Statistical Applications in Genetics and Molecular Biology

Volume 9, Issue 1

2010

Article 41

Predicting Patient Survival from Longitudinal Gene Expression

Yuping Zhang*

Robert J. Tibshirani[†]

Ronald W. Davis[‡]

*Stanford University, yupingz@stanford.edu

[†]Stanford University, tibs@stat.stanford.edu

[‡]Stanford University, dbowe@stanford.edu

Predicting Patient Survival from Longitudinal Gene Expression*

Yuping Zhang, Robert J. Tibshirani, and Ronald W. Davis

Abstract

Characterizing dynamic gene expression pattern and predicting patient outcome is now significant and will be of more interest in the future with large scale clinical investigation of microarrays. However, there is currently no method that has been developed for prediction of patient outcome using longitudinal gene expression, where gene expression of patients is being monitored across time. Here, we propose a novel prediction approach for patient survival time that makes use of time course structure of gene expression. This method is applied to a burn study. The genes involved in the final predictors are enriched in the inflammatory response and immune system related pathways. Moreover, our method is consistently better than prediction methods using individual time point gene expression or simply pooling gene expression from each time point.

KEYWORDS: prediction, time course, gene expression, survival

*We wish to acknowledge the efforts of many individuals at participating institutions of the Glue Grant Program that generated the clinical and genomic data reported here. This study was supported by NIH U54 GM-062119 and P01-HG000105.

1 Introduction

Microarray technology is dramatically changing studies in molecular biology, permitting the generation of high-throughput data on a global scale. In particular, one of the most exciting areas of application has been its use in clinic studies. Microarray technology helps investigators gain a fresh understanding of disease processes at the molecular level, and aids in the classification of diseases as well as the identification of new subtypes of diseases. Furthermore, depending on an individual's genetic information, investigators will be able to diagnose and prognose, as well as come up with a treatment plan. This will lead to the realization of "personalized medicine".

In methodology, finding associations between expression profiles and phenotypic data is the key to applying gene expression to a clinical study. In the stationary gene expression scenario, the main challenge is the well-known " $p \gg n$ " paradigm. Dimension reduction methods have been studied extensively when mRNA abundances are measured from DNA microarrays at individual time points. In summary, two strategies are distinguished in dimension reduction: feature selection and feature extraction. Partial least squares regression, see Nguyen and Rocke (2002), Aastveit and Martens (1986), Martens and et al. (1986), Jacobsen, Kolset, and et al. (1986), Cramer, Bunce, and et al. (1988), Helland (1990), Faigle, Poppi, and et al. (1991), Wakeling and Morris (1993), Yeh and Spiegelman (1994), Young (1994), Holcomb, Hjalmarsson, and et al. (1997), Chun and Keles (2010) and ridge regression Jain (1985), Lecessie and Vanhouwelingen (1992), Hoerl and Kennard (1990), Vangaans and Vriend (1990), Pliskin (1990), Hertz (1991), Gana (1995), Forrester and Kalivas (2004), Hu (2005), Farkas and Heberger (2005), Maruyama and Strawderman (2005), Ozkale (2008), Akdeniz and Tabakan (2009) are techniques of dimension reduction. Compared with the L_2 - *penalty* used in ridge regression, Tibshirani developed the Lasso method using L_1 - *penalty* in 1996, see Tibshirani (1996). The L_1 - *penalty* has a feature selection procedure that shrinks some of the coefficients to zero. In 2002, Tibshirani et al proposed a nearest shrunken centroid method to classify subtypes of a disease, see Bair and Tibshirani (2004). Another competitive method is the supervised principal component analysis (SPC) which was developed by Bair, Hastie, and et al. (2006). It contains two dimension reduction steps. This method first performs an univariate supervised feature selection, then constructs the principal components accompanied by a feature extraction. This method can be used in both quantitative outcome and survival outcome scenarios. These stationary gene expression data based methods have been successfully applied to identify biomarkers and predict patient outcomes especially for cancer disease, see Korde, Lusa, and et al. (2010), Desmedt, Haibe-Kains, and et al. (2009), Pandit, Kennette, and et al. (2009), Pandit et al. (2009), Karlsson,

Delle, and et al. (2008), Schorge, D., and et al. (2009), Naderi, Teschendorff, and et al. (2007), Lin, Friederichs, and et al. (2007), Schneeweiss, Lichter, and et al. (2006), Nimeus-Malmstrom, Ritz, and et al. (2006).

