Supervised wavelet method to predict patient survival from gene expression data

Maryam Farhadian (MSc)[a]

[a] Department of Epidemiology & Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan. Iran
Email: m.farhadian@umsha.ac.ir

PauloJ.G. Lisboa (PhD)[b]

[b] School of Computing and Mathematical Sciences, Liverpool John Moores University, UK
Email: P.J.Lisboa@ljmu.ac.uk

Abbas Moghimbeigi (PhD)[c]

[c] Modeling of Noncommunicable Disease Research Center, Department of Biostatistics and Epidemiology, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran
Email: moghimbeigi@umsha.ac.ir

Jalal Poorolajal (MD, PhD)[d]

[d] Modeling of Noncommunicable Diseases Research Center, Department of Epidemiology & Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran
E-mail: poorolajal@umsha.ac.ir

Hossein Mahjub (PhD)[f]

[f] Research Center for Health Sciences and Department of Biostatistics and Epidemiology, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran
E-mail: mahjub@umsha.ac.ir

Corresponding author: Prof. Hossein Mahjub (PhD)
Research Center for Health Sciences and Department of Biostatistics and Epidemiology, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran
Tel: +98(81)38380090
Fax: +98(81)38380509
E-mail: mahjub@umsha.ac.ir

Abstract

In microarray studies, the number of samples is relatively small compared to the number of genes per sample. An important aspect of microarray studies is the prediction of patient survival based on their gene expression profile. This naturally calls for the use of a dimension reduction procedure together with the survival prediction model. In this study, a new method based on combining wavelet approximation coefficients and Cox regression was presented. The proposed method was compared with supervised principal component and supervised partial least squares methods. The different fitted Cox models based on supervised wavelet approximation coefficients, the top number of supervised principal components and partial least squares components were applied to the data. The results showed that the prediction performance of the Cox model based on supervised wavelet feature extraction was superior over the supervised principal components and partial least squares components. The results suggested that the possibility of developing new tools based on wavelets for the dimensionally reduction of microarray data sets in the context of survival analysis.

1. Introduction

Microarray studies are widely used in biological and medical studies because they allow researchers to monitor tens of thousands of gene expression profiles simultaneously. Much of the interest in microarray data analysis derives from the potential of identifying the genes that relate to biological processes, the classification of tumor types, the stages based on gene expression patterns, and the study of gene interactions [1,2]. However, because microarray data sometimes include patients survival data, it is important to study patients survival times (response) in terms of their corresponding gene expression levels (predictors). The discovery of the relationship between time to event (survival time) and gene expression profiles as covariates provide the possibility to obtain more accurate diagnosis and advanced treatment [3]. It is estimated that high-dimensional gene expression data could noticeably enhance the predictive ability of such survival models [4].

Survival analysis is a statistical method that especially deals with the modeling and analysis time from a well-defined time origin until the occurrence of some event or end point of interest. A major complexity of analyzing such data is right censoring, where the event of interest is known to occur only after a certain time point. One popular regression model that takes into account the censored response is the Cox Proportional Hazards (CPH) regression model [5]. A substantial challenge in this setting comes from the fact that the number of genomic variables p is usually much larger than the number of subjects n (i.e., $p>>n$). Existing statistical methods such as CPH model require fewer predictors than cases [4]. Thus, a crucial step towards the application of microarrays in survival prediction is the dimensionality reduction from the gene expression profiles. In recent years, both feature selection and feature extraction methods have been widely used to predict the survival of cancer patients based on gene expression data [6].

Rosenwald et al. described a feature selection approach for identifying genes related to survival time that fits CPH models to each gene and selected those that pass a threshold for significance [7]. Liu et al. presented the adaptive $L_{1/2}$ shooting regularization method, which is used for variable selection in the CPH model [8]. Alizadeh et al. described an approach in which he first clustered the genes and then fitted a CPH model using the average expression level of each cluster as a covariate [1]. Nguyen and Rocke and Park et al. considered the problem of relating survival time to gene expression by reducing the dimensionality via partial least squares method. The first

a few linear combinations of gene expressions obtained via PLS were subsequently used in a CPH regression model for predicting the survival probabilities [9,10]. Li and Luan developed a penalized estimation procedure for the CPH model using kernels, under the assumption that the covariate effects were smooth functions of gene expression levels [11].

Several studies have compared dimension reduction methods in survival prediction based on microarray data. Bøvelstad et al. applied seven dimension reduction methods in order to predict survival in patients with diffuse large B-cell lymphoma (DLBCL) using gene expression dataset. Totally, their results showed that the ridge regression had best performance [4].

One of the methods used for feature extraction from the high dimensional data is wavelet transform. Normally, one dimensional discrete wavelet transform (DWT) is used to reduce dimensionality in the analysis of high dimensional biomedical data [12]. The primary intuition for applying wavelets in the case of gene expression is that genes are often co-expressed in groups. It would be useful to treat the group as a single variable, akin to the motivation behind methods such as principal component analysis [12]. Studies showed that this method has acceptable performance in the field of dimension reduction in the classification framework [14-17].

However, few studies have used wavelet transform in the area of survival analysis. For example, Liu used continuous wavelet transform combined with a genetic algorithm to select genes related to survival in colon cancer [16]. This study aimed to introduce a dimension reduction strategy for transforming the high-dimensional gene expression data into a low dimensional space based on wavelet transform. Accordingly, a predictive survival model was built upon the reduced dimensional space. Then, the proposed novel supervised method of feature extraction was compared with the supervised principal component analysis (PCA) and the supervised partial least squares (PLS) method.

2. Material and methods

2.1. Simulation Setup:

We performed simulation study to evaluate and compare the performance of the proposed supervised wavelet method with Supervised PCA and Supervised PLS. The simulated data set was first presented by Bair et al., for evaluation purposes [18]. Following Bair et al. simulated data set X consisted of 5000 genes and 100 samples. All expression values were generated as standard normal random numbers with a few exceptions. Genes 1–50 in samples 1–50 had a mean of 1.0. We randomly selected 40% of the samples to have a mean of 2.0 in genes 51–100, 50% of the samples to have a mean of 1.0 in genes 101–200, and 70% of the samples to have a mean of 0.5 in genes 201–300.

