ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

An Alternative Dimension Reduction Approach to Supervised Principal Components Analysis in High Dimensional Survival Data

Çok Boyutlu Sağkalım Verilerinde Denetimli Temel Bileşenler Analizine Alternatif Bir Boyut İndirgeme Yaklaşımı

Elvan AKTÜRK HAYAT,[a]
Mevlüt TÜRE,[b]
Şanslı ŞENOL[c]

[a]Department of Econometrics, Adnan Menderes University Aydın Faculty of Economics,
[b]Department of Biostatistics and Medical Informatics, Adnan Menderes University Faculty of Medicine, Aydın
[c]Department of Statistics, Ege University Faculty of Science, İzmir

Geliş Tarihi/*Received:* 27.01.2016
Kabul Tarihi/*Accepted:* 22.02.2016

Abstract of this article was presented at International Attended 15th National Biostatistics Congress (20-23 August 2013, Aydın, Turkey).

Yazışma Adresi/*Correspondence:*
Elvan AKTÜRK HAYAT
Adnan Menderes University Aydın Faculty of Economics, Department of Econometrics, Aydın, TÜRKİYE/TURKEY
elvanakturk@gmail.com

ABSTRACT Objective: This study aims at comparing the performances of supervised principal component analysis (SPCA) which is used for dimension reduction and an alternatively proposed approach of nonlinear principal component analysis using artificial neural networks performed by gene selection with survival tree (survival tree based NLPCA- NN). **Material and Methods:** Gene expression data set from Rosenwald et al.(2002) pertaining to 240 patients with diffuse large B-cell lymphoma (DLBCL) is used. While Cox scores are used for determining important genes from high dimensional gene expression data in SPCA, in survival tree based NLPCA-NN approach, importance values of the survival tree are used. Important genes according to the Cox scores are reduced to three principal components by singular value decomposition. Important genes determined by the survival tree are taken as input variables in neural networks and reduced to three principal components. The performances of SPCA and survival tree based NLPCA-NN are compared by using Cox regression models (CRM). C index is calculated to compare obtained Cox regression models. **Results:** According to Cox scores, 121 genes are determined; according to importance values of survival tree 114 genes are determined. The percentages of variances explained by SPCA and survival tree based NLPCA-NN were 18.2% and 35.1% respectively. Harrell's C indexes are calculated as 0.726 for CRM-1, 0.687 for CRM-2. **Conclusion:** As a result, while SPCA takes only the linear relationships into consideration, survival tree based NLPCA-NN also takes non-linear relationships into account and has more variance explanation and NLPCA-NN can be evaluated as an alternative method to SPCA.

Key Words: Dimension reduction; supervised principal component analysis; survival tree; artificial neural networks; Cox regression analysis, gene expression data

ÖZET Amaç: Bu çalışmada, boyut indirgemede kullanılan denetimli temel bileşenler analizi (D-TBA) ile bu yönteme alternatif bir yaklaşım olarak önerilen sağkalım ağacıyla gen seçerek uygulanan yapay sinir ağlarıyla doğrusal olmayan temel bileşenler analizinin (sağkalım ağacı temelinde YSA-DOTBA) performanslarının karşılaştırılması amaçlandı. **Gereç ve Yöntemler:** Çalışmada Rosenwald ve diğerlerinin (2002) çalışmasından elde edilen yaygın B-hücreli lenfoma (DLBCL) hastası 240 bireye ilişkin gen ekspresyon verileri kullanıldı. D-TBA'da çok boyutlu gen ekspresyon verilerinden önemli genlerin belirlenmesinde Cox skorlar kullanılırken, Sağkalım ağacı temelinde YSA-DOTBA yaklaşımında önemli genlerin belirlenmesinde sağkalım ağacının önemlilik değerleri kullanıldı. Cox skorlara göre önemli olduğu belirlenen genler tekil değer ayrışması ile 3 temel bileşene indirgendi. Sağkalım ağacıyla önemli bulunan genler YSA'da girdi değişkeni olarak alınarak, 3 temel bileşene indirgendi. D-TBA ve sağkalım ağacı temelinde YSA-DOTBA'nın performansları Cox regresyon modeli (CRM) ile karşılaştırıldı. Elde edilen Cox regresyon modellerini karşılaştırmak için Harrell'in C indeksi hesaplandı. **Bulgular:** Cox skorlara göre 121 adet genin, sağkalım ağacının önemlilik değerlerine göre 114 adet genin önemli genler olduğu belirlendi. D-TBA'nın varyans açıklama oranı %18.2, sağkalım ağacı temelinde YSA-DOTBA'nın varyans açıklama oranı %35.1 bulundu. Harrell'in C indeksi CRM-1 için 0.726, CRM-2 için 0.687 olarak hesaplandı. **Sonuç:** Sonuç olarak D-TBA, sadece doğrusal ilişkileri göz önüne alırken, sağkalım ağacı temelinde YSA-DOTBA, doğrusal olmayan ilişkileri de dikkate alması ve daha fazla varyans açıklama oranına sahip olması açısından D-TBA'ya alternatif bir yöntem olarak değerlendirilmelidir.