In clinical studies, more and more investigators monitor the status of patients and measure the longitudinal gene expression. In the study of longitudinal gene expression, many methods have been employed that extract temporal patterns of differential expression. Storey, Xiao, and et al. (2005) used a spline-based approach to detect changes in expression over time within a single biological group or to detect differences of the behavior of expression over time between two groups. Yuan (2006), Yuan, Li, and et al. (2008) applied hidden Markov models to analyze microarray time course data under multiple biological conditions. Tai (2006) developed a multivariate empirical Bayes statistic to detect differentially expressed genes. Ma, Zhong, and Liu (2009) proposed a functional ANOVA mixed-effect model to model time course gene expression observations and identify differentially expressed genes. Zhou, Xu, Herndon, Tompkins, Davis, Xiao, and Wong (2010) proposed TANOVA to simultaneously handle the time course and factorial structure in microarray data. However, to our knowledge, no prediction method using the longitudinal structure has been published yet.

In this paper, we propose a novel prediction framework using longitudinal gene expression. We applied our method to the survival prediction of burn patients. By utilizing the longitudinal structure, we obtain much better performance of prediction than using the individual time points or simply pooling the gene expression from all time points.

2 Survival prediction using longitudinal gene expression

We assume there are p features (e.g. genes) measured on N observations across t time points. Let X_g denote the $N \times t$ matrix, which is the expression matrix of gene g across t time points. Let Y be an N -vector of outcome measurements. The outcome can be the survival times subject to censoring. To make use of the time course information, we evaluate gene response based on information pooled across time. Specifically, we search for a direction in the t -dimensional time space that has the strongest response signal of interest, and extracts a predictor based on the projection of this direction. We name the direction “optimal direction”. This direction captures the gene response related to the outcome. The paragraphs below are ordered as follows. First, we give the estimation of “optimal direction”. Second, we introduce the procedure for model selection and the extraction of final predictors.

2.1 Estimating the optimal direction

The unknown optimal direction is gene specific and dependent upon the outcome variable of interest as well as the objective function we choose. To estimate the direction, we first define an appropriate objective function. The objective function is required to capture the correlation between the outcome and longitudinal gene expression. We propose a gene specific multivariate Cox regression model to capture the correlation between time course gene expression and survival outcome. For each gene, we treat gene expression of each time point as a variable. Consider a model in which we start with a baseline hazard function $\lambda_0(t)$. We then want to model the influence of time course gene expression on this baseline hazard. We ignore the subscript of genes. We use an exponential function to model the influence, i.e. $\lambda(t) = \lambda_0(t)\exp(X\beta)$, X is an $N \times t$ matrix. As showed in this model, the hazards are proportional, which was first proposed by Cox (1972). Let R_j denote the risk set at time y_j , that is, $\{y_i : y_i \leq y_j, i \in \{1, \dots, N\}\}$. The Cox proportional regression model relates the time passes before the event (death of patients) occurs to the covariates (gene expression) that may be associated with that quantity. The unique effect of a unit increase in gene expression is multiplicative with respect to the hazard rate. The Cox proportional model avoids making potentially untenable distributional assumptions about the hazard based on conditional probabilities – given “an event happened at a particular time y_j ”, what the probability of observing l (with covariates X_l) is.

$$\frac{\lambda_l(y_j)}{\sum_{i \in R_j} \lambda_i(y_j)} = \frac{h_0(y_j)\exp(X_l\beta)}{\sum_{i \in R_j} h_0(y_j)\exp(X_i\beta)} = \frac{\exp(X_l\beta)}{\sum_{i \in R_j} \exp(X_i\beta)}, \quad (1)$$

where the numerator of the most left fraction is the hazard for individual l at time y_j , the denominator is the sum of all the hazards at y_j for all the individuals who were at risk at y_j .

Each observed event contributes one term like above to the partial likelihood. The log partial likelihood is shown as follows:

$$l(\beta) = \sum_{y_i \text{ uncensored}} \{X_i\beta - \log[\sum_{y_i \leq y_j} \exp(X_j\beta)]\}, \quad (2)$$

where $i, j \in \{1, \dots, N\}$.

We use the above log partial likelihood function as the objective function and seek to find a direction to maximize this objective function. Let $\hat{\beta}$ denote the estimated coefficients by maximizing the log partial likelihood function. The amplitude of $\hat{\beta}$ reflects the relationship between the hazard function and gene expression from each time point. Signs of $\hat{\beta}$ reflect the relationship among gene expression of all time points. We then project the gene expression of individual time points

to this direction and obtain the weighted gene expression i.e. $\hat{x} = X\hat{\beta}$. We use the estimated weighted $\hat{x}$ as a new feature representing that gene. The weighted gene expression matrix $\hat{X}$ consist of p vectors $\hat{x}$ with length of N .

