Sri Lankan Journal of Applied Statistics (Special Issue)
Modern Statistical Methodologies in the Cutting Edge of Science

Predicting Quantitative Outcomes of Patients Using Longitudinal Gene Expression

Yuping Zhang[1*,2,3] **and Zhengqing Ouyang**[4,5,6]

[1]Department of Statistics, University of Connecticut,USA
[2]Institute for Systems Genomics, University of Connecticut, USA
[3]Center for Health, Intervention, and Prevention,
University of Connecticut, USA
[4]The Jackson Laboratory for Genomic Medicine, Farmington, USA
[5]Department of Biomedical Engineering, University of Connecticut, USA
[6]Department of Genetics and Developmental Biology, University of
Connecticut Health Center, USA
* Corresponding Author: yuping.zhang@uconn.edu

Received: 15, August 2013 / Revised: 2, February 2014 / Accepted: 22, March 2014

ABSTRACT

We present a new statistical method using longitudinal gene expression to predict quantitative outcomes of patients. Through simulations and application on a dataset of burn patients, the proposed method outperforms prediction methods using gene expression from individual time points or simply combining them together.

Keywords: Prediction, longitudinal gene expression, quantitative outcomes.

1. Introduction

High-throughput genomic technologies are dramatically advancing research in biology and translational medicine. Many existing literature have focused on prediction of patient outcomes using stationary gene expression. However, gene expression changes dynamically for various biological functions. Measuring gene expression across time can play important roles for monitoring the status of patients. Thus, it is essential to develop statistical methods for analyzing longitudinal gene expression data from patients. Many existing methods are for detecting differentially expressed genes across time [4, 6, 7, 11, 5, 3], or unsupervised learning of temporal structure of longitudinal gene expression [8]. Towards the goal of genomic medicine, one challenge is to predict patient

outcomes based on gene expression across time. Previous research has demonstrated that using temporal structure of gene expression can improve prediction accuracy of survival outcomes [9] and categorical outcomes [10]. However, to our knowledge, no method has been developed for prediction of quantitative outcomes of patients using longitudinal gene expression. It is common that clinical outcomes of interest are quantitative measurements or scores, e.g., scores of multiple organ failure (MOF). In this paper, we will present a new prediction method for quantitative outcomes of patients using longitudinal gene expression.

2. Method

Let $\mathbf{X}_g$ denote a $N \times t$ matrix, which is a gene expression matrix with columns centered. We assume there are p features (e.g. genes) measured on N observations across t time points, and each column is centered. Let $\mathbf{y}$ be an N-vector of quantitative outcome measurements or scores. To make use of temporal information, we evaluate the association between longitudinal gene expression and outcomes. Specifically, we find the direction that maximizes the association between the transformation of longitudinal gene expression and outcomes. Since genes are not functionally equal on the recovery of patients, the unknown optimal direction should be gene specific. To find the optimal direction, we need to define an objective function that maximizes the correlation between the transformed longitudinal gene expression and outcomes. It is natural to use a linear regression model. For each gene g, we perform the linear regression

$$\mathbf{y} = \mathbf{X}_g \beta_g. \tag{2.1}$$

The coefficient vector of this linear regression model is the optimal direction for the projection of longitudinal expression of the gene. Let $\hat{\beta}$ denote the estimated coefficient vector. The amplitude of $\hat{\beta}$ reflects the relationship between the quantitative outcomes 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.

$$\hat{\mathbf{x}}_g = \mathbf{X}_g \hat{\beta}_g. \tag{2.2}$$

The estimated weighted gene expression $\hat{\mathbf{x}}_g$ is used as a new feature representing that gene. We denote the weighted gene expression matrix by $\hat{\mathbf{X}}_g$, which consists of p vectors $\hat{\mathbf{x}}$ with length of N.

