Analysis of High Dimensional Gene Data Combining Correlation Principal Component Regression and Additive Risk Model

Yichuan Zhao and Guoshen Wang
Department of Mathematics and Statistics, Georgia State University, Atlanta, GA, 30303, USA
matyiz@langate.gsu.edu, gwang5@student.gsu.edu

Abstract
One problem of interest is to relate genes to survival outcomes of patients for the purpose of building regression models to predict future patients’ survival based on their gene expression data. Applying semiparametric additive risk model of survival analysis, we propose a new approach to conduct the analysis of gene expression data with the focus on model’s predictive ability. The method modifies the correlation principal component regression to handle the censoring problem of survival data. In addition, we employ the time dependent AUC and RMSEP to assess how well the model predicts the survival time. Furthermore, the proposed method is able to identify significant genes that are related to the disease. Finally, this proposed approach is illustrated by the diffuse large B-cell lymphoma (DLBCL) data set. The results show that the model fits the data set very well.

Keywords: Correlation principal component regression, Additive risk model, Gene expression data, Right censoring

1. Introduction
One of main advantages of microarray technology is that it permits simultaneous measurements of expression levels of thousands of genes. In the analysis of microarray gene expression data, one of the major challenges to statisticians is the number of variables (genes) needed to deal with. The other is how to model the survival data with high dimensional covariates due to the presence of censoring.

[1] suggested to first apply the PLS approach to gene expression and survival data to reduce the dimension and then fit the Cox regression model to the resulting PLS components with the survival data. [2] proposed a method to transform the Cox survival analysis to a generalized linear regression problem and then apply the PLS techniques to the transformed problem. [3] also developed a clever analysis procedure that allows one to perform the PLS analysis under the Cox model. [4] suggested to first apply the PLS approach to the gene expression and survival data to reduce the dimension and then fit the additive risk model to the resulting PLS components with the survival data. [5] have employed PLS to accelerated failure time (AFT) model to re-analyze Michigan lung cancer data. The proposed mean imputation algorithm is a very clever way to handle right censored life time. [6] applied the standard principal component regression (SPCR) to survival data and use the resulting PCR in the additive risk model. [7] applied a variant correlation principal component regression (CPCR) for the Cox survival analysis. So far, there is no method which has been applied to semiparametric additive risk model based on CPCR.

In this paper, we propose a new approach for the semiparametric additive risk model in survival analysis. Instead of PCA or PLS, the approach uses a variant of CPCR, which is the correlation principal component additive regression (CPCAR). Details of the methodology are presented in Section 2. In Section 3, we introduce the additive risk model and explicit solution for the coefficients of the additive risk model, and the main procedure to obtain CPCAR components. We also discuss model assessment procedure that applies time dependent AUC and root mean squared error for prediction (RMSEP). In Section 4, we apply the proposed method to the diffuse large B-cell lymphoma (DLBCL) data set and compare the results by using different models. In Section 5, conclusion and discussion are included.

2. Correlation principal component regression
The standard principal component regression (SPCR) has been used as a regression technique for a long time mainly for the situation where there exist strong multi-collinearities among predictor variables and/or the number of predictor variables is close to or larger than the sample size such as in gene expression data. The idea of principal component regression is to regress a response variable on principal components ordered by their variability rather than original predictor variables, while the idea behind the correlation principal component regression (CPCR) is to regress a response variable on principal components ordered by their correlation with the response variable. In both cases, the number of principal
components used in the final regression model is typically much less than the number of predictor variables.

SPCR assumes that the main information of interest is contained in the directions of the predictor space with high variations. In some cases, however, the high variations may be generated by sources that are not related to the response variable under study and it is common that a component with lower variance is a good predictor in a regression model ([9]). In these situations, it is natural to use only the principal components that represent relevant directions in the regression as in CPCAR. There always exists noise in the measurement, and thus the more principal components are used in the model, the larger noise is incorporated. This is partly the motivation for CPCAR, which employs only the relevant principal components (with large correlations) and ignores the irrelevant principal components.