2.2 Prediction model

After projecting the gene expression of individual time points to the “optimal direction”, one can obtain the weighted gene expression. The relationship between survival and the weighted gene expression varies across genes. If we could discover the underlying genetic factors of the patients, often reflected by a group of genes acting together in pathways, then we would do a better job of predicting patient survival. Thus, variable selection is necessary. Traditional variable selection techniques for stationary gene expression data can be applied to the weighted gene expression. Among the many variable selection methods, SPC and the Lasso are two of the most important ones. Bair et al. (2006) extended and modified the SPC by preselecting a subset of genes based on a score statistic. In survival prediction, the score statistic for predictor g can be chosen as

$$s_g = \frac{U_g(0)^2}{I_g(0)},$$

where $U_g = l'_g$, $I_g = -l''_g$ and l_g is the partial likelihood relating the data for a single predictor and the outcome. The Lasso is a shrinkage and selection method for linear regression. It minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients. It has connections to soft-thresholding of wavelet coefficients, forward stagewise regression, and boosting methods. Instead of minimizing the sum of squared errors, Park and Hastie (2007) and Friedman, Hastie, and et al. (2010) extended the Lasso technique in Tibshirani (1996) to the Cox regression model by maximizing the log-likelihood function. Let $\{(\hat{x}_i, y_i, \delta_i) : \hat{x}_i \in R^p, y_i \in R^+, \delta_i \in \{0, 1\}, i = 1, \dots, N\}$ be N triples of p estimated weighted gene expression, a response indicates the survival time and a binary indicator $\delta_i = 1$ for complete (died) observations and $\delta_i = 0$ for right-censored patients. R_j is the risk set at time y_j . The objective function with L_1 constraint is showed as following.

$$l(\alpha) = \sum_{i=1}^N \delta_i \alpha' x_i - \sum_{i=1}^N \delta_i \log \left\{ \sum_{j \in R_j} \exp(\alpha' x_j) \right\}, \quad (3)$$

subject to $\sum |\alpha_i| \leq s$ Coefficients are found to maximize the above objective function. Variable selection is archived via the coordinate descent algorithm. Variable with non-zero coefficients will be selected as the predictors.

We can use either of these two methods to do variable selection based on the weighted gene expression. In the supervised principal component analysis scenario, the algorithm is as follows:

- Compute the multivariate Cox proportional hazard regression coefficients for all time points of each feature. Obtain the weighted predictors $\hat{X}$, with X^i denoting the usual vector of predictors $(\hat{x}_1, \dots, \hat{x}_p)$ for the i^{th} individual.
- Center the weighted gene expression
- Compute the score statistic for each gene based on the weighted gene expression.
- Form a reduced data matrix consisting of only those genes whose score statistic exceeds a threshold θ in absolute value (θ is estimated by cross-validation).
- Compute the first (or first few) principal components of the reduced data matrix.
- Use these principal component(s) in a Cox proportional hazard regression model to predict the outcome.

To choose the value of the tuning parameter θ , we use two-fold cross-validation. One note about cross-validation is that we need to recalculate the coefficients in cross-validation, but not use the coefficients obtained from the training data. For selecting an appropriate value of θ , we recalculate the $\hat{X}^*$ using the coefficients $\hat{\beta}$ obtained in the whole training data. We then center each component of $\hat{X}^*$ using the means derived from the training data $\hat{X}$. Instead of doing SVD on $\hat{X}^*$, we derive the predictor using the right singular vectors derived from training data.

Alternatively, we can choose the Cox regression model with the Lasso penalty Park and Hastie (2007), Friedman et al. (2010) to select variables based on the weighted gene expression. The use of the L_1 – *penalty* Cox regression model in the second stage is carried out as follows:

- Compute the multivariate Cox proportional hazard regression coefficients for all time points of each feature. Obtain the weighted predictors $\hat{X}$, with X^i denoting the usual vector of predictors $(\hat{x}_1, \dots, \hat{x}_p)$ for the i^{th} individual.
- Variable selection is archived by maximizing the following objective function.

$$l(\alpha) = \sum_{i=1}^N \delta_i \alpha' x_i - \sum_{i=1}^N \delta_i \log \left\{ \sum_{j \in R_j} \exp(\alpha' x_j) \right\}, \quad (4)$$

subject to $\sum |\alpha_i| \leq s$. The tuning parameter s is estimated by cross-validation.

- The predictors with non-zero $\hat{\alpha}$ are the selected predictors. Use the selected non-zero predictors in a Cox proportional hazard regression model to predict the outcome.