The survival times of samples 1–50 were generated as normal random numbers with a mean of 10.0 and a standard deviation of 2.0, and the survival times of samples 51–100 were generated as normal random numbers with a mean of 8.0 and a standard deviation of 3.0. For each sample, a censoring time was generated as a normal random number with a mean of 10.0 and a standard deviation of 3.0. If the censoring time turned out to be less than the survival time, the observation was considered to be censored [18].

2.2. Real-life datasets

We applied the supervised wavelet transform method to a set of gene expression data with survival information on two real datasets. The first dataset was related to the diffuse large B-cell lymphoma (DLBCL) dataset of Rosenwald et al. and the second dataset was related to the lung cancer dataset of Beer et al. [7,20].

The DLBCL dataset included expression measurements of 7,399 genes on 240 patients, together with their survival times. A total of 138 deaths were observed during the study with the median death time of 2.8 years. The dataset is available at http://llmpp.nih.gov/lymphoma/data.shtml.

The lung cancer dataset also included expression measurements of 7,129 genes on 86 lung adenocarcinoma patients, together with their survival times. The survival times were observed in 24 patients, and the censored times, in 62 patients. A detailed description of lung cancer dataset can be found in the original publication [20]. We used the dataset from the study conducted by Zhao et al. in 2008[21].

2.3. Cox proportional hazards model

The CPH model is the most commonly used model in survival analysis. It is also known as the Cox regression model. It factorizes the time dependence of the event rate from the covariate dependence, as follows:

$$h(t,x) = h_0(t)\exp(\beta^T x) \quad (1)$$

where $h(t,x)$ represents the hazard function at time t for a subject with covariates x. For different covariates, CPH regression models the hazard as a proportional factor applied to time-dependent baseline hazard that corresponds to a reference population for which the covariate values are all zero. This baseline hazard function is $h_0(t)$ and the effect of the covariates x is modeled linearly using $\beta^T x$, which is known as the risk score. The coefficient vector β is estimated by maximizing the partial likelihood:

$$l(\beta) = \prod_{j=1}^{k}\left(\frac{\exp(\beta^T x_j)}{\sum_{l\in R_j}\exp(\beta^T x_l)}\right) \quad (2)$$

R_j represents all patients at risk at the jth failure time and k the number of distinct failure times. The hazard ratio between different observations i and j by Eq (1) is assumed to be constant and independent of time:

$$\frac{h_i(t,x_i)}{h_j(t,x_j)} = \frac{exp(\beta^T x_i)}{exp(\beta^T x_j)} \quad (3)$$

Consequently, the Cox regression model is a proportional hazards model [5].

2.4. Wavelet transform

A wavelet is a "small wave", which has its energy concentrated in time. In signal processing, a transformation technique is used to transfer a data in another domain where hidden information can be extracted. Wavelets have a nice feature of local description and separation of signal characteristics, and give a tool for the analysis of transient or time-varying signal [12]. A wavelet is a set of orthonormal basis functions generated from dilation and translation of a single scaling function or father wavelet (φ), and a mother wavelet (ψ).

Wavelet transforms are classified into two different categories: the continuous wavelet transforms (CWT) and the discrete wavelet transforms (DWT). DWT is a linear operation that operates on a data vector, transforming it into

a wavelets coefficient. The idea underlying DWT is to express any function $f\ (t) \in L^2\ (R)$ in terms of φ (*t*) and ψ (*t*) as follows:

$$f(t) = \sum_k c_0(k)\,\varphi(t-k) + \sum_k \sum_{\mathrm{j}=1} d_{\mathrm{j}}(k) 2^{\frac{-\mathrm{j}}{2}} \psi\left(2^{-\mathrm{j}}t - k\right)$$

$$= \sum_k c_{\mathrm{j}_0}(k) 2^{\frac{-\mathrm{j}0}{2}} \varphi\left(2^{-\mathrm{j}0}t - k\right) + \sum_k \sum_{\mathrm{j}=\mathrm{j}_0} d_{\mathrm{j}}(k) 2^{\frac{-\mathrm{j}}{2}} \psi\left(2^{-\mathrm{j}}t - k\right) \quad (4)$$

where φ(t), ψ(t), c_0 and d_{j}represent the scaling function, mother wavelet function, scaling coefficients (approximation coefficients) at scale 0, and detail coefficients at scale *j*, respectively. The variable *k* is the translation coefficient for the localization of gene expression data. The scales denote the different (low to high) scale bands. The variable symbol j_0 is scale (level) number selected.

One-dimensional discrete wavelet transform decomposes a signal as a sum of wavelets at different time shifts and scales (frequencies) using DWT. For this purpose, the signal is passed through series of high pass and low pass filters in order to analyze low as well as high frequencies in the signal as follows:

$$c_{j+1} = \sum_m h(m-2k) c_j(m) \quad (5)$$

$$d_{j+1} = \sum_m h_1(m-2k) c_j(m) \quad (6)$$

where $h(m-2k)$ and $h_1(m-2k)$ are the low-pass filters and high-pass filters.

The whole process of obtaining the wavelet transform of $f(t)$ using the pyramid algorithm is shown in Fig. 1.

At each level, the high pass filter produces detail coefficients (wavelet coefficients) d_1, while the low pass filter associated with scaling function produces approximation coefficient (scaling coefficients) c_1. Then the approximation coefficients c_1 are split into two parts by using the same algorithm and are replaced by c_2 and d_2, and so on. This decomposition process is repeated until the required level is reached. The coefficient vectors are produced by down sampling and are only half the length of the signal or the coefficient vector at the previous level.

The main advantage of the wavelet transform is that each basis function is localized jointly in both the time and frequency domains. From a viewpoint of time-frequency, the approximation coefficients are corresponding to the larger-scale low-frequency components, and the detail coefficients are corresponding to the small-scale high-

frequency components. Generally, the former can be used to approximate the original signal, and the latter represents some local details of the original signal [12-15].