Anahtar Kelimeler: Boyut indirgeme; denetimli temel bileşenler analizi; sağkalım ağacı; yapay sinir ağları; Cox regresyon analizi, gen ekspresyon verisi

doi: 10.5336/biostatic.2016-50294

Turkiye Klinikleri J Biostat 2016;8(1):21-9

In parallel to technological developments, data collection and storage capacity increases have brought along the problem of high-dimensionality. High number of variables and interactions between variables make it harder to interpret and summarize the results in data set analysis.[1-3] Currently, high dimensional regression problems where the number of predictors exceeds the number of observations have become the focus of statistical research.[4,5] Survival prediction from gene expression data and other high-dimensional genomic data have been the research topic of numerous studies recently.[6-10]

Traditionally, the Cox proportional hazard model, which is one of the survival prediction methods, is applied in a situation where the number of observations exceeds the number of independent variables. In situations where the number of predictors greatly exceeds the number of observations Cox proportional hazard model cannot be applied since multicollinearity between variables in predicting survival with gene expression data would increase.[9-11] Therefore, dimension reduction becomes inevitable in survival analysis with high dimensional data. In the dimension reduction stage Wald statistics or Cox score statistics are generally used for determining genes that are associated with survival from gene expression data. Survival trees can be used as an alternative method to Wald statistics and Cox score statistics, being a method determining the variables associated with survival time in order of importance.

Principal components analysis (PCA), a linear dimension reduction method, is an effective method for explaining the correlated structures in gene expression data. PCA explains the variance-covariance structure of a set of variables through a few linear combinations of these variables.[1,12] However, the irrelevance of the first PC of PCA to the dependent variable causes insufficient results of PCA.[13] Supervised principal components analysis (SPCA) proposed by Bair and Tibshirani (2004) for gene set analysis is a method that performs PCA to the independent variables which are related to the dependent variables. Artificial neural networks which is an alternative method to PCA, is a computer algorithm that describes and classifies data structures by modelling the behavior and the structure of human brain.[14,15] While PCA determines linear relationships, nonlinear principal components analysis using neural networks (NLPCA-NN) is a dimension reduction method that considers both linear and nonlinear relationships.[16-20]

This study aims at analyzing the high dimensionality problem of gene data and comparing the performances of SPCA and survival tree based NLPCA-NN by taking a new dimension reduction approach into consideration which uses survival tree and NLPCA-NN as an alternative to Cox scores and PCA respectively.

MATERIAL AND METHODS

PATIENTS

In this study, gene expression data set from Rosenwald et al. (2002) pertaining to 240 samples from patients with diffuse large B-cell lymphoma (DLBCL) is used.[21] The dataset is available at http://llmpp.nih.gov/DLBCL/.

The data set comprises of 7399 gene expression values, survival times (years) and "International Prognostic Index- IPI" risk score for each patient. 130 (57.5%) patients are completed observation (dead), while 102 (42.5%) patients are uncompleted observations (alive). According to IPI risk score, 82 (34.2%) patients are in low risk group, 108 (45%) are in medium risk group, 32 (13.3%) are in high risk group and 18 (7.5%) are unknown.

STATISTICAL ANALYSIS

We randomly divided the samples into a training set of size 160 and a test set of size 80. SPCA and survival tree based NLPCA-NN are applied to this data. The flow chart for both methods is given Figure 1. We used R (2.11.1) software packages (http://www.r-project.org/) '*super pc'* for SPCA, '*rpart'* for survival tree. NLPCA-NN is applied with Hsieh's (2007) algorithm in Matlab7.1. and CRM is performed with SPSS packages.