3. Main procedure

3.1 Additive risk model

Suppose that one observes right-censored survival data given by \( \{Y_i = \min(T_i, C_i), \delta_i = I(Y_i = T_i), X_i; i = 1, ..., n\} \), where \( T_i \) denotes the survival time, \( C_i \) the censoring time independent of the \( T_i \), and \( X_i \) the vector of covariates or genes associated with subject \( i \) in the study. Let \( N_i(t) = \delta_i I(Y_i \leq t) (i = 1, ..., n) \) be a counting process for the \( i \)th subject, which indicates that the failure time of the \( i \)th subject is observed up to time \( t \). Let \( W_i(t) = I(Y \geq t) \) denote the predictable process indicating whether or not the \( i \)th subject is at risk just before time \( t \). Let \( \tau \) satisfies \( P(Y > \tau) > 0 \), \( \lambda(t | X_i) \) denote the hazard function for the life time \( T \) under covariate \( X_i(t) \). The semiparametric additive risk model ([8]) has the following form:

\[
\lambda(t | X_i) = \lambda_0(t) + \beta^T X_i, \tag{3.1}
\]

where \( \lambda_0(t) \) is an unknown baseline hazard function corresponding to subject with \( X_i = 0 \), \( \beta \) is the \( p \)-vector of unknown regression parameters. The objective is to make inferences about \( \beta \) and predict future survival.

Proposed by [8], the regression coefficients \( \beta \) is estimated by

\[
\hat{\beta} = \left[ \sum_{i=1}^{n} W_i(t)(X_i(t) - \bar{X}(t))^2 dt \right]^{-1} \left[ \sum_{i=1}^{n} \int (X_i(t) - \bar{X}(t)) dN_i(t) \right] \tag{3.2}
\]

where \( \bar{X}(t) = \sum_{i=1}^{n} \frac{W_i(t)X_i(t)}{\sum_{i=1}^{n} W_i(t)} \)

3.2 The proposed procedure

The proposed procedure is similar to that given in [7], instead of using Cox survival model, we propose to use additive risk model. Using CPCAR to first reduce the dimension of gene expression data by finding the principal components and order them based on the strength of their association with the survival time. Specifically, Let \( X = (x_1, ..., x_n) (n\times p) \) denote the gene expression data matrix, \( Y (n \times 1) \) the response vector of interest representing survival times to a certain event such as death due to cancer or censoring times, and \( \delta (n \times 1) \) the vector of censoring indicators. We apply CPCAR to \( X, Y, \) and \( \delta \) to obtain the principal components \( Z^*_1 = (z_1^*, ..., z_n^*) \), ordered from the smallest to largest based on the \( p \)-values from testing \( \beta = 0 \) under model (3.1) when regressing \( Y \) on the principal components individually. Note that we propose to use \( p \)-values instead of correlation coefficients between the true survival time and individual components because the coefficients cannot be computed due to censoring. For \( X \) with centered columns, \( A = n-1 \). For each \( K (1 \leq k \leq A) \), replacing predictor variables in model (3.1) with \( Z^*_k = (z_1^*, ..., z_n^*) \) gives

\[
\hat{\lambda}(t | Z_k^*) = \lambda_k(t) + \beta_k^T Z_k^*(t) = \lambda_k(t) + f(X), \tag{3.3}
\]

\[
f(X) = \beta_k^T Z_k^*(t), \tag{3.4}
\]

where the risk score function \( f(X) \) is a linear combination of the original data matrix \( X \). Then one applies (3.2) to obtain an estimate \( \hat{\beta}_k \) in model (3.3). The estimate, say \( \hat{\beta}_k \), of the original \( \beta \) in model (3.1) can then be obtained by transforming \( Z_k^* \) back to \( X \).

3.3 Model evaluation

The predictive performance of fitted model can be evaluated using root mean squared error for prediction (RMSEP), and area under a time dependent ROC curve (AUC).