The association between quantitative outcomes and the weighted gene expression varies across genes. We perform variable selection to select those genes with higher association to outcomes. We use the score statistic to filter the genes [1]. The score statistic for predictor g is defined as follows:

$$s_g = \frac{U_g(0)^2}{I_g(0)}, \tag{2.3}$$

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.

We select those genes with high score statistic and then perform principal component analysis on the weighted gene expression matrix of the selected genes. The whole algorithm is as below:

- Compute the coefficient vector in the regression model Eq. 2.1.
- Obtain the weighted predictors $\hat{\mathbf{X}}$(Eq. 2.2), with $\hat{\mathbf{X}}^{\mathbf{i}}$ denoting the usual vector of predictors $(\hat{\mathbf{x}}_{\mathbf{1}}, ..., \hat{\mathbf{x}}_{\mathbf{p}})$ for the i^{th} individual.
- Center the weighted gene expression
- Compute the score statistic using Eq. 2.3 for each gene based on the weighted gene expression.
- Form a reduced data matrix consisting of only selected genes whose score statistic exceeds a threshold θ in absolute value (θ is estimated by cross-validation).
- Compute the first (or first few) principal component(s) of the reduced data matrix.
- Use these principal component(s) in a linear regression model to predict the quantitative outcome.

To choose appropriate tuning parameter θ, we use two-fold cross-validation. It is necessary to recalculate coefficient vectors in cross-validation, instead of using the one obtained from the whole training data. After an appropriate value of θ is chosen, we calculate weighted gene expression for test data $\hat{\mathbf{X}}^*$ using the coefficient vector $\hat{\beta}$ obtained in the whole training data. We then center each component of $\hat{\mathbf{X}}^*$ using the means derived from the training data $\hat{\mathbf{X}}$. Instead of

doing singular vector decomposition on $\hat{\mathbf{X}}^*$, we derive the predictor using the right singular vectors derived from training data.

3. RESULTS

3.1. Simulation

To validate our method, we performed a simulation study. We simulated longitudinal data sets with 1000 features and 100 subjects from two time points. Let X^k denote the data matrix at time point k consisting of 1000 "genes" (rows) and 100 "patients" (columns). Each value in this matrix is denoted by x^k_{gj}, $k \in 1, 2$, which represents the "expression level" of the gene g and patient j at time point k. We generated the data as follows:

$$x^0_{gj} = \begin{cases} 6 + \varepsilon_b & \text{if } g \leq 500, j \leq 50 \\ 6.1 + \varepsilon_b & \text{if } g \leq 500, j > 50 \\ 3.5 + \varepsilon_b & \text{else ,} \end{cases}$$

and

$$x^1_{gj} = \begin{cases} 0.5 + \varepsilon_t + x^0_{gj} & \text{if } g \leq 500, j \leq 50 \\ -0.3 + \varepsilon_t + x^0_{gj} & \text{if } g \leq 500, j > 50 \\ \varepsilon_t + x^0_{gj} & \text{else ,} \end{cases}$$

where, $\varepsilon_b \in N(0, 2.5)$ and $\varepsilon_t \in N(0, 2)$. We introduced the time course structure in genes indexed from 1 to 500.

The outcomes (denoted by y_j, $j \in \{1, \ldots, 100\}$) were generated as the following:

$$y_j = \begin{cases} 6 + \varepsilon & j \leq 50 \\ 10 + \varepsilon & j > 50, \end{cases}$$

where $\varepsilon \sim N(0, 1)$. After generating the training data sets, we generated the test data sets as the same manner independently.

The training data was used to train the parameters in three methods: 1) our method using longitudinal gene expression, 2) the SPC method [1] using gene expression from individual time points, and 3) using gene expression matrix with time points simply combined. 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 the 10 independent simulations. The performance of prediction was characterized by the Pearson correlation coefficient between the real values from test data and predicted values. To make the prediction performance comparable, we used the number of features as the tuning parameter. The results are shown in Figure 3.1. One can see that our method has the best performance. It suggests that considering the temporal structure of longitudinal data provides better prediction accuracy for quantitative outcomes of patients than those not considering such structure.

