Supervised Principal Component Analysis Approach Based on Artificial Neural Networks in Gene Expression Data

Gen Ekspresyon Verilerinde Yapay Sinir Ağlarına Dayalı Denetimli Temel Bileşenler Analizi Yaklaşımı

Mevlüt Türe, İmran Kurt Ömürlü
Adnan Menderes Üniversitesi Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Aydın, Türkiye

Abstract: The aim of this study is dimension reduction of multidimensional gene expression data using supervised principal component analysis (S-PCA) and –proposed as a new approach- supervised principal component analysis with artificial neural networks (S-ANN-PCA) and to compare performances of these two methods by using random survival forests (RSF). In simulation application 5000 genes were generated according to multivariate normal distribution and then survival time that is correlated to these gene data were generated for 100 units. Simulation step was carried out with 1000 repetitions.

In addition, gene expression data for 240 individuals with extensive B-cell lymphoma (DLBCL) were used. Dimension reduction was done using Wald statistic in selection of important genes. The new data sets obtained from the methods were analyzed using RSF analysis.In the simulation application, it was obtained that the explanatoriness of S-PCA was significantly different from S-ANN-PCA ($p<0.001$). In the DLBCL data application, it was found that the error rate for the S-PCA was 36.78% and 43% for the S-ANN-PCA as a result of RSF. The importance value of S-PCA method was found to be higher and its error rate was found to be lower than the other method.S-PCA performed better than S-ANN-PCA in analyzing gene expression data experiencing a multidimensional problem.

Key Words: dimension reduction, neural networks, supervised principal component analysis, random survival forests, gene expression

Ture M , Omurlu Kurt I. 2018, Supervised Principal Component Analysis Approach Based on Artificial Neural Networks in Gene Expression Data , *Osmangazi Tıp Dergisi*, 40(1): 20-27 **Doi:** 10.20515/otd. 371882

Özet: Bu çalışmada, denetimli temel bileşenler analizi (D-TBA) ile yeni bir yaklaşım olarak önerilen yapay sinir ağlarıyla denetimli temel bileşenler analizi (D-YSA-TBA) kullanılarak çok boyutlu gen ekspresyon verilerinin boyutunun indirgenmesi ve random survival forests (RSF) analizi kullanılarak performansların karşılaştırılması amaçlandı. Simülasyon uygulamasında çok değişkenli normal dağılımdan 100 birim için 5000 gen ve bu gen verisi ile ilişkili yaşam süresi verisi türetildi. Simülasyon aşaması 1000 tekrarlı olarak gerçekleştirildi. Ayrıca yaygın B-hücreli lenfoma (DLBCL) hastası 240 bireye ilişkin gen ekspresyon verileri kullanıldı. Önemli genlerin seçiminde Wald istatistiği kullanılarak boyut indirgemesi yapıldı. Yöntemlerden elde edilen yeni veri setleri RSF analizi kullanılarak analiz edildi. Simülasyon uygulamasında D-TBA ve D-YSA-TBAyöntemlerinin açıklayıcılıkları arasında anlamlı bir fark olduğu görülmüştür ($p<0.001$). DLBCL verisi ile yapılan uygulamada D-TBA yönteminin hatasının %36.78, D-YSA-TBA yönteminin ise RSF sonucu- %43 olduğu bulunmuştur. D-TBA yönteminin önem değeri diğer yöntemden daha büyük, hatası ise daha düşük çıkmıştır. Çok boyutluluk problemi yaşanan gen ekspresyon verilerinin analizinde D-TBA, D-YSA-TBA'ya göre daha iyi performans göstermiştir.

Anahtar Kelimeler: boyut indirgeme, yapay sinir ağları, denetimli temel bileşenler analizi, random survival forests, gen ekspresyon

Türe M , Ömürlü Kurt İ. 2018, Gen Ekspresyon Verilerinde Yapay Sinir Ağlarına Dayalı Denetimli Temel Bileşenler Analizi Yaklaşımı, *Osmangazi Journal of Medicine* 40(1):20-27 **Doi:** 10.20515/otd.371882

Geliş Tarihi / Received
27.12.2017

Kabul Tarihi / Accepted:
22.02.2018

Yazışma Adresi / Correspondence Address
Mevlüt Türe
mail: mevlutture@gmail.com

1. Introduction

Along with the evolving technology, the increase in data collection and data storage possibilities brought the problem of multi-dimensionality with it. It is difficult to analyze especially multidimensional gene expression data in which the number of variables is more than the number of units according to classic statistical methods. For this reason, the examination of multi-dimensional problems has become focus of interest of statistical researches nowadays. Due to the problems caused by theoretical structure of classic statistical methods, it is necessary to analyze gene expression data by making dimension reduction instead of examining them together.

Due to the elimination of dependency between variables and the reduction of the number of components, the classic principal component analysis (PCA) continues its popularity even today. However, classic PCA is not able to find enough solution for multi-dimensional data problems like gene expression data which involve survival time. For this purpose, supervised principal component analysis (S-PCA) was developed by Bair and Tibshirani (1). S-PCA uses only genes which are strongly correlated to dependent variable, instead of applying principal components by using all of the genes in the dataset. Since S-PCA consults to dependent variable for identifying subset of gene expression data, it is known as a supervised method (2) and uses Cox scores calculated based on the Cox proportional hazards model to identify genes that may be associated with survival. Beer et. al. (3) stated in their study that Cox scores was used to determine important genes and since iterative calculation of parameter estimations of each variable caused disadvantage, Wald scores could be used instead of Cox scores.

In our study, we will also consider the principal component analysis with supervised artificial neural networks (S-ANN-PCA) method proposed by Kramer (4) in multidimensional gene expression data, in which survival time is considered as an alternative to S-PCA. S-ANN-PCA is a flexible generalization of the classic PCA and based on artificial neural networks. (4-7).

Dong and McAvoy (8) compared PCA and ANN-PCA in terms of image compression and stated that ANN-PCA showed higher performance. Monahan (6) determined using a dataset which is about climate that ANN-PCA was more robust method after comparing the dimension reduction performances of PCA and ANN-PCA. Aktürk Hayat et al. (9) compared the performances of S-PCA which is used for dimension reduction and an alternatively proposed approach of nonlinear PCA using ANN performed by gene selection with survival tree. They reported that explained variance ratio of survival tree based on ANN-PCA was higher than S-PCA.

Albanis and Batchelor (10), in a dataset about assessment of long-term credit continuity, showed that dimension reduction performances of linear and non-linear PCA based on ANN were better than PCA. Hsieh (11) compared PCA and ANN-PCA