Given the test data, we project the time course gene expression of the test data and obtain the weighted gene expression using the optimal direction estimated from the training data. We use these weighted gene expressions in a Cox proportional model. The coefficients in the Cox proportional model are the estimated coefficients $\hat{\alpha}$ obtained from training data. To separate patients into two groups, we calculate the mean of $\alpha' \hat{X}_{train}$ and use this value as a cutoff. For each patient in the test data set, we compare $\alpha' \hat{X}_{test}$ with the cutoff. Patients with $\alpha' \hat{X}_{test} > E(\alpha' \hat{X}_{train})$ belong to the “high-risk” group; the rest belong to the “low-risk” group. We use the log-rank statistic to test the significance of the separation. We call our method using SPC in the model selection step as TSPC and use Lasso in model selection as TCoxL1.

3 Results

3.1 Simulation study

We performed a simulation study to validate the performance of our method. The simulated data set X^k at time point k consisted of 1000 “genes” (rows) and 100 “patients” (columns). Here we considered three time points. Let x_{gj}^k , $k \in 1, 2, 3$ denote the “expression level” of the gene g and patient j at time point k . We generated the data as follows:

$$x_{gj}^0 = \begin{cases} 6 + \varepsilon_b & g \leq 100, j \leq 50 \\ 6.1 + \varepsilon_b & g \leq 100, j > 50 \\ 3.5 + 0.2I(\mu_j < 0.4) + \varepsilon_b & 200 < g \leq 300 \\ 3.5 + \varepsilon_b & \text{else,} \end{cases}$$

$$x_{gj}^1 = \begin{cases} 0.5 + \varepsilon_t + x_{gj}^0 & g \leq 100, j \leq 50 \\ -0.3 + \varepsilon_t + x_{gj}^0 & g \leq 100, j > 50 \\ \varepsilon_t + x_{gj}^0 & \text{else,} \end{cases}$$

and

$$x_{gj}^2 = \begin{cases} -0.8 + \varepsilon_t + x_{gj}^1 & g \leq 100, j \leq 50 \\ -0.7 + \varepsilon_t + x_{gj}^1 & g \leq 100, j > 50 \\ \varepsilon_t + x_{gj}^1 & \text{else} \end{cases}$$

Here, $\varepsilon_b \in N(0, 2.5)$ and $\varepsilon_t \in N(0, 2)$. The μ_j is a uniform random variable on $(0, 1)$ and $I(x)$ is an indicator function. We introduced the time course structure in genes indexed from 1 to 100.

The survival times were generated as follows. First, we generated the true survival times (denoted by y_j , $j \in \{1, \dots, 100\}$) of the patients. For patient $1 \sim 50$, the true survival times were generated as normal random numbers with a mean of 6 and a standard deviation of 1. For patient $51 \sim 100$, the true survival times were generated as normal random numbers with a mean of 10 and a standard deviation of 1. Second, we generated censoring times, which were denoted by c_j , $j \in \{1, \dots, 100\}$. For each patient, a censoring time was generated as a normal random number with a mean of 12 and a standard deviation of 2. Third, the observed survival times of patients were generated by $t_j = \min(y_j, c_j)$. If the censoring time turned out to be less than the true survival time, the observation was considered as being censored. After generating the training data sets, we generated the test data sets by the same manner independently.

According to the above survival time generation method, patients $1 \sim 50$ were defined as “high-risk”, while patients $51 \sim 100$ were defined as “low-risk”. We generated the simulation data sets 10 times independently, and applied the prediction on each simulation. For each tuning parameter, we recorded prediction performance and averaged them across 10 times independent simulation.

We compared TSPC using the time course data with SPC using individual time points. The performance of prediction was characterized by the log-rank test p-values. In the original SPC method, the tuning parameter is the Cox-score, which has different ranges for different time points. To make the prediction performance comparable, we used the number of features (genes) as the tuning parameter. As shown in (Figure 1), SPC on individual time points 1, 2, and 3 gives unsatisfied prediction results, irrespective of how many features are used. To investigate whether simply combining the three time points (without considering the time course structure as TSPC does) can generate better results, we generated a big matrix by pooling the simulated gene expression from the three time points. Again, applying SPC on the pooled data set does not increase the prediction performance (Figure 1). On the contrary, TSPC gives much better results over SPC on individual time points, and simply pooled data set across the three time points (Figure 1). It suggests that TSPC successfully captures the time course structures of genes reflecting to the outcome.

We again compared TCoxL1 with the ordinary L_1 - *penalty* Cox regression model using individual time points. Similarly, TCoxL1 has better performance over the L_1 - *penalty* Cox regression model on the three individual time points, as well as the simply pooled data set (Figure 2). The prediction performance of TCoxL1, however, is not as high as TSPC. In the subsequent analysis, we focused on the application of TSPC on real data.