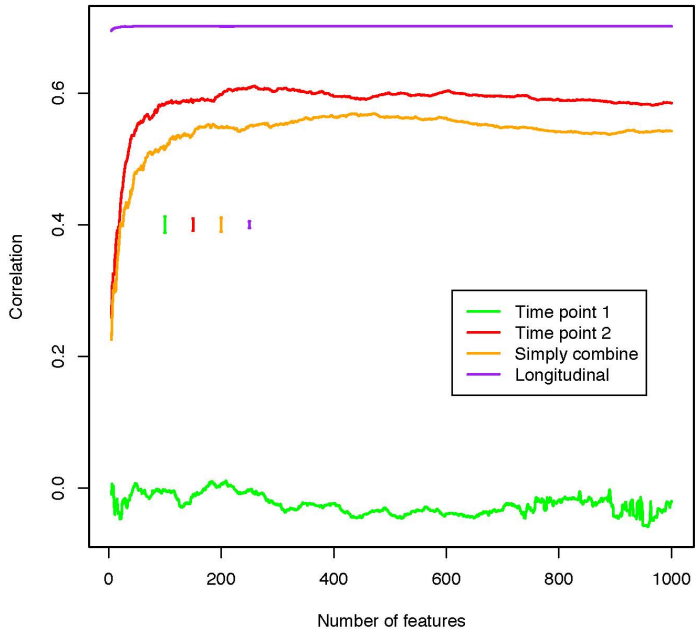

Figure 3.1. Methods comparison using simulated data. Green: SPC using the first time point; red: SPC using the second time point; orange: applying SPC to the data set simply combing all the three time points; purple: our approach using longitudinal gene expression. Y-axis is the Pearson correlation coefficient between the real values of the test data and the predicted values. The four short bars in green, red, orange and purple colors mean the average of standard error of correlation coefficient from each method respectively.

3.2. Application on a data set of burn patients

We applied our method to a cohort of burn patients (www.gluegrant.org). The patients were monitored for observing MOFs and gene expression over time. The time points can be divided into three stages - the early stage (0 day to 10 days with three days median time), the middle stage (11 days to 49 days with 19 days median time), and the late stage (larger than 49 days). Gene expression values were measured over time from their blood samples using the Affymetrix HU133 Plus 2.0 arrays. In our study, we used gene expression data from the early stage and the middle stage. The outcomes of interest are the maximum organ failure scores of patients during the late stage. Gene expression values were calculated and normalized by dChip ([2]) and further reduced to 7354 probe sets with 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 total number of patients in our study is 121.

We performed the following study to show the performance of our method and make comparison with prediction approaches using individual time points or all time points simply combined. We first divided the patients randomly with 63 patients for training and 58 patients for test. We used the training data to learn the parameters in three methods: 1) our method using longitudinal gene expression, 2) the SPC method of [1] using gene expression from individual time points, and 3) using gene expression matrix with time points simply combined. Then we make predictions on the test data sets under different tuning parameters, which are the numbers of features used in the prediction models. We performed the study 10 times by randomly splitting the whole data into training and test data sets, and showed results in Figure 3.2. We used Pearson correlation coefficient between the real values of the test data and the predicted values to demonstrate the performance of the methods. One can see that our approach outperforms the others using individual time points or simply combining them together.