There are different families of wavelets symlet, coiflet, daubechies and biorthogonal wavelets. They vary in various basic properties of wavelets, like compactness. Among them, Haar wavelets belonging to Daubechies wavelet family are most commonly used wavelets in database literature because they are easy to comprehend and fast to compute.

2.4.1. Supervised wavelet transform

The proposed method starts by adopting a univariate Cox model for each gene:

$h(t, x_g) = h_0(t) \exp(\beta^T x_g)$, for each gene $g = 1,2, \dots, 7399$,

The covariates, each representing a different gene, are then sorted by increasing absolute values of the Wald's statistic $\frac{\beta}{se(\beta)}$, which are measures of the correlation between the gene expression level and patient survival. Then, in each step we pick out the top number of genes included with higher Wald's statistic. Then, this reduced set of genes is modeled by the one-dimensional discrete wavelet transform to extract the relevant information and finally, the wavelet approximation coefficients in the first levels of decomposition are used in a multiple Cox regression model (Eq (1)). Note that, numbers of selected genes in this stage are considered proportional to the sample size. The Haar wavelet transform in the first level is applied on the preselected genes.

2.5. Supervised principal components analysis

Bair and Tibshirani and Bair et al. proposed the supervised principal components regression [18, 19]. This procedure first picks out a subset of the gene expressions that is correlated with survival by using univariate selection, and then applies PCA to this subset. In our analysis, we pick out top number of genes with higher Wald's statistic. Then, we apply principal components analysis to this subset of genes and in each step, include the top number of principal components that will be comprised of at least 75% of the total variance into a multivariate Cox model.

2.6. Partial least squares method

Partial least squares (PLS) is a supervised dimension reduction technique that is usually employed to correlate a response variable to the explanatory variables. PLS components are linear combinations of the predictor variables, constructed to maximize an objective criterion based on the sample covariance between response and covariates. PLS finds components that are both dependent on the variance of the gene expressions and the covariance between the gene expressions and the survival, whereas the components in PCA only depend on the variance of the gene expressions [9]. Many methods have been suggested to perform PLS for Cox regression. We used the method which was provided by the plsRcox package. In this study, the number of PLS components was fixed like for the Supervised PCA method.

2.6. Model building and model evaluation criteria

In order to evaluate the proposed method, in all experiments (simulation and real-life), data set was randomly divided into training (2/3 of the data) and test (1/3 of the data) sets for 50 times. The methods (supervised wavelet, supervised PCA and supervised PLS) were applied to the training set and the test set was used to calculate the evaluation measures. These data sets included 66 samples from 100 samples for simulated data, 160 samples from 240 patients for DLBCL data and 60 samples from 86 patients for lung cancer data.

For predicting survival of patients based on gene expression, we applied the proposed dimension reduction method, supervised PCA and supervised PLS in stage 1 in each data set, and then used the data in the reduced subspace to apply in the multiple CPH model in stage 2. In fact, following the evaluation scheme proposed by Bøvelstad et al. in each experiment, the parameters were estimated ($\widehat{\beta_{train}}$) from the training data set for a given method. Then, in the test set for each patient, the obtained estimates were used to derive a prognostic index (PI) ($PI = \acute{x}\,\widehat{\beta_{train}}$). Then, this PI index was used in the Cox model for calculating the evaluation criteria. The above procedure was repeated for 50 times [3, 4]. It is noted that various numbers of preselected genes were tested in each situation. Next, the results of model evaluation criteria were computed for each dataset. These methods were compared in terms of the mean of the criteria values. MATLAB r2012a software and R statistical package were used for data analysis.

The predictive performance of a fitted Cox model based on supervised wavelet coefficients, supervised principal components, and supervised partial least squares components were evaluated using R^2 statistic, Concordance Probability Estimate(CPE), Likelihood ratio test statistic, Integrated Brier Score and C index.

Moreover, in order to evaluate the effect of adding clinical information to genomic data on the performance of model for a lung cancer data set, clinical information was added to genomic data. The clinical features for each patient were included age, sex, stage, tumor size and nodal status.

2.6.1. R^2 statistic

R^2statistic measures the proportion of variation in survival data that may be explained by the predictor. A predictor with good predictive performance can explain a high proportion of variation in the survival data. On the other hand, a poor predictor may explain only a little variation in the data. Accordingly, when comparing models, the model with the larger R^2 statistic is usually preferred [6]. Nagelkerke et al. suggested a general definition of the R^2 statistic that may be employed for Cox proportional hazard model as follows:

$$R^2 = 1 - \exp(-\frac{2}{n}(l(\hat{\beta}) - l(0))) \quad (7)$$

where $l(.)$ indicates the log-likelihood function [22]. In the present study, R^2 values are those which were provided by the coxph() R function.

2.6.2. Concordance Probability Estimate

The discriminatory power of a statistical model is assessed by concordance probability estimate (CPE). This estimator is merely a function of the regression parameters and the covariate distribution without using the observed event and censoring times. A value of one for CPE denotes the perfect discrimination [23].

2.6.3. C index

Concordance, or C-statistic, is a valuable measure of model discrimination in analyses involving survival time data. In general, consider selecting random pairs of patients and for each pair note, whether the model correctly predicts an order, e.g., a higher model score for the better result. Concordance is then the fraction of pairs for

which the model is correct. A completely random prediction would have a concordance of 0.5, a perfect rule a concordance of one [24].

2.6.4. Likelihood ratio test statistic

The likelihood ratio test is a global goodness-of-fit test statistic for a Cox regression model**.** The test statistic for the likelihood ratio test is given as follows:

$$LR = -2lnL_R - (-2lnL_F) \quad (8)$$

Where R denotes the reduced (PH) model obtained when all β's are 0, and F denotes the full model. Thus, the performance is good when LR is large [5].