3.2 Survival analysis of burn patients

We applied our method to burn disease. Patients can have dramatically different outcomes after suffering a burn injury. We aim to uncover the biological reasons and improve the current medical prognosis by building genetic predictors. Patients

Figure 2: Methods comparison using simulated data. Black: TCoxL1, using Cox model with L_1 penalty as the second step; red: CoxL1 1, using the first time point; green: CoxL1 2, Cox model with L_1 using the second time point; blue: CoxL1 3, Cox model with L_1 using the third time point; cyan: CoxL1 C, applying Cox model with L_1 to the data set combining all the three time points. Y-axis is the pvalue of Log rank test. The five short bars in black, red, green, blue and cyan colors mean the average of standard error of p-values from each method respectively.

are monitored for changes in their clinical status, such as organ failure. Blood samples are drawn over the first days, months and years after the injury to obtain gene expression over time. We focus on the survival prediction of the burn patients using the number of days between injury and death or the day we stop monitoring

as the outcome. We begin by introducing preprocessing gene expression data, then describe the application of our method on this burn data as well as the comparison with straightforward prediction methods using individual time points.

In the Glue Grant project, a large scale collaborative effort, the blood samples of 123 burn patients were collected and the gene expression levels were measured by the Affymetrix HU133 Plus 2.0 arrays. The data was downloaded from the Glue Grant website (<http://www.gluegrant.org/>). Sample preparation was described in Zhou et al. (2010). Each array consisted of about 50,000 probe sets. Patients were monitored according to time. In our study, we used the data from the early stage (0 day to 10 days with three days median time) and the middle stage (11 days to 49 days with 19 days median time). Gene expression data was calculated and normalized by dChip (Li and Wong (2001)) and further reduced to 7354 probe sets with the coefficient of variation (standard deviation/mean) larger than 0.8. We used gene expression from early and middle stages to build the predictors. For patients with several measurements during early or middle stages, we took the median gene expression. The outcome of our prediction study is “the death day since the injury” of patients. The maximum death day is 195. If patients survived, we treat them as censored at the 196th day.

We first applied TSPC on the survival analysis of burn patients using the early and middle stage gene expression data set. The 123 patients were randomly divided into a training set with a size of 62 and a test set with a size of 61. The training set was used to learn the weights ($\hat{\beta}_1$ and $\hat{\beta}_2$) for the early stage and the middle stage gene expression, respectively. Predictors were constructed as the weighted gene expression data. Two-fold cross-validation was employed to select the number of predictors from the training set, based on the first principal component. This resulted in 259 genes as the predictors. Then, the first principal component was used for the prediction on the test data. Based on the scores of predictors, patients in the test data set were divided into two groups – the high-risk group and the low-risk group, with the cutoff of the median of scores obtained from training data. We applied the log-rank test on the two group patients. The log-rank test on the test set is very significant with a p-value 4.3×10^{-5} (Figure 3). It suggests that with TSPC the early and middle stage gene expression levels of burn patients are predictive for the survival time.

We next investigated whether the predictions of TSPC were implicated in other important clinical outcomes. We studied the distributions of the clinical outcomes on the TSPC predictions of the high-risk and low-risk patients. In figure 4 we plotted the distribution of continuous clinical outcomes on the high-risk and low-risk patients. The two clinical outcomes are the maximum Denver multiple organ failure (MOF) score (MAX_DENVER_2_SCORE) which grades four organ dysfunctions (lung, kidney, liver, and heart), acute care length of stay and discharged to

Figure 3: Log-rank test on the test data set of burn patients by TSPC.

inpatient rehabilitation (ACUTE_CARE_LENGTH_OF_STAY). One-sided Wilcoxon-test indicates that high-risk patients have higher MOF ($p\text{-value} = 1.5 \times 10^{-3}$). Figure 5 illustrates the distribution of categorical outcomes on predicted high-risk and low-risk patients. It shows that high-risk patients have higher death rate, higher noninfectious complication rate, lower acute care and lower burn wound infection. The results indicate that the TSPC predictions are suggestive of related clinical outcomes.

We also studied the biological meanings of the 259 selected predictor genes. We used the Ingenuity pathway analysis tool (<http://www.ingenuity.com/>) to perform functional analysis on the genes. The enriched canonical pathways were shown in Figure 6. These pathways are involved in the functions of signal transduction, inflammatory response, immune response functions, etc., consistent with the nature of the burn disease. We showed the heatmap of the gene expression of the

Figure 4: Box plots of continues outcomes of burn patients from predicted high-risk and low-risk patients.

