Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Mathematical Biosciences 193 (2005) 119–137

Mathematical Biosciences

www.elsevier.com/locate/mbs

Partial least squares dimension reduction for microarray gene expression data with a censored response

Danh V. Nguyen *

Division of Biostatistics, Public Health Sciences, School of Medicine, University of California, One Shields Avenue, Davis, CA 956168638, USA

Received 24 November 2003; received in revised form 8 September 2004; accepted 21 October 2004

Abstract

An important application of DNA microarray technologies involves monitoring the global state of transcriptional program in tumor cells. One goal in cancer microarray studies is to compare the clinical outcome, such as relapse-free or overall survival, for subgroups of patients defined by global gene expression patterns. A method of comparing patient survival, as a function of gene expression, was recently proposed in [Bioinformatics 18 (2002) 1625] by Nguyen and Rocke. Due to the (a) high-dimensionality of microarray gene expression data and (b) censored survival times, a two-stage procedure was proposed to relate survival times to gene expression profiles. The first stage involves dimensionality reduction of the gene expression data by partial least squares (PLS) and the second stage involves prediction of survival probability using proportional hazard regression. In this paper, we provide a systematic assessment of the performance of this two-stage procedure. PLS dimension reduction involves complex non-linear functions of both the predictors and the response data, rendering exact analytical study intractable. Thus, we assess the methodology under a simulation model for gene expression data with a censored response variable. In particular, we compare the performance of PLS dimension reduction relative to dimension reduction via principal components analysis (PCA) and to a modified PLS (MPLS) approach. PLS performed substantially better relative to dimension reduction via PCA when the total predictor variance explained is low to moderate (e.g. 40%–60%). It performed similar to MPLS and slightly better in some cases.

* Tel.: +1 530 754 6510; fax: +1 530 752 3239.
E-mail address: ucdnguyen@ucdavis.edu

doi:10.1016/j.mbs.2004.10.007

Additionally, we examine the effect of censoring on dimension reduction stage. The performance of all methods deteriorates for a high censoring rate, although PLS–PH performed relatively best overall.

Keywords: DNA Microarray; Dimension reduction; Gene expression; Partial least squares; Principal components; Proportional hazard regression

1. Introduction and motivating applications

DNA microarray technologies have found broad applications, especially in biomedical research. For an overview of the technological and biological aspects of DNA microarrays, including applications, see [1] by Nguyen et al. and references therein. The application of DNA microarray technologies in cancer research is one specific area of interest. In this paper we examine a method for analyzing censored patient survival times (the censored response variable) with their corresponding gene expression profiles as covariates (predictors). To be more concrete, consider the following two motivating applications:

1. *Example: Diffused large B-Cell lymphoma*. In a study of diffused large B-cell lymphoma (DLBCL), mRNA expression for over 5622 gene probes were measured from microarray experiments [2]. In addition to the gene expression data collected, patient survival times were ascertained for $N = 40$ DLBCL patients. However, not all survival times could be observed by the end of the study. Of the 40 observed survival times, 22 were times of death; thus the percentage of censored observations is 55%.
2. *Example: Breast carcinomas*. Similarly, in a prospective breast carcinomas study, thousands of mRNA gene expression measurements were obtained simultaneously from microarray experiments [3]. Tissue samples for the microarray experiments were obtained from patients in a prospective study on locally advanced breast cancer with no distant metastases. Survival data was available for $N = 49$ patients with approximately 61% censoring.

Similar cancer microarray studies with survival data can be found in [4] for central nervous system embryonal tumors, [5] for DLBCL, [6] and [7] for prostate cancer, and [8] for lung carcinomas among others.

The data structure in the examples presented above can be more formally described as follows. Suppose that Y_i is the true survival time of the ith patient. The variate of interest at the end of a study, the survival time, cannot be observed completely. Instead, we are only able to observe $T_i = \min(Y_i, Z_i)$, where Z_i is a censored value. Also, recorded for the ith patient are p gene expression values, denoted by $\mathbf{x}'_i = (x_{i1}, \ldots, x_{ip})$ $(i = 1, \ldots, N)$, obtained from microarray experiments. This vector of gene expression values is called the (patient-specific) gene expression profile or pattern. We refer to the expression profile generally as covariates or predictors. Thus, the typical data set, illustrated by the examples above, consists of N samples. In addition, each sample contains the triple $\{T_i, \delta_i, \mathbf{x}_i\}$, where $\mathbf{x}_i$ is the gene expression pattern, T_i is the survival time if $\delta_i = 1$, and it is the right-censored time if $\delta_i = 0$. Note that the number of observed survival times that