2.6.5. Integrated Brier Score (IBS)

At a given time point t, the Brier score for a single subject is defined as the squared difference between observed survival status (e.g., 1 = alive at time t and 0 = dead at time t) and a model based prediction of surviving time t. The Brier score is given by:

$$BS(t) = \frac{1}{N}\sum_{i=1}^{n}(p_i(t) - o_i(t))^2 W \quad (9)$$

Where N is the sample size, $o_i(t)$is the observed survival at time t and $p_i(t)$ is the predicted probability at time t. The weight W is used to remove a large censoring bias. The Integrated Brier Score (IBS) is a summary of the prediction error over event time by integrating the formula (9). The smaller the Brier score, the better the survival prediction would be [25].

3. Results

The results of the predictive performance of the fitted Cox models based on approximation wavelet coefficients, the top number of principal components and partial least squares components for simulated, DLBCL and lung datasets are shown in Tables 1 to 3, respectively. In general, the results showed that the Cox model based on supervised wavelet feature extraction method was superior over the supervised principal components and partial least squares components in terms of different evaluation criteria for three data sets. Although, in simulated data set all methods have a similar performance in terms of the Integrated Brier Score.

The results showed the spread of mean values of five evaluation measures over the 50 data sets are fairly large. These variations caused by selecting the data at random into 50 data sets as well as the variations of the prediction methods performance for the given datasets. In order to determine how much of the variation was due to the prediction methods, we used the supervised wavelet method as a benchmark, and for each of the two other methods computed the difference between the evaluation criteria in each of the conditions.

Fig.2 to Fig.6 showed the boxplots of these differences in each evaluation criterion for the 50 data sets. The median values for R^2, C index, CPE and LR were positive, which showed supervised wavelet method performed better than other methods. In addition, the median values for the Integrated Brier Score criterion in the different conditions was negative. Totally, simulation results and real data analysis confirmed the suitable performance of the supervised wavelet method.

The results of the predictive performance of the fitted Cox models based on combination of clinical and genomic information for a lung data set are shown in Tables 4. The results showed that adding clinical information leads to an increase in the predictive ability of the model in three mentioned methods (supervised wavelet, supervised PCA, supervised PLS).

4. Discussion

This study employed the supervised dimension reduction method based on wavelet transform and modeled survival times in the presence of right censoring and taking into account the microarray data information. The proposed method was evaluated by simulations and applied to the Rosenwald et al's DLBCL dataset and Beer et al's Lung cancer dataset [7,19].

Considering the fact that most genes are irrelevant to patients' survival, we analyzed the reduced dataset given by selecting genes that were significantly related to survival time based on the Wald's statistic. If the wavelet transform is performed directly by using all of the genes in a data set, there is no guarantee that the resulting wavelet coefficients will be related to survival [21,22]. Thus, this study introduced a supervised form of wavelet transform that can be considered as supervised wavelet. After extracting supervised wavelet approximation coefficients using discrete Haar wavelet transform, the coefficients had higher predictive performance than the top

number of principal components and the top number of partial least squares components. Hence, our results suggested that the wavelet coefficients are an efficient way to characterize the features of high dimensional microarray data. It seems that, these results exhibit the possibility of developing more efficient tools using wavelets for the dimensionally reduction of microarray data sets in the context of survival analysis.

The main purpose of the feature extraction method using wavelet transform is that the approximation coefficients usually comprise the majority of the important information [15].In addition, this method can usually condense or de-noise a signal without appreciable degradation due to using a different view of data than those presented by conventional methods. In addition, the powerful capability of the DWT to compress the signal energy makes it a good candidate for feature extraction applications. The DWT compresses most of the energy from the input signal and concentrates it in a few high-magnitude coefficients in the transformed matrix. The DWT also reduces the size of the input signal to half of its original size. Keeping only a number of these high-magnitude coefficients (in addition to their locations) while discarding the rest of the coefficients in the transformed signal can produce a valid feature vector representation of the input signal [14].

The wavelet feature extraction method does not depend on the training dataset to obtain the basis of feature space compared to PCA and PLS methods. Therefore, the wavelet feature extraction method reduces the computation load compared to PCA and PLS [16].

The flexible characteristic of our proposed method makes it appropriate not only for correlating censored patient survival data with microarray gene expression data but also with large-scale biological data stemming from other high-throughput technologies such as DNA copy number analysis and proteomics.

Although the proposed method was better than supervised principal components and supervised partial least squares components based on two popular data sets and brief simulation, it is suggested that comprehensive simulation is used in future studies in order to evaluate of this method compared with the other dimension reduction methods.

The future investigations can focus on different ways of preselecting gene in the first stage of the proposed method. For example, rather than ranking genes based on their Wald's statistic, one would use a different metric to measure

the association between a given gene and survival time. On the other hands, another mother wavelet and different level of decomposition can be studied.

5. Conclusion

This study showed the Cox model based on supervised wavelet feature extraction method had superior predictive performance over the supervised principal components and supervised partial least squares components based on top selected genes. These results exhibit the possibility of developing more advanced tools using wavelets for the dimension reduction of microarray data sets in the context of survival analysis.

Conflict of interest statement

None declared.

Acknowledgements

This study is a part of PhD thesis in Biostatistics. Therefore, the authors thank the Vic-chancellor of Research and Technology of Hamadan University of Medical Sciences, Iran, for approving the project and providing financial support.

References

[1] A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Losses and A. Resenwald, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*,vol. 403, pp. 503-511, 2000.

[2] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, P. Mesirov et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531–537, 1999.

[3] H. M. Bovelstad, S. Nygard and Ø. Borgan, "Survival prediction from clinico-genomic models-a comparative study," *BMC Bioinformatics*, vol. 10, pp. 1-9, 2009.

[4] H. M. Bovelstad, S. Nygard and H. L. Storvold, "Predicting survival from microarray data--a comparative study," *Bioinformatics*, vol. 23, pp. 2080-2087, 2007.

[5] J. P. Klein and M. L. Moeschberger, "Survival analysis: Techniques for censored and truncated data", 2nd ed. Springer-Verlag, New York, 2003.

[6] N. Wessel, V. Wieringen, D. Kuna, R. Hampelb and A. L. Boulesteix, "Survival prediction using gene expression data: A review and comparison," *Computational Statistics & Data Analysis*, vol. 53, pp. 1590-1603, 2009.