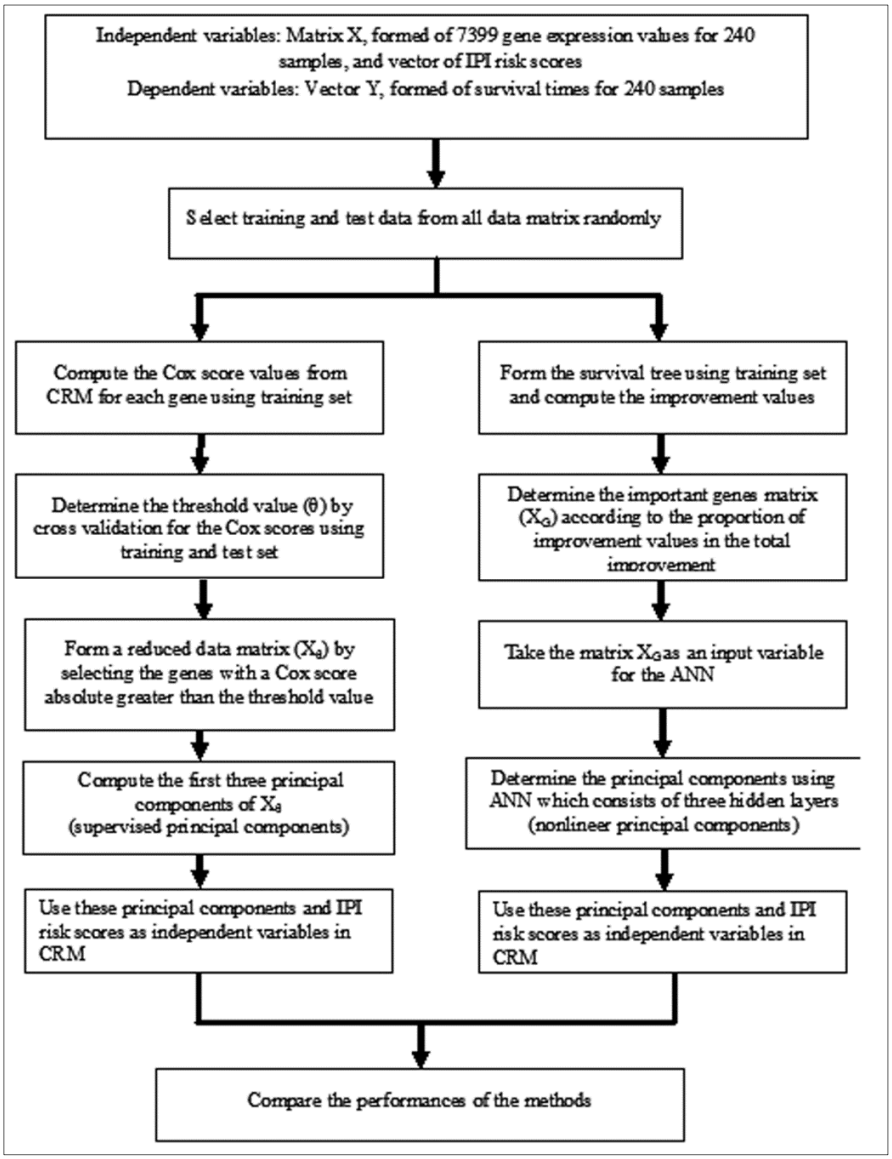

FIGURE 1: The flow chart for SPCA and survival tree based NLPCA-NN.

GENE SELECTION WITH COX SCORES

In order to determine the genes that can be related to survival from multidimensional data, the score statistics or the Cox score values are calculated by obtaining the CRM values for each gene, Cox scores measures whether an independent variable is a good predictor in survival predictor. The score statistic for the j[th] variable is calculated using the $S_j = \left(\partial l(0)/\partial \beta_j\right)/\left(\partial^2 l(0)/\partial \beta_j^2\right)^{1/2}$equity and shows a standard normal distribution. In this equation, $l(.)$ indicates a logarithmic partial likelihood function and β_j indicates Cox regression parameters for j[th] variable. This value's being greater than the table value enables the denial of the null hypothesis which is constructed as variable j is not related to the dependent variables. Thus, the genes which are important in survival prediction could be determined.[22] In the gene selection step, the genes with a Cox score absolute greater than a certain threshold value are considered as important genes. In order to determine this threshold value, the training set is divided into k cross-validation tests, the CRM is tried for k-1 and the model is tested for the one left in each step; and the logarithmic likelihood ratio statistics are calculated. The greatest value for the

logarithmic likelihood ratio statistics is selected as the threshold value.[6,7,23]

SUPERVISED PRINCIPAL COMPONENT ANALYSIS

Principal components in SPCA, proposed by Bair and Tibshirani (2004) for gene expression data analysis, are estimated from a subset of genes which are related to the dependent variables.[6] Let X which is formed of p variables for n observations be an n x p matrix of independent variables and y be the n vector of dependent variable.

SPCA briefly comprises of the following steps:[4,7]

i. Univariate regression coefficients are calculated for each variable (e.g. Cox regression coefficients for survival data)

ii. A reduced data matrix comprising of only those variables whose univariate coefficient exceeds a threshold θ in absolute value (θ is estimated by cross-validation) is formed.

iii. The first principal component/components from the reduced data matrix are calculated.

iv. These principal components are used in Cox regression model in order to predict the dependent variable.

