hu T, Li J. Tests for coefficients in high-dimensional additive hazard models. Scand J Stat 2015;42:649–64.
5. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data, 2nd ed. Ne York: John Wiley, 2002.
6. Lin DY, Ying Z. Semiparametric analysis of the additive risk model. Biometrika 1994;81:61–71.
7. Gaiffas S, Guilloux A. High-dimensional additive hazards models and the Lasso. Electron J Stat 2012;6:522–46.
8. Lin W, Lv J. High-dimensional sparse additive hazards regression. J Am Stat Assoc 2013;108:247–64.
9. Martinussen T, Scheike TH. Covariate selection for the semiparametric additive risk model. Scand J Stat 2009a;36:602–19.
10. Martinussen T, Scheike TH. The additive hazards model with high-dimensional regressors. Lifetime Data Anal 2009b;15:330–42.
11. Goeman J, Geer VD, Houwelingen V. Testing against a high-dimensional alternative. J R Stat Soc Ser B 2006;68:477–93.
12. Zhong PS, Chen SX. Tests for high dimensional regression coefficients with factorial designs. J Am Stat Assoc 2011;106:260–74.
13. Goeman J, Houwelingen V, Finos L. Testing against a high dimensional alternative in the generalized linear model: asymptotic type I error control. Biometrika 2011;98:381–90.
14. Goeman J, Oosting J, Cleton-Jansen AM, Anninga JK, van Houwelingen HC. Testing association of a pathway with survival using gene expression data. Bioinformatics 2005;21:1950–7.
15. Lan W, Wang H, Tsai CL. Testing covariates in high dimensional regression. Ann Inst Stat Math 2014;66:279–301.
16. Wang S, Cui H. Generalized F-test for high dimensional linear regression coefficients. J Multivariate Anal 2013;117:134–49.
17. Anderson TW. An introduction to multivariate statistical analysis, 2nd ed. New Jersey: John Wiley & Sons, 2003.
18. Sultmann H, Heydebreck A, Huber W, Kuner R, Bune A, Vogt M, et al. Gene expression in kidney cancer is associated with cytogenetic abnormalities, metastasis formation, and patient survival. Clin Cancer Res 2005;11:646–55.
19. Klein JP, Moeschberger ML. Survival analysis: Techniques for censored and truncated data, 2nd ed. New York: Springer, 2003.