“response to wounding” and “immune response” pathways in Figure 7. It demonstrates that although either of the early and middle stages of gene expression are slightly distinguishable between the high-risk and low-risk patients, the weighted gene expression levels are very distinct between the two groups.

The weights of the early and middle stage expression of the 259 genes were shown in Figure 8. The amplitude of $(\hat{\beta}_1, \hat{\beta}_2)$ reflects the contribution of gene expression from each time point. The sign of $\hat{\beta}_1 \times \hat{\beta}_2$ reflects the relationship between two time points. If the sign of $\hat{\beta}_1 \times \hat{\beta}_2$ is positive, it means that the two time points have additive effects on the outcome. If the sign of $\hat{\beta}_1 \times \hat{\beta}_2$ is negative, it suggests that the outcome is related to the trend of that gene along time. The amplitude of the weights acts as a significant factor to the time course effect because the odd ratio is 2.24 with 95% confidence interval (1.16, 4.33). The percentage of contrary-sign weights is 0.47 within the “inflammatory response” pathways; while the percentage of $|\hat{\beta}_2| > |\hat{\beta}_1|$ is 0.84. This is consistent with the fact shown in Figure 7 that gene expression of the patients with different outcomes in the middle stage is higher than the early stage.

To investigate the significance of gene expression on burn patients’ survival analysis, we compared the prediction performance of longitudinal gene expression

data with those of clinical variables. Ryan, Schoenfeld, and et al. (1998) studied the probability of death of burn patients based on clinical variables such as age, sex, and the injury percentage of the body-surface area, etc. We performed survival analysis

Table 1: Comparing with clinical variables.

Variable	p-value	Hazard ratio	lower .95	upper .95
Gene Predictor	< 0.0001	2.72	1.64	4.51
AGE	0.09	1.03	1	1.06
SEX	0.15	0.43	0.14	1.36
WEIGHT	0.5	1.01	0.99	1.03
ER_BASE_DEFICIT	0.73	0.98	0.88	1.09
ER_LACTATE	0.19	1.16	0.93	1.45
NOSO	0.25	0.3	0.04	2.35
APACHESC	0.85	1.01	0.94	1.08
DENV2_24	0.26	1.3	0.82	2.06
BURN_PERCENT_TOT	0.002	1.05	1.02	1.08
INJ_INHAL	0.44	1.57	0.51	4.86
BWI	0.99	1.01	0.27	3.74

using the univariate Cox model for the clinical variables. The p-values based on the Wald test and the Hazard ratio were calculated for each clinical variable, as well as the longitudinal gene expression levels. As shown in Table 1, the gene predictors are much better than the tested clinical variables.

Finally, to check whether predictors using the longitudinal gene expression are better than using the individual time points, we compared the prediction perfor-

Figure 7: Heatmap of gene expression in the “response to wounding” and “immune response” pathways. Left panel, gene expression from early stage and middle stage; right panel, the projected gene expression. For each heatmap, rows indicate genes, columns indicate patients which are ordered by increasing survival time. One way hierarchial clustering is applied on the gene expression.

mance of TSPC and SPC on the early or middle stages or the pooled data from the two stages. We repeated the prediction by randomly splitting the 123 samples into the training and test sets for 30 times. In each randomization, the model parameters were learned from the training set and tested on the test set for all methods. The averaged prediction performance on test data was shown in Figure 9. One can see that our method using the longitudinal gene expression has the best performance.

4 Discussion

We have proposed a new statistical prediction method in the survival analysis scenario using the longitudinal gene expression. The key idea is first projecting time

course gene expression to an “optimal direction”, then incorporating model selection and dimension reduction strategies to extract predictors. As the weights for the projection of test data are the same weights obtained from training data, the homogenousness of data is critical. If there are strong batch effects between training data and test data, the estimated “optimal direction” of test data can be very much different from true “optimal direction”. Batch effects should be removed in the first place. In the second stage, we used two particular methods – SPCA and Cox regression with L_1 penalty. However, procedures of the second stage are not limited to these two methods. Other methods like forward stagewise regression and boosting methods can be used to do variable selection. One can choose an appro-

Figure 9: Methods comparison using burn data. Black: TSPC; red: SPC E, SPC using the early time point; green: SPC M, SPC using the middle time point; blue: SPC C, applying SPC to the data set combining the early and middle time points. Y-axis is the $-\log(\text{pvalue of Log rank test})$

appropriate variable selection method depending on the situation and preference. Besides survival outcomes, it will be interesting to predict other types of outcomes such as binary, category and continuous outcomes using longitudinal gene expression in future.