[7] A. Rosenwald, G. Wright, W. C. Chan, J. M. Connors and E. Campo, "The use of molecular profiling to predict survival after chemotherapy for diffuse large Bcell lymphoma," *New England Journal of Medicine*, vol. 346, pp. 1937–1947, 2002.

[8] X. Y. Liu, Y. Liang, Z. B. Xu, H. Zhang and K. S. Leung, "Adaptive L1/2 Shooting Regularization Method for Survival Analysis Using Gene Expression Data"*The Scientific World Journal*, Vol. 2013, no. 475702, 2013.

[9] D. V. Nguyen and D. M. Rocke, "Partial least squares proportional hazard regression for application to DNA microarray survival data," *Bioinformatics*, vol. 18, no.12, pp. 1625-1633, 2002.

[10] P. J. Park, L. Tian and I. S. Kohane, "Linking gene expression data with patient survival times using partial least squares," *Bioinformatics*, vol. 20, pp. 208-215, 2002.

[11] H. Li and Y. Luan, "Kernel Cox regression models for linking gene expression profiles to censored survival data," *Pacific Symposium of Biocomputing*, vol. 8, pp. 65-76, 2003.

[12] Y. Liu, "Feature extraction and dimensionality reduction for mass spectrometry data," *Computers in Biology and Medicine*, vol. 39, pp. 818–823, 2009.

[13] T. A. Tokuyasu, D. Albertson, D. Pinkel and A. Jain, "Wavelet transforms for the analysis of microarray experiments," *Proceedings of the Computational Systems Bioinformatics*, 2003.

[14] Y. Liu, "Dimensionality reduction and main component extraction of mass spectrometry cancer data," *Knowledge-Based Systems*, vol. 26, pp. 207–215, 2012.

[15] L. Nanni and A. Lumini, "Wavelet selection for disease classification by DNA microarray data," *Expert Systems with Applications*, vol.38, pp. 990-995, 2011.

[16] Y. Liu, U. Aickelin, J. Feyereisl and L. G. Durrant, "Wavelet feature extraction and genetic algorithm for biomarker detection in colorectal cancer data," *Knowledge-Based Systems*, vol. 37, pp. 502–514, 2013.

[17] A. M. Sarhan, "Wavelet-based feature extraction for DNA microarray classification," *Artificial Intelligence Review*, vol. 39, no. 3, pp. 237-249, 2013.

[18] E. Bair and R. Tibshirani, "Semi-Supervised methods to predict patient survival from gene expression data," *PLoS Biology*, vol. 2, pp. 0511-0522, 2004.

[19] E. Bair, T. Hastie, D. Paul and R.Tibshirani, "Prediction by supervised principal components," *Journal of the American Statistical Association*, vol. 101, pp. 119-136, 2006.

[20] D. G. Beer, S. L. Kardia, C. Huang, T. J. Giordano, A. M. Levin et al., "Gene-expression profiles predict survival of patients with lung adenocarcinoma.," *Nature Medicine*, vol.8, no. 8, pp. 816-824, 2002.

[21] Y. Zhao and R. Simon, "BRB ArrayTools data archive for human cancer gene expression: a unique and efficient data sharing resource," *Cancer Informatics*, vol. 6, pp. 9-15, 2008.

[22] N. J. S. Nagelkerke, "A note on a general definition of the coefficient of determination," *Biometrika*, vol. 78, pp. 691-692, 1991.

[23] M. Gonen and G. Heller, "Concordance Probability and Discriminatory Power in Proportional Hazards Regression," *Biometrika*, vol. 92, no. 4, pp. 965-970, 2005.

[24] M. J. Pencina and R. B. Agostino, "Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation," *Statistics in Medicine* , vol. 23, pp. 2109–2123, 2004.

[25] E. Graf, C. Schmoor, W. Sauerbrei and M. Schumacher, "Assessment and comparison of prognostic classification schemes for survival data," *Statistics in Medicine*, vol. 18, pp. 2529–2545, 1999.

TABLE 1: Performance of different Cox models for simulated dataset.

#Gene	Method	Cindex $\pm$ se	CPE $\pm$ se	$R^2 \pm$ se	$LR \pm se$	$IBS \pm se$
40	Supervised Wavelet	0.924± 0.002	0.904± 0.003	0.766± 0.006	96.906± 1.729	0.153± 0.000
	Supervised PCA	0.907± 0.002	0.850± 0.003	0.709± 0.003	81.564± 0.700	0.153± 0.000
	Supervised PLS	0.919± 0.005	0.865± 0.005	0.739± 0.005	89.083± 1.311	0.155± 0.000
30	Supervised Wavelet	0.914± 0.002	0.877± 0.004	0.720± 0.009	83.313± 2.284	0.150± 0.000
	Supervised PCA	0.897± 0.003	0.842± 0.014	0.684± 0.007	76.448± 1.410	0.151± 0.004
	Supervised PLS	0.910± 0.003	0.853± 0.016	0.711± 0.008	82.436± 1.791	0.151± 0.004
20	Supervised Wavelet	0.899± 0.006	0.837± 0.030	0.682± 0.005	72.253± 2.233	0.153± 0.003
	Supervised PCA	0.886± 0.004	0.827± 0.025	0.648± 0.009	69.357± 1.873	0.154± 0.004
	Supervised PLS	0.895± 0.003	0.835± 0.027	0.669± 0.011	73.691± 2.273	0.154± 0.003
10	Supervised Wavelet	0.870± 0.006	0.823± 0.023	0.618± 0.013	65.800± 1.419	0.154± 0.004
	Supervised PCA	0.855± 0.011	0.810± 0.002	0.582± 0.008	58.072± 1.845	0.154± 0.003
	Supervised PLS	0.866± 0.009	0.818± 0.001	0.609± 0.009	62.484± 1.767	0.156± 0.003

TABLE 2: Performance of different Cox models for DLBCL dataset.