The singular value decomposition (SVD) of X can be written as below:[4,6,7]

$$X = UDV^T \quad (1)$$

In Equation (1);

Where U, D and V are (n x m), (m x m), and (m x p) and m=min (n-1, p) is the rank of X. Here D is a diagonal matrix containing the singular values d_i , and the columns of U are the principal components $u_1, u_2, \ldots, u_m$ these are assumed to be ordered, so that $d_1 \geq d_2 \geq \ldots \geq d_m \geq 0$.

s is the p dimensional vector of the standardized regression coefficients in which the univariate effects of each gene on y separately are calculated, and is calculated as $s_j = (x_j^T y)/\sqrt{x_j^T x_j}$. Let C_θ be the sum of $|s_j| > \theta$ indexes and let the matrix comprising the columns of X pertaining to C_θ be X_θ. Singular value decomposition of X_θ is shown as $X_\theta = U_\theta D_\theta V^T$. Here $U_\theta = (u_{\theta,1}, u_{\theta,2}, \ldots, u_{\theta,m})$ and $u_{\theta,1}$ are the first supervised principal component of X .[7]

AN ALTERNATIVE APPROACH TO SUPERVISED PRINCIPAL COMPONENT ANALYSIS

Gene Selection with Survival Tree

Tree classification is the non-parametric alternative of semi-parametric CRM for the prognostic groups of survival data. Tree based methods recursively partition the covariate space into different regions and the corresponding data into groups. With the tree structure starting from the root node at the top; all possible partitions for independent variables are evaluated, and points that would partition the nodes homogenously are selected.[24] Recursive partition algorithms partition the covariate space into regions according to a rule which maximizes the improvement measures. LeBlanc and Crowley's (1992) algorithm partitions the covariate space into regions by enabling the maximization of the decreases in the one-step full likelihood deviance.[25] The partitions for only one variable are discussed in the algorithm. Let N be the total number of observations in learning sample. The s improvement for partition values for the h node into l(h) and r(h) sub-nodes are calculated as below:

$$R(s,h) = R(h) - [R(l(h)) + R(r(h))] \quad (2)$$

$$R(h) = \frac{1}{N} \sum_{i \in S_h} \left[\delta_i \log\left(\frac{\delta_i}{\hat{\Lambda}_0^1(t_i)\hat{\theta}_h} \right) - (\delta_i - \hat{\Lambda}_0^1(t_i)\hat{\theta}_h) \right]$$

$R(h)$ is the deviance for node h.

N , is the total number of observations in the learning sample.

(t_i, δ_i) , is the vector of observation time and failure indicator for individual i.

S_h , is the set of observation labels, $\{i : x_i \in x_h\}$, for observations in the region X_h corresponding to node h.

$\hat{\Lambda}_0^j(t)$ is the Breslow cumulative hazard estimate for iteration j.

θ_h is a nonnegative parameter of the proportional hazards model.

$\hat{\theta}_h$ is the estimate of θ_h.

All binary partitions are found until a few observations are left in each node and a great tree is obtained.[25]

Nonlinear Principal Component Analysis with Neural Networks

The NLPCA-NN method proposed by Kramer (1991), is a non-linear generalization of PCA. According to Kramer (1991) if there are non-linear relations between the variables, NLPCA-NN partitions the data into its components using a smaller number of factors and with a higher explanatoriness. In order to implement NLPCA-NN, a Multi-Layer Perceptron (MLP) is used, comprising of three hidden layers between the input layer (X) and the output layer (X').[3,16,17,19,20,26,27] In order to minimize the mean square error of this approach, the objective function $E=\sum_{i=1}^{n}(x-x')^2$ should be minimized.[16,17,19,20,26,27]

The f_1 activation function in Figure 2 makes a prediction according to the $h_k = f_1((W^{(x)}x+\phi^{(x)})_k)$, k=1,..,r equity from x to h_k, the hidden layer 1 (Figure 2). Here $W^{(x)}$ shows the $r\times p$ dimensional weight matrix and $\phi^{(x)}$, the r dimensional column vector, shows the biased parameters. The second activation function f_2 makes a prediction according to the $b = f_2(w^{(x)}h_k+\bar{\phi}^{(x)})$ equity. Here

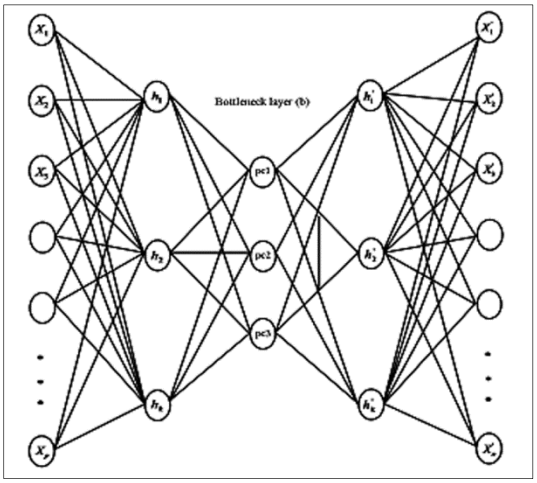

FIGURE 2: Auto-associative neural network schematic for NLPCA-NN.