Our studies on the burn data and simulated data show that appropriate use of longitudinal structure of gene expression can improve the power of prediction. For the clinical study perspective, it is straightforward to monitor patient status and

measure the time course gene expression. The genome-wide dynamic regulation of gene expression is attracting more and more interest. As more time course data are generated, our prediction approach will have wider applications.

References

- Aastveit, A. H. and H. Martens (1986): "Anova interactions interpreted by partial least-squares regression," *Biometrics*, 829–844, 42(4):.
- Akdeniz, F. and G. Tabakan (2009): "Restricted ridge estimators of the parameters in semiparametric regression model," *Communications in Statistics-Theory and Methods*, 1852–1869, 38(11):.
- Bair, E., T. Hastie, and et al. (2006): "Prediction by supervised principal components," *Journal of the American Statistical Association*, 119–137, 101(473):.
- Bair, E. and R. Tibshirani (2004): "Semi-supervised methods to predict patient survival from gene expression data," *PLoS Biol*, 2(4): E108.
- Chun, H. and S. Keles (2010): "Sparse partial least squares regression for simultaneous dimension reduction and variable selection," *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 3–25, 72:.
- Cox, D. R. (1972): "Regression models and life tables," *Journal of the Royal Statistical Society Series B*, 187–220, 34(2):.
- Cramer, R. D., J. D. Bunce, and et al. (1988): "Cross-validation, bootstrapping, and partial least-squares compared with multiple-regression in conventional qsar studies," *Quantitative Structure-Activity Relationships*, 18–25, 7(1):.
- Desmedt, C., B. Haibe-Kains, and et al. (2009): "Gene expression signatures can predict the efficacy of anthracyclines in her2-negative and her2-positive breast cancer (bc) patients: The results of the top trial and their validation in the big1-00 trial, on behalf of the top trial investigators," *Annals of Oncology*, 45–45, 20:.
- Faigle, J. F., R. J. Poppi, and et al. (1991): "Multicomponent principal component regression and partial least-squares analyses of overlapped chromatographic peaks," *Journal of Chromatography*, 123–132, 539(1):.
- Farkas, O. and K. R. Heberger (2005): "Comparison of ridge regression, partial least-squares, pairwise correlation, forward- and best subset selection methods for prediction of retention indices for aliphatic alcohols," *Journal of Chemical Information and Modeling*, 339–346, 45(2):.
- Forrester, J. B. and J. H. Kalivas (2004): "Ridge regression optimization using a harmonious approach," *Journal of Chemometrics*, 7–8, 18(): 372-384.
- Friedman, J., T. Hastie, and et al. (2010): "Regularization paths for generalized linear models via coordinate descent," *journal of statistical software* 33(1), .

- Gana, R. (1995): "Ridge-regression estimation of the linear probability model *journal of applied statistics* 22(4): 537-539," .
- Helland, I. S. (1990): "Partial least-squares regression and statistical-models," *Scandinavian Journal of Statistics*, 97–114, 17(2):.
- Hertz, D. (1991): "Sequential ridge-regression," *Ieee Transactions on Aerospace and Electronic Systems*, 571–574, 27(3):.
- Hoerl, A. E. and R. W. Kennard (1990): "Ridge-regression - degrees of freedom in the analysis of variance," *Communications in Statistics-Simulation and Computation*, 1485–1495, 19(4):.
- Holcomb, T. R., H. Hjalmarsson, and et al. (1997): "Significance regression: A statistical approach to partial least squares," *Journal of Chemometrics*, 283–309, 11(4):.
- Hu, H. C. (2005): "Ridge estimation of a semiparametric regression model," *Journal of Computational and Applied Mathematics*, 215–222, 176(1):.
- Jacobsen, T., K. Kolset, and et al. (1986): "Partial least-squares regression and fuzzy clustering - a joint approach," *Mikrochimica Acta*, 1–6, 2: 125-138.
- Jain, R. K. (1985): "Ridge regression and its application to medical data," *Comput Biomed Res*, 363–368, 18(4):.
- Karlsson, E., U. Delle, and et al. (2008): "Gene expression variation to predict 10-year survival in lymph-node-negative breast cancer," *Bmc Cancer*, 8.
- Korde, L. A., L. Lusa, and et al. (2010): "Gene expression pathway analysis to predict response to neoadjuvant docetaxel and capecitabine for breast cancer," *Breast Cancer Research and Treatment*, 685–699, 119(3):.
- Lecessie, S. and J. C. Vanhouwelingen (1992): "Ridge estimators in logistic-regression," *Applied Statistics-Journal of the Royal Statistical Society Series C*, 191–201, 41(1):.
- Li, C. and W. H. Wong (2001): "Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection, *proc natl acad sci u s a* 98(1): 31-36," .
- Lin, Y. H., J. Friederichs, and et al. (2007): "Multiple gene expression classifiers from different array platforms predict poor prognosis of colorectal cancer," *Clinical Cancer Research*, 498–507, 13(2):.
- Ma, P., W. Zhong, and J. S. Liu (2009): "Identifying differentially expressed genes in time course microarray data," *Statistics in Biosciences*, 144–159, 1(2):.
- Martens, I. L., H. and et al. (1986): "Partial least-squares regression on design variables as an alternative to analysis of variance," *Analytica Chimica Acta*, 133–148, 191:.
- Maruyama, Y. and W. E. Strawderman (2005): "A new class of generalized bayes minimax ridge regression estimators," *Annals of Statistics*, 1753–1770, 33(4):.