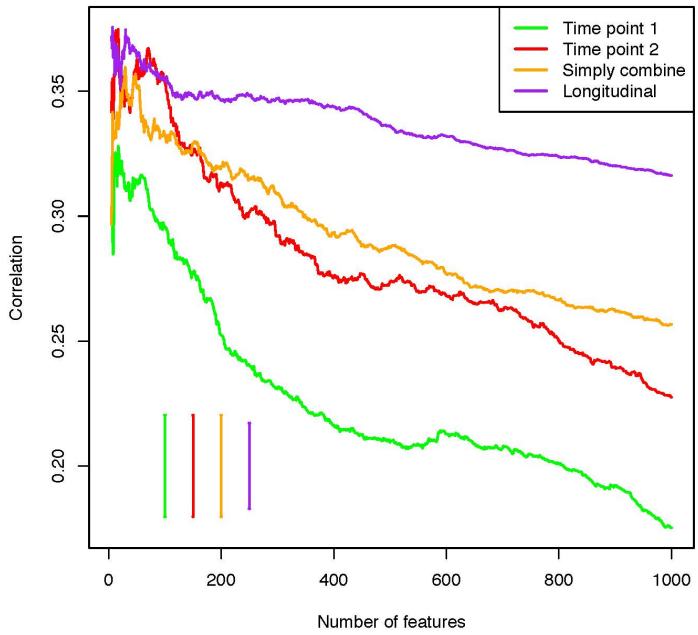

Figure 3.2. Method comparison on the data set of burn patients. Green: SPC using the first time point; red: SPC using the second time point; orange: applying SPC to the data set simply combing both time points; purple: our approach using longitudinal gene expression. Y-axis is the Pearson correlation coefficient between the real values of the test data and the predicted values. The four short bars in green, red, orange and purple colors mean the average of standard error of correlation coefficient from each method respectively.

4. Conclusion

We have proposed a new statistical method for predicting quantitative outcomes using longitudinal gene expression. The key idea is that we first perform supervised transformation of the longitudinal data set to capture temporal structures associated with outcomes. Then, we perform model selection and dimension reduction to extract the predictors. Our studies on the burn data and simulated data show that capturing of longitudinal structure of gene expression can improve the power of predicting quantitative outcomes of patients.

The weights used for transformation of test data are the same as those obtained from training data. Thus, it is critical that input data should be homogeneous. In our approach, we used supervised principal component analysis to select predictors. However, it is not limited to this approach. We can use other methods to do variable selection, e.g., lasso, forward stagewise regression, and boosting. One can choose an appropriate variable selection method depending on real situations and the preference.

Acknowledgements

We wish to acknowledge the efforts of many individuals at participating institutions of the Glue Grant Program that generated the clinical and genomic data used here.

References

1. Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, **101**(473).

2. Li, C. and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences*, **98**(1), 31–36.

3. Ma, P., Zhong, W., and Liu, J. S. (2009). Identifying differentially expressed genes in time course microarray data. *Statistics in Biosciences*, **1**(2), 144–159.

4. Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G., and Davis, R. W. (2005). Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(36), 12837–12842.

5. Tai, Y. C., Speed, T. P., *et al.* (2006). A multivariate empirical bayes statistic for replicated microarray time course data. *The Annals of Statistics*, **34**(5), 2387–2412.

6. Yuan, M. and Kendziorski, C. (2006). Hidden markov models for microarray time course data in multiple biological conditions. *Journal of the American Statistical Association*, **101**(476), 1323–1332.

7. Yuan, Y. Y., Li, C. T., and et al. (2008). Partial mixture model for tight clustering of gene expression time-course. *Bmc Bioinformatics*. 9.

8. Zhang, Y. and Davis, R. (2013). Principal trend analysis for time-course data with applications in genomic medicine. *The Annals of Applied Statistics*, **7**(4), 2205–2228.

9. Zhang, Y., Tibshirani, R., and Davis, R. (2010). Predicting patient survival from longitudinal gene expression. *Statistical applications in genetics and molecular biology*, **9**(1), Article41.

10. Zhang, Y., Tibshirani, R., and Davis, R. (2013). Classification of patients from time-course gene expression. *Biostatistics*, **14**(1), 87–98.

11. Zhou, B., Xu, W., Herndon, D., Tompkins, R., Davis, R., Xiao, W., and Wong, W. (2010). Analysis of factorial time-course microarrays with application to a clinical study of burn injury. *Proceedings of the National Academy of Sciences*, **107**(22), 9923.