$w^{(x)}$ is the weight matrix of the 2th layer and f_2 reveals the non-linear principal component data by making a prediction from the hidden layer 1 to the dimension reduction layer (b). f_3 activation function, on the other hand, makes a prediction from b to the last hidden layer using the h'_k, $h'_k = f_3((w^{(b)}b+\phi^{(b)})_k)$, k=1,.., r equity. Here, $w^{(b)}$ is the weight matrix of Bottleneck layer. f_4 makes a prediction from h'_k to x'; here x', is the p dimensional output vector, $x'_i = f_4((W^{(b)}h'_k+\bar{\phi}^{(b)})_i)$. $W^{(b)}$ is the weight matrix of the last hidden layer.

f_1 and f_3 are generally hyperbolic tangent or sigmoid functions. f_2 and f_4 are generally similarity functions.[20,26]

HARRELL'S CONCORDANCE INDEX

Harrell's concordance index (C index) is a measure of survival performance. The C index is defined as the proportion of all usable patient pairs in which the predictions are in concordance with the observation values.[28,29] C index which measures the predictor data obtained from independent variables set in a model, predicts the probability of consistency between the predicted dependent variables and observed dependent variables. 0.5 value indicates that there is not any predictor partition and 1 value indicates that patients with different observation values have excellent partition.[30] C index is calculated using the C=(E+T/2)/D equity.[29] Here E is the number of the pairs with concordant observations and predictions when predicted survival time is different; T is the number of concordant pairs with the same survival time; D is the total number of concordant pairs. Error rate is calculated as 1-C. Error rates may vary between 0 and 1. A value closer to 0 indicates an excellent consistency.[31-33]

RESULTS

SUPERVISED PRINCIPAL COMPONENT ANALYSIS RESULTS

Cox score values for 7399 genes are calculated using the training set. In order to determine θ, likelihood ratio statistics and corresponding threshold values are calculated from the test

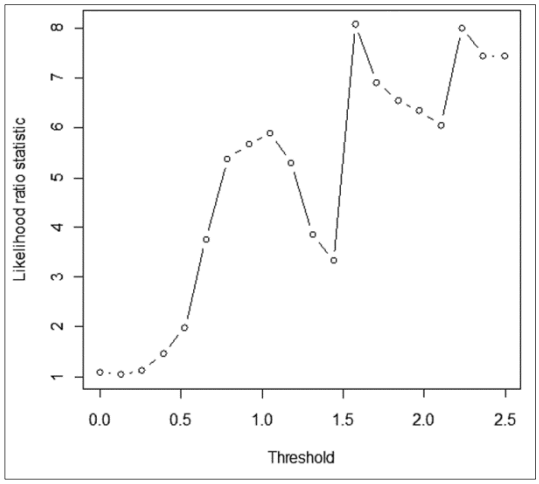

FIGURE 3: Threshold values versus logarithmic likelihood ratio statistics graph derived from test set.

data (Figure 3). The threshold value maximizing the logarithmic likelihood ratio test statistics for Cox scores are found 1.57. 121 genes are determined which have greater Cox score values from 1.57, the absolute value of Cox scores. The reduced data matrix obtained from the gene expression values pertaining to the genes selected according to Cox scores. The principal components, obtained from the singular value decomposition of this matrix, are supervised principal components. The variance explanation ratio for these principal components is calculated as 18.2%.

SURVIVAL TREE BASED NLPCA-NN APPROACH RESULTS

The training set is analyzed using the survival tree algorithm and it is found that 114 genes were related to survival. The gene expression values for these 114 important genes are taken as input variables for the ANN. ANN's input and output layers are constructed from 114 neurons, the second and the fourth layers are constructed from 3 neurons each, and the dimension reduction layer is constructed from 3 neurons. The second and the fourth layers of the network are hyperbolic tangent functions and the rest are linear functions. 114 genes are reduced to 3 nonlinear principal components at the end of NLPCA-NN, and the total variance explanation ratio is calculated as 35.1%.

COMPARISON OF THE PERFORMANCES OF SPCA AND SURVIVAL TREE BASED NN-PCA

Two separate CRM models which are CRM-1 and CRM-2, in order to determine the factors that influence the survival of the patients, have been applied. CRM-1 is applied using the three principal components obtained from the SPCA and the IPI risk scores as explanatory variables. CRM-2 is applied using the three principal components obtained from survival tree based NLPCA-NN methods and the IPI risk scores as explanatory variables.

The forward stepwise likelihood ratio method is used in CRMs. The performances of CRM-1 and CRM-2 models are compared using the Harrell's concordance index.