#Gene	Method	Cindex $\pm$ se	CPE $\pm$ se	$R^2 \pm$ se	$LR \pm se$	$IBS \pm se$
40	Supervised Wavelet	0.755 $\pm$ 0.005	0.744 $\pm$ 0.004	0.401 $\pm$ 0.011	78.739$\pm$ 1.815	0.237$\pm$ 0.007
	Supervised PCA	0.711 $\pm$ 0.004	0.695 $\pm$ 0.003	0.270$\pm$ 0.000	42.636$\pm$ 1.762	0.245$\pm$ 0.005
	Supervised PLS	0.723$\pm$ 0.003	0.698$\pm$ 0.003	0.294$\pm$ 0.007	55.883$\pm$ 1.449	0.250$\pm$ 0.005
30	Supervised Wavelet	0.723 $\pm$ 0.005	0.727 $\pm$ 0.007	0.325 $\pm$ 0.013	70.303$\pm$ 2.618	0.244$\pm$ 0.004
	Supervised PCA	0.709 $\pm$ 0.004	0.692 $\pm$ 0.003	0.262 $\pm$ 0.008	42.087$\pm$ 1.825	0.245$\pm$ 0.003
	Supervised PLS	0.713$\pm$ 0.002	0.697$\pm$ 0.002	0.289$\pm$ 0.007	54.898$\pm$ 1.418	0.251$\pm$ 0.004
20	Supervised Wavelet	0.730$\pm$ 0.002	0.714 $\pm$ 0.002	0.323 $\pm$ 0.009	59.708$\pm$ 2.699	0.243$\pm$ 0.004
	Supervised PCA	0.709 $\pm$ 0.003	0.688 $\pm$ 0.003	0.260$\pm$ 0.008	41.327$\pm$ 2.079	0.245$\pm$ 0.003
	Supervised PLS	0.719$\pm$ 0.002	0.696$\pm$ 0.003	0.282$\pm$ 0.006	53.130$\pm$ 1.486	0.249$\pm$ 0.004
10	Supervised Wavelet	0.703$\pm$ 0.004	0.686 $\pm$ 0.005	0.255 $\pm$ 0.007	49.838$\pm$ 1.832	0.248$\pm$ 0.003
	Supervised PCA	0.699 $\pm$ 0.005	0.686 $\pm$ 0.003	0.254 $\pm$ 0.013	41.056$\pm$ 2.045	0.252$\pm$ 0.004
	Supervised PLS	0.701$\pm$ 0.003	0.684$\pm$ 0.003	0.255$\pm$ 0.007	45.648$\pm$ 2.241	0.254$\pm$ 0.006

TABLE 3: Performance of different Cox models for Lung cancer dataset.

#Gene	Method	Cindex $\pm$ se	CPE $\pm$ se	$R^2 \pm$ se	$LR \pm se$	$IBS \pm se$
20	Supervised Wavelet	0.923 $\pm$ 0.005	0.876 $\pm$ 0.007	0.582 $\pm$ 0.014	54.986$\pm$ 2.130	0.328$\pm$ 0.015
	Supervised PCA	0.892 $\pm$ 0.003	0.796 $\pm$ 0.010	0.471 $\pm$ 0.014	38.609$\pm$ 1.637	0.353$\pm$ 0.009
	Supervised PLS	0.909$\pm$ 0.005	0.801$\pm$ 0.005	0.498$\pm$ 0.008	40.77$\pm$ 1.439	0.365$\pm$ 0.011
15	Supervised Wavelet	0.905 $\pm$ 0.004	0.846 $\pm$ 0.005	0.531 $\pm$ 0.007	45.466$\pm$ 1.838	0.343$\pm$ 0.007
	Supervised PCA	0.894 $\pm$ 0.003	0.801 $\pm$ 0.007	0.469 $\pm$ 0.010	38.263$\pm$ 1.678	0.349$\pm$ 0.007
	Supervised PLS	0.900$\pm$ 0.002	0.803$\pm$ 0.005	0.483$\pm$ 0.008	39.954$\pm$ 1.382	0.353$\pm$ 0.009
10	Supervised Wavelet	0.889 $\pm$ 0.006	0.813 $\pm$ 0.006	0.462 $\pm$ 0.018	38.357$\pm$ 1.641	0.330$\pm$ 0.010
	Supervised PCA	0.878 $\pm$ 0.005	0.784 $\pm$ 0.009	0.441 $\pm$ 0.008	34.217$\pm$ 1.671	0.335$\pm$ 0.008
	Supervised PLS	0.885$\pm$ 0.003	0.788$\pm$ 0.004	0.448$\pm$ 0.007	36.087$\pm$ 1.356	0.350$\pm$ 0.007
5	Supervised Wavelet	0.873 $\pm$ 0.006	0.795 $\pm$ 0.005	0.429 $\pm$ 0.001	31.906$\pm$ 1.786	0.297$\pm$ 0.007
	Supervised PCA	0.853 $\pm$ 0.005	0.775 $\pm$ 0.006	0.387 $\pm$ 0.012	29.241$\pm$ 1.784	0.315$\pm$ 0.006
	Supervised PLS	0.858$\pm$ 0.005	0.771$\pm$ 0.006	0.386$\pm$ 0.010	29.650$\pm$ 1.313	0.323$\pm$ 0.006

TABLE 4: Performance of different Cox models for Lung cancer dataset (Clinical + Genomic data).