- Naderi, A., A. E. Teschendorff, and et al. (2007): "A gene-expression signature to predict survival in breast cancer across independent data sets," *Oncogene*, 1507–1516, 26(10):.
- Nguyen, D. V. and D. M. Rocke (2002): "Partial least squares proportional hazard regression for application to dna microarray survival data," *Bioinformatics*, 1625–1632, 18(12):.
- Nimeus-Malmstrom, E., C. Ritz, and et al. (2006): "Gene expression profilers and conventional clinical markers to predict distant recurrences for premenopausal breast cancer patients after adjuvant chemotherapy," *European Journal of Cancer*, 2729–2737, 42(16):.
- Ozkale, M. R. (2008): "A jackknifed ridge estimator in the linear regression model with heteroscedastic or correlated errors," *Statistics and Probability Letters*, 3159–3169, 78(18):.
- Pandit, T. S., W. Kennette, and et al. (2009): "Lymphatic metastasis of breast cancer cells is associated with differential gene expression profiles that predict cancer stem cell-like properties and the ability to survive, establish and grow in a foreign environment," *International Journal of Oncology*, 297–308, 35(2):.
- Park, M. and T. Hastie (2007): "L1-regularization path algorithm for generalized linear models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 659–677.
- Pliskin, J. L. (1990): "The mean squared error of the generalized ridge-regression estimator and the orientation of beta," *Communications in Statistics-Simulation and Computation*, 1477–1484, 19(4):.
- Ryan, C. M., D. A. Schoenfeld, and et al. (1998): "Objective estimates of the probability of death from burn injuries, *n engl j med* 338(6): 362-366," .
- Schneeweiss, A., P. Lichter, and et al. (2006): "Gene expression profiling to predict chemotherapy response in primary breast cancer," *Breast Care*, 362–367, 1(6):.
- Schorge, J. O., D. R. D., and et al. (2009): "Gene expression profiling to predict early relapse in ovarian cancer," *Gynecologic Oncology*, 112(2): S105-S106.
- Storey, J. D., W. Xiao, and et al. (2005): "Significance analysis of time course microarray experiments," *Proc Natl Acad Sci U S A*, 2837–1284, 102(36): 12.
- Tai, S. T., YC (2006): "A multivariate empirical bayes statistic for replicated microarray time course data," *Annals of Statistics*, 2387–2412, 34:.
- Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B-Methodological*, 267–288, 58(1):.
- Vangaans, P. F. M. and S. P. Vriend (1990): "Multiple linear-regression with correlations among the predictor variables - theory and computer algorithm ridge (fortran-77)," *Computers and Geosciences*, 933–952, 16(7):.
- Wakeling, I. N. and J. J. Morris (1993): "A test of significance for partial least-squares regression *journal of chemometrics* 7(4): 291-304," .

- Yeh, C. H. and C. H. Spiegelman (1994): “Partial least-squares and classification and regression trees *chemometrics and intelligent laboratory systems* 22(1): 17-23,” .
- Young, P. J. (1994): “A reformulation of the partial least-squares regression algorithm,” *Siam Journal on Scientific Computing*, 225–230, 15(1):.
- Yuan, K. C., M. (2006): “Hidden markov models for microarray time course data under multiple biological conditions (with discussion),” *Journal of the American Statistical Association*, 1323–1340, 101:.
- Yuan, Y. Y., C. T. Li, and et al. (2008): “Partial mixture model for tight clustering of gene expression time-course,” *Bmc Bioinformatics*, 9.
- Zhou, B., W. Xu, D. Herndon, R. Tompkins, R. Davis, W. Xiao, and W. Wong (2010): “Analysis of factorial time-course microarrays with application to a clinical study of burn injury,” *Proceedings of the National Academy of Sciences*, 107, 9923.