It is found that the first supervised principal component (S-PC1) variables had significant effects on survival time ($p<0.001$), which are calculated using the IPI risk score and SPCA in the reduced model, obtained in the second step of the CRM's forward stepwise likelihood ratio selection method. It is also found that S-PC1 ($p<0.001$) variable and the medium ($p<0.001$) and high ($p<0.001$) levels of and the IPI risk score had substantial effects of survival time. It is determined that death risk increased by 2.872 times (95% CI: 2.064-3.997) as the value of S-PC1 variable increased; and also the death risk of patients in the IPI medium risk group increased by 2.457 times (95% CI: 1.583-3.812) when compared to IPI low risk group, and it increased by 4.641 times (95% CI: 2.712-7.942) in IPI high risk group (Table 1).

It is found that the second principal component variables (NN-PC2) had significant effects on survival time ($p<0.001$), which are calculated using the IPI risk score and survival tree based NLPCA-NN in the reduced model, obtained in the second step of the CRM's forward stepwise likelihood ratio selection method. It is also found that NN-PC2 and the medium ($p<0.001$) and high ($p<0.001$) levels of IPI risk scores had substantial effects on survival time. It is concluded that death risk increased by 1.749 times (95% CI: 1.321-2.316) as the value of NN-PC2 variable increased. It is also found that the death risk of pa-

TABLE 1. CRM-1 results for SPCA's principal components.

Independent variable	$\hat{\beta}$	$\hat{\sigma}_{\hat{\beta}}$	HR	95% CI for HR	p
S-PC1	1.055	0.169	2.872	(2.064-3.997)	<0.001
IPI Group (medium)	0.899	0.224	2.457	(1.583-3.812)	<0.001
IPI Group (high)	1.535	0.274	4.641	(2.712-7.942)	<0.001

TABLE 2. CRM-2 results for survival tree based NLPCA-NN's principal components.

Independent variables	$\hat{\beta}$	$\hat{\sigma}_{\hat{\beta}}$	HR	95% CI for HR	p
NN-PC2	0.559	0.143	1.749	(1.321-2.316)	<0.001
IPI Group (medium)	0.789	0.230	2.201	(1.403-3.455)	<0.001
IPI Group (high)	1.552	0.274	4.720	(2.758-8.076)	<0.001

tients in the IPI medium risk group was increased by 2.201 times (95% CI: 1.403-3.455) when compared to IPI low risk group, and it increased by 4.720 times (95% CI: 2.758-8.076) in IPI high risk group (Table 2).

The 2log-likelihood values calculated from both models are low and this indicates the suitability of the CRM to the data set. While the 2log-likelihood value for CRM-1 is 1177.707, it is 1201.733 for CRM-2. Accordingly, CRM-1, which comprises of principal components obtained from SPCA, could be accepted as a more convenient model. Survival function values for both models are very close (Figure 4). Harrell's concordance index is calculated as 0.726 for CRM-1, and 0.687 for CRM-2. As for that, the probability of concordance between the predicted and observed variables in CRM-1 is found higher than the probability of concordance in CRM-2.

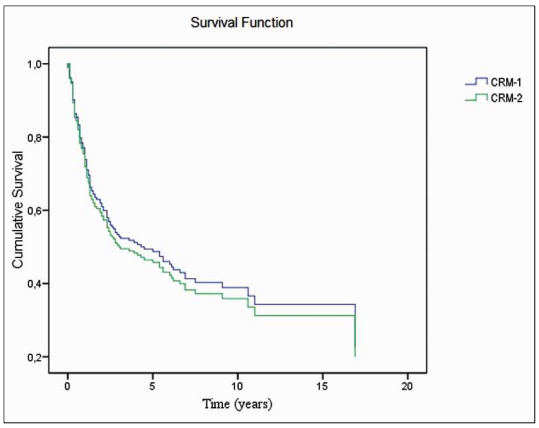

FIGURE 4: Survival function graphs for CRM-1 and CRM-2.

DISCUSSION

Gene expression data are characterized with high dimensionality. Numerous approaches have been proposed which develop a model on predicting the survival times by including gene profiles and the survival times of patients among these kinds of data. These approaches generally use dimension reduction and Cox regression analyses.[6-10,34] Van Wieringen et al. (2009) is recently reviewed survival prediction methods for gene expression data.[10] In addition, several authors have also proposed penalized partial likelihood approaches for the Cox PH model to cope with the high dimensionality of the gene expression data. Li and Luan (2003) used kernel transformations of the Cox partial likelihood in the framework of a penalization method.[35] Nyugen and Rocke (2002) used the multivariate partial minimum square method.[36] Nguyen and Rojo (2009), proposed a variant of Partial Least Squares, denoted as Rank-based Modified Partial Least Squares (RMPLS), that is insensitive to outlying values of both the response and the gene expressions.[34]

However, the disadvantage of these methods is that they use all genes in survival prediction. Since most of the genes in the data set are irrelevant of survival, irrelevant variables enter into the model and thus the prediction accuracy of the model decreases.[6]

SPCA and NLPCA-NN methods used in this study do not take all the genes into the model;

they choose the genes which are relevant to survival only. While SPCA makes a selection using Cox Scores to determine the genes that affect survival time; survival tree based NLPCA-NN uses survival trees to select the most important gene groups. As a result, the predictions obtained are more accurate.