#Gene	Method	Cindex $\pm$ se	CPE $\pm$ se	$R^2 \pm$ se	$LR \pm se$	$IBS \pm se$
20	Supervised Wavelet	0.949± 0.006	0.924± 0.010	0.669± 0.031	72.304± 2.589	0.431± 0.007
	Supervised PCA	0.907± 0.008	0.844± 0.009	0.553± 0.033	52.020± 2.208	0.432± 0.007
	Supervised PLS	0.914± 0.007	0.849± 0.009	0.564± 0.035	53.814± 2.366	0.435± 0.009
15	Supervised Wavelet	0.916± 0.005	0.855± 0.011	0.558± 0.031	56.318± 3.017	0.433± 0.010
	Supervised PCA	0.903± 0.007	0.836± 0.010	0.540± 0.034	53.478± 2.585	0.435± 0.009
	Supervised PLS	0.908± 0.007	0.842± 0.012	0.552± 0.041	55.526± 2.398	0.435± 0.006
10	Supervised Wavelet	0.906± 0.006	0.848± 0.008	0.552± 0.027	52.746± 2.872	0.426± 0.006
	Supervised PCA	0.892± 0.009	0.831± 0.008	0.521± 0.029	48.092± 2.119	0.426± 0.007
	Supervised PLS	0.905± 0.009	0.842± 0.009	0.542± 0.031	51.472± 2.562	0.430± 0.005
5	Supervised Wavelet	0.895± 0.008	0.818± 0.011	0.499± 0.036	51.472± 2.760	0.352± 0.008
	Supervised PCA	0.883± 0.009	0.803± 0.010	0.445± 0.042	46.336± 2.113	0.359± 0.008
	Supervised PLS	0.879± 0.007	0.814± 0.010	0.481± 0.029	49.976± 2.152	0.355± 0.006

Figure Captions:

FIGURE 1: The 1D wavelet decomposition process

FIGURE 2: Box plot of the difference in model evaluation criteria between the supervised wavelet and the two other methods for simulated dataset with different number of preselected genes.

FIGURE 3: Box plot of the difference in model evaluation criteria between the supervised wavelet and the two other methods for DLBCL dataset with different number of preselected genes.

FIGURE 4: Box plot of the difference in model evaluation criteria between the supervised wavelet and the two other methods for Lung dataset with different number of preselected genes.

Figure 1

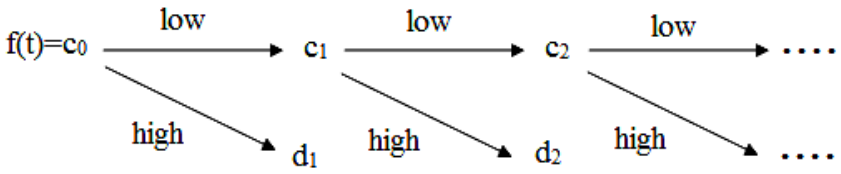


Figure 2

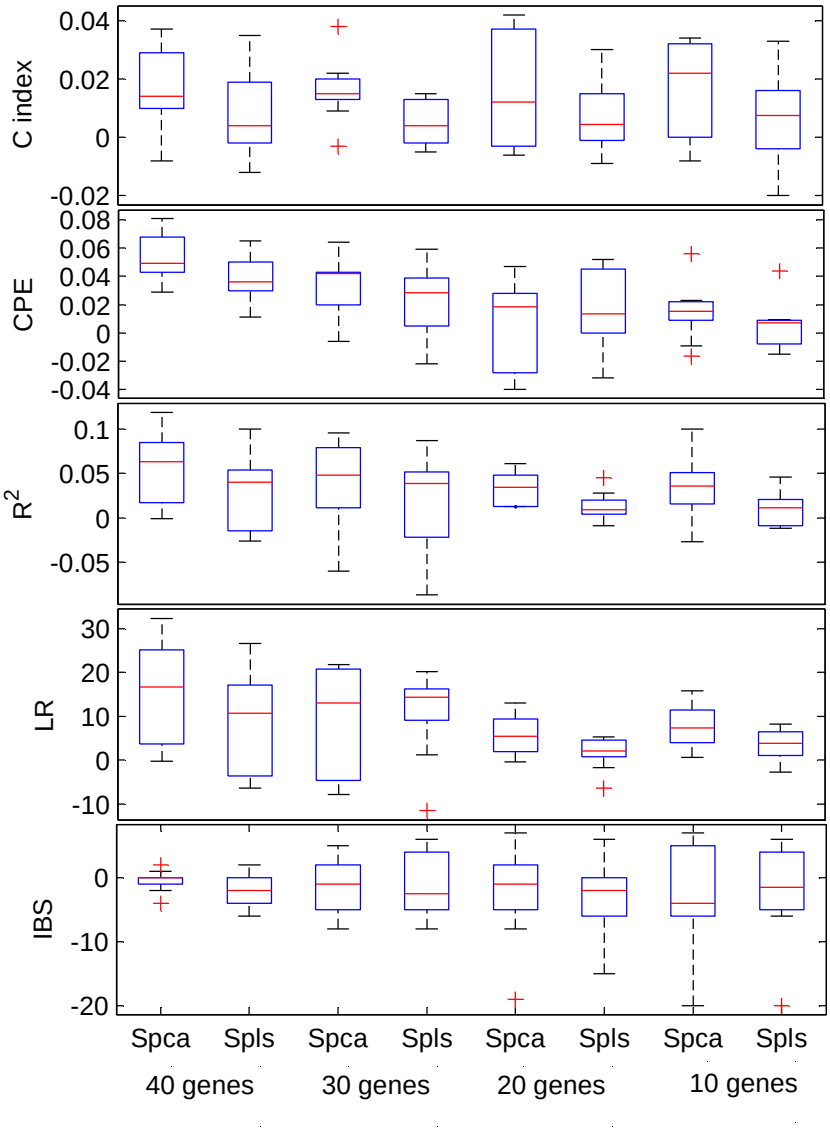

C index
0.04
0.02
0
-0.02
CPE
0.08
0.06
0.04
0.02
0
-0.02
-0.04
R2
0.1
0.05
0
-0.05
LR
30
20
10
0
-10
IBS
0
-10
-20
Spca
Spls
Spca
Spls
Spca
Spls
Spca
Spls
40 genes
30 genes
20 genes
10 genes