To obtain the RMSEP, let \( R \) denote a random subset of \( \{1, ..., n\} \) with \( m \) elements and \( \hat{\beta}_{K,R} \) the partial likelihood estimator obtained based on training data \( \{(Y_{R}, Z^*_{K,R})\} \), where \( \{(Y_{R}, Z^*_{K,R})\} \) denotes \( \{(Y, Z^*)\} \) with the information on subjects \( i \in R \) removed. That is, \( \hat{\beta}_{K,R} \) is \( \hat{\beta}_K \) with the information from the subjects \( i \in R \) removed. For additive risk model, for each \( i \in R \) in the validation set, define the martingale residual as

\[
M_{i,t} = \delta_i - \sum_{a} \int \left[ M_{i,t} \left[ (\hat{\beta}_{K,R}^a X_i(t) + \hat{\beta}_{K,a} Z_{K,a}^*(t)) dt \right] \right], \tag{3.5}
\]

For the Cox regression model, following [7], for each \( i \in R \) in the validation set, define the martingale residual as
\[ M_{i,R} = \delta_i - \int_0^\infty I(Y_i \geq t) \exp(\hat{\beta}_{X,R}^* Z_{X,R,i}) \, d\bar{N}_i(t, \hat{\beta}^*_{X,R}), \quad (3.6) \]

Then we propose to define the prediction error of a model as

\[ \text{RMSEP}(K, m) = \left[ \frac{1}{m} \sum_{i=1}^m M_{i,R}^2 \right]^{1/2}, \quad (3.7) \]

the delete-m root mean square error of prediction.

The second criterion is to use AUC, originally developed by [10], following [3], we can define

Sensitivity (c, t | f(X)) = P { f(X) > c | \delta(t) = 1 }, \quad (3.8)

and

Specificity (c, t | f(X)) = P { f(X) \leq c | \delta(t) = 0 }, \quad (3.9)

where f(X) is a given risk score function illustrated in (3.3), c is a cutoff value from risk score, t is a time point varying from (-\infty, +\infty). We define the ROC curve for any time t as the plot of sensitivity vs. 1 - specificity with cutoff point c varying. The larger the AUC is, the better the model performance is.

4. Application to DLBCL data set

To illustrate the procedure presented in the previous section, we compare the partial Cox regression in [3], referred as PCR method, correlation principal component Cox regression (CPCCR) methods in [7], supervised principal component analysis (SPCA) in [11], and our proposed correlation principal component additive regression (CPCAR). We reanalyzed the DLBCL gene expression data that were analyzed by [3] among others. The dataset includes gene expression measurements of 7399 genes on 240 patients with DLBCL and their survival times. Among the 240 patients, 138 (57.5%) deaths were observed during the study with the median death time of 2.8 years and, for others, right-censoring times were observed.

Table 1 contains a comparison among the partial Cox regression (PCR), correlation principal component Cox regression (CPCCR), and correlation principal component additive regression (CPCAR). The first column of Table 1 shows the component number. The second column of Table 1 shows the p-values from univariate additive risk model for the first 10 components with the whole 7399 genes from training data set. As we can see, the first 7 CPCAR components are significant at \( \alpha = 0.05 \) level. As a comparison, we also list p-values for other methods in the same table.

To evaluate how well the additive risk model with CPCAR components, Cox model with PCR components as described in [3], Cox model with correlation principal components as described in [7], supervised principal component analysis (SPCA) predict the survival time, we built several models using the training set with different k, and predicted the survival probabilities for patients in the testing data set. Based on (3.4), we estimated the risk score for each patient in the testing set. By choosing the mean risk score of the patients in the testing set as the cutoff point, we divided those patients into high risk group and low risk group and plot the Kaplan-Meier survival curves.