In recent years, various unsupervised methods have also been developed to detect clusters of high dimensional gene expression data. Xu et al. (2010), used two step method to reduce the dimension of gene expression data. They extract a subset of genes based on statistical characteristics of their corresponding gene expression levels and fuzzy ART, a neural network clustering theory, is applied to the resulting data to generate clusters of cancer samples.[37]

Zhao et al. (2006) determined the genes that predict the survival times of the patients accurately from the gene expression profiles of survival related genes in kidney cancer patients, using hierarchical clustering and SPCA.[8] Quackenbush (2001) argued that PCA, when used together with other classification techniques, is a powerful technique in the analysis of gene expression data.[38] Berrar et al. (2005) pointed out that survival trees, one of the classification techniques, grouped patients into similar survival times and determined the genetic profiles in these groups, and could be used as a prediction model.[39] Zhang et al. (2001) indicated that the methods used in the analysis of gene expression data, such as cluster analysis, were unsupervised methods and especially clustering aimed at collecting similar genes together and this would pose a disadvantage in diagnosing the diseases. They also showed that recursive partition selected variables from different classes by comparing all variables, and it could be used in classifying the tissues using gene expression data. They stated that this technique yielded better results than other techniques due to its ease of application and flexibility.[40] Khan et al. (2001) used the principal components they obtained from PCA for classifying cell tumors into subclasses as input variable and implemented a linear ANN.[41] While PCA determines the linear relations between the variables, ANN is a method that reduced dimension by handling both linear and non-linear relations.[16,19,20,26] In this case, components with high explanatory ratio are obtained.

As for this study, an alternative approach to dimension reduction based survival prediction using the 7399 gene expression data from 240 individuals with diffuse large B-cell lymphoma taken from Rosenwald et al. (2002). In this approach which determines the important genes using survival trees, components with higher explanatory ratios, compared to SPCA, are obtained. While the variance explanation ratio for the three principal component obtained with SPCA is 18.2%, the variance explanation ratio for the three principal component obtained with survival tree based NLPCA-NN is 35.1%. According to Harrell's concordance index, the probability of concordance between the predicted and observed dependent variables in CRM-1 (SPCA components and IPI score) is 0.726, this value is 0.678 for CRM-2 (NLPCA-NN components and IPI score). Accordingly, it is found that the prediction performance of CRM-1 model was better than CRM-2.

In conclusion, although the CRM results which took the SPCA components and IPI scores as independent variables are better than the CRM results which took survival based NLPCA-NN components and IPI scores as independent variables, the survival tree based NLPCA-NN approach should be considered as an alternative to SPCA method, since it is important in terms of considering non-linear relations and having greater variance explanation ratio.