Figure 3

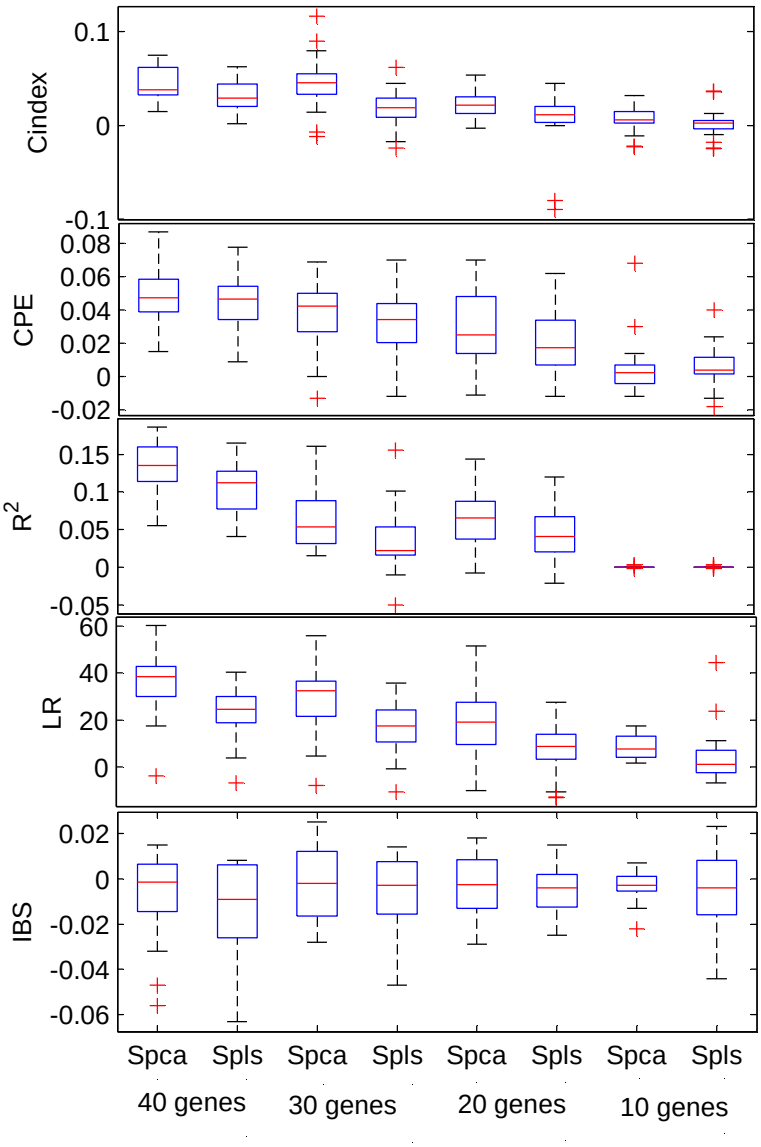
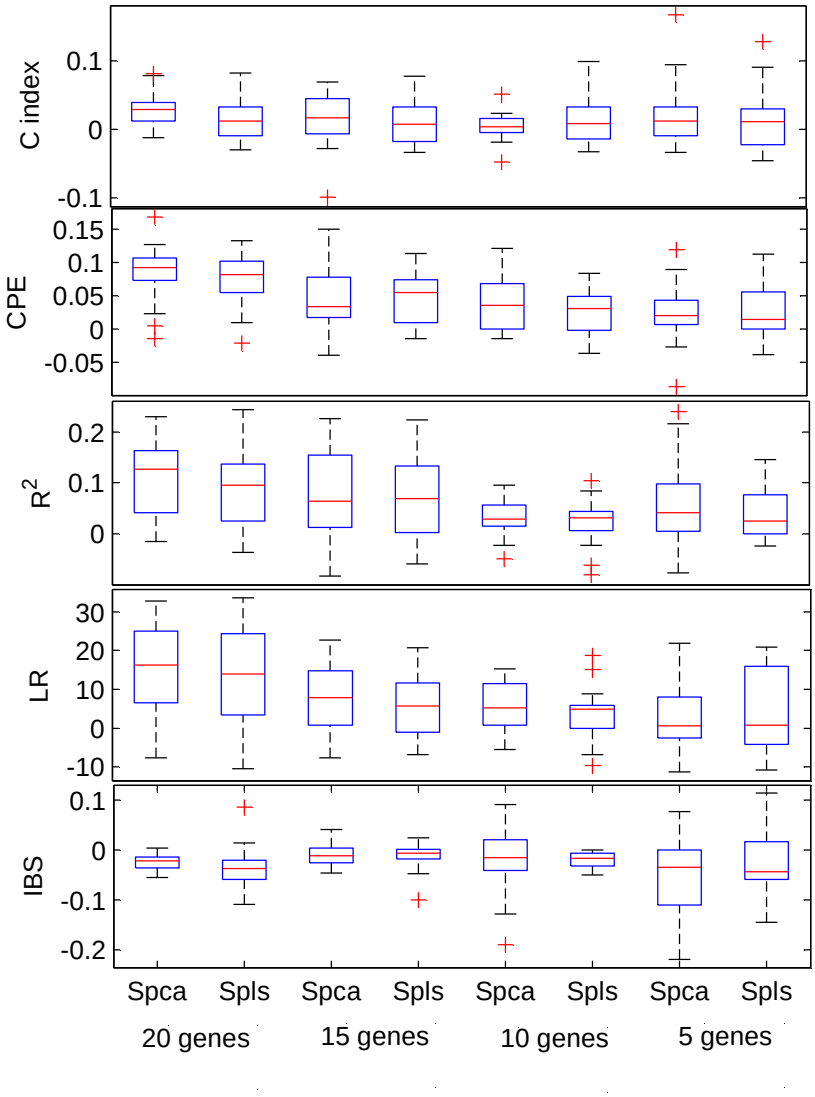

Cindex
0.1
0
-0.1
CPE
0.08
0.06
0.04
0.02
0
-0.02
R2
0.15
0.1
0.05
0
-0.05
LR
60
40
20
0
IBS
0.02
0
-0.02
-0.04
-0.06
Spca
Spls
Spca
Spls
Spca
Spls
Spca
Spls
40 genes
30 genes
20 genes
10 genes

Figure 4


C index
0.1
0
-0.1
CPE
0.15
0.1
0.05
0
-0.05
R^2
0.2
0.1
0
LR
30
20
10
0
-10
IBS
0.1
0
-0.1
-0.2
Spca
Spls
Spca
Spls
Spca
Spls
Spca
Spls
20 genes
15 genes
10 genes
5 genes