Figure 1 A to D shows the Kaplan-Meier survival curves for the two risk groups with different methods. (C, D are not shown due to space limit) The Chi-square test p-value of the difference between two groups is \( p = 0.0518 \) with CPCAR analysis, and 0.284 with CPCCR, 0.983 with PCR. The CPCAR method splits the two groups more significantly than CPCCR, PCR and SPCA do. With CPCCR, the two groups are even not separated from each other. Figure 1 E shows the areas under the ROC curve with time varying. One can observe the additive risk model performs the best prediction on the testing patients. Other methods may be OK for this data set.

<table>
<thead>
<tr>
<th>K</th>
<th>CPCAR</th>
<th>CPCCR</th>
<th>PCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.551791e-05</td>
<td>1.4e-07</td>
<td>2.3e-15</td>
</tr>
<tr>
<td>2</td>
<td>1.000816e-03</td>
<td>7.2e-04</td>
<td>2.6e-01</td>
</tr>
<tr>
<td>3</td>
<td>4.599242e-03</td>
<td>2.6e-03</td>
<td>7.5e-02</td>
</tr>
<tr>
<td>4</td>
<td>3.190365e-03</td>
<td>2.1e-03</td>
<td>1.1e-02</td>
</tr>
<tr>
<td>5</td>
<td>2.241105e-02</td>
<td>2.2e-02</td>
<td>1.2e-04</td>
</tr>
<tr>
<td>6</td>
<td>1.787239e-03</td>
<td>1.3e-02</td>
<td>8.1e-02</td>
</tr>
<tr>
<td>7</td>
<td>1.894647e-02</td>
<td>1.5e-02</td>
<td>1.8e-02</td>
</tr>
<tr>
<td>8</td>
<td>1.569958e-01</td>
<td>6.8e-02</td>
<td>1.1e-01</td>
</tr>
<tr>
<td>9</td>
<td>1.905000e-02</td>
<td>1.5e-02</td>
<td>1.8e-04</td>
</tr>
<tr>
<td>10</td>
<td>1.116449e-03</td>
<td>4.7e-02</td>
<td>7.8e-01</td>
</tr>
</tbody>
</table>

Table 1: Comparison of the p-values of first 10 components from CPCAR, CPCCR, and PCR on DLBCL data set

![Graph](image-url)
In addition, we set up criteria as proposed by [5], in which genes are selected based on the largest absolute coefficients of normalized expressions in the semiparametric additive risk model. We finally get overlapped genes under at least two different CPCAR components model. (Results are not shown here due to space limit).

**5. Conclusion and discussion**

No matter what are the purposes of the gene expression data analysis, one has to deal with this multi-collinearity first by transformation and/or dimension reduction.

This paper proposes an approach to the analysis of gene expression data and it combines CPCR and partial likelihood approach together under additive risk model.

Note that due to censoring, p-values are used in determining correlation instead of correlation coefficient calculated by treating censoring time as survival time or using imputed survival time. The model selection criteria, RMSEP (RMedSEP) and AUC, were also discussed. Note that if estimation of \( \beta \) in the semiparametric additive risk model is the main objective instead of prediction, one may wish to use a different model selection criterion, such as, Akaike's information criterion (AIC), or Bayesian information criterion (BIC).

The proposed CPCAR chooses PCs with strongest association with the survival time to be included in the semiparametric additive risk model. A direction for future investigation may focus on different ways of selecting PCs. In addition, we will investigate AFT model, semiparametric transformation model, Aalen’s additive hazards model, and additive-multiplicative hazard model in the future.

**ACKNOWLEDGMENT**

The authors would like to thank the two reviewers for helpful comments.

**References**


[5] Datta, S. Rademacher, L. and Datta, S. Predicting patient survival from microarray data by Accelerated failure time modeling using partial least squares and LASSO. Biometrics. 2007 Mar; 63(1) 259-


[7] Zhao, Q. and Sun, J. Cox survival analysis of microarray gene expression data using correlation principal component regression, Statistical Applications in Genetics and Molecular biology: Vol. 6: 2007, Iss. 1, Article 16