REFERENCES

1. Mardia KV, Kent JT, Bibby JM. Principal component analysis. In: Birnbaum ZW, Lukacs E, eds. Multivariate Analysis. 1st ed. New York: Academic Press; 1979. p.213-46.
2. Tatlıdil H. [Principal component analysis]. Uygulamalı Çok Değişkenli İstatistiksel Analiz. Ankara: Akademi Matbaası; 2002. p.138-62.
3. Ture M, Kurt I, Akturk Z. Comparison of dimension reduction methods using patient satisfaction data. Expert Systems with Applications 2007;32(2):422-6.
4. Hastie T, Tibshirani R, Friedman JH. High dimensional problems. Springer Series in Statistics. The Elements of Statistical Learning: Data Mining, Inference and Prediction. 2nd ed. Stanford, California: Springer; 2009. p.649-94.
5. Tibshirani R. Univariate shrinkage in the Cox model for high dimensional data. Stat Appl Genet Mol Biol 2009;8(1):Article 21.
6. Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. PloS Biol 2004;2(4):e108.
7. Bair E, Hastie T, Paul D, Tibshirani R. Prediction by supervised principal components. J Am Stat Assoc 2006;101(473):119-37.
8. Zhao H, Ljungberg B, Grankvist K, Rasmuson T, Tibshirani R, Brooks JD. Gene expression profiling predicts survival in conventional renal cell carcinoma. PloS Med 2006;3(1):115-24.
9. Bøvelstad HM, Nygard S, Størvold HL. Aldrin M, Borgan Ø, Frigessi A, et al. Predicting survival from microarray data-a comparative study. Bioinformatics 2007;23(16):2080-7.
10. Van Wieringen WN, Kun D, Hampel R, Boulesteix AL. Survival prediction using gene expression data: A review and comparison. Computational Statistics and Data Analysis 2009;53(5):1590-603.
11. Witten DM, Tibshirani R. Covariance-regularized regression and classification for high dimensional problems. Journal of the Royal Statistical Society Series B (Statistical Methodology) 2009;71(3):615-36.
12. Jolliffe IT. Interpreting principal components: Examples. Principal Component Analysis. 2nd ed. New York: Springer-Verlag; 2002. p.63-77.
13. Chen Xi, Wang L, Smith JD, Zhang B. Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes. Bioinformatics 2008;24(21):2474-81.
14. Bishop CM. Principal component analysis. Neural Networks for Pattern Recognition. 1st ed. New York: Oxford University Press; 1995. p.310-19.
15. Haykin S. Principal component analysis. Neural networks: A Comprehensive Foundation. 2nd ed. India: Prentice Hall; 2001. p.414-62.
16. Kramer MA. Nonlinear principal compon ent analysis us ing autoassociative neural networks. AIChE J 1991;37(2):233-43.
17. Monahan AH. Nonlinear principal component analysis by neural networks: theory and application Lorenz system. Journal of Climate 2000;13:821-35.
18. Monahan AH. Nonlinear principal component analysis: Tropical Indo-Pacific sea surface temperature and sea level pressure. Journal of Climate 2001;14(2):219-33.
19. Scholz M, Fraunholz M, Selbig J. Nonlinear principal component analysis: Neural network models and applications. In: Gorban AN, Kégl B, Wunsch DC, Zinovyev A, eds. Principal Manifolds for Data Visualization and Dimension Reduction. Vol. 58. LNCSE. Berlin Heidelberg: Springer; 2007. p.44-67.
20. Hsieh WW. Nonlinearprincipal component analysis. Machine Learning Methods in the Environmental Sciences: Neural Networks and Kernels. 1st ed. New York: Cambridge University Press; 2009. p.213-51.
21. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, et al; Lymphoma/Leukemia Molecular Profiling Project. The use of molecular profiling to predict survival after chemotherapy for diffuse large B-cell lymphoma. N Engl J Med 2002; 346(25):1937-47.
22. Witten DM, Tibshirani R. Survival analysis with high-dimensional covariates. Stat Methods Med Res 2010;19(1):29-51.
23. Boulesteix AL. Wilcox CV: an R package for fast variable selection in cross-validation. Bioinformatics 2007;23(13):1702-4.
24. Huang X, Chen SD, Soong SJ. Piecewise exponential survival trees with timedependent covariates. Biometrics 1998;54(4):1420-33.
25. Le Blanc M, Crowley J. Relative risk trees for censored survival data. Biometrics 1992; 48(2):411-25.
26. Hsieh WW. Nonlinear principal component analysis by neural networks. Tellus 2001;53A:599-615.
27. Hsieh WW. Nonlinear principal component analysis of noisy data. Neural Networks 2007;20(4):434-43.
28. Kattan MW, Hess KR, Beck JR. Experiments to determine whether recursive partitioning (CART) or an artificial neural network overcomes theoretical limitations of Cox proportional hazards regression. Comput Biomed Res 1998;31(5):363-73.
29. Newson RB. Comparing the predictive power of survival models using Harrell's C or Somers' D. Stata Journal 2010;10(3):339-58.
30. Harrell FE Jr, Lee KL, Daniel BM. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15(4):361-87.
31. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati R. Evaluating the yield of medical tests. JAMA 1982;247(18):2543-6.
32. Ishwaran H, Kogalur UB. Random survival forests for R. R News 2007;7(2):25-31.
33. Kurt Omurlu I, Ture M, Tokatli F. The comparisons of random survival forests and Cox regression analysis with simulation and an application related to breast cancer. Expert Systems with Applications 2009;36(4):8582-8.
34. Nguyen TS, Rojo J. Dimension reduction of microarray data in the presence of censored survival response: a simulation study. Stat Appl Genet Mole Biol 2009;8(1):Article 4.
35. Li H, Luan Y. Kernel Cox regression models for linking gene expression profiles to censored survival data. Pacific Symposium on Biocomputing 2003;8:65-76.
36. Nguyen DV, Rocke DM. Multi-class cancer classification via partial least squares with gene expression profiles. Bioinformatics 2002;18(9):1216-26.
37. Xu R, Damelin S, Nadler B, Wunsch DC 2nd. Clustering of high-dimensional gene expression data with feature filtering methods and diffusion maps. Artif Intell Med 2010;48(2-3): 91-8.
38. Quackenbush J. Computational analysis of microarray data. Nat Rev Genet 2001; 2(6):418-27.
39. Berrar D, Sturgeon B, Bradbury I, Downes CS, Dubitzky W. Survival trees for analyzing clinical outcome in lung adenocarcinomas based on gene expression profiles: identification of neogenin and diacylglycerol kinase alpha expression as critical factors. J Comput Biol 2005;12(5):534-44.
40. Zhang H, Yu CY, Singer B, Xiong M. Recursive partitioning for tumor classification with gene expression microarray data. PNAS 2001;98(12):6730-5.
41. Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 2001;7(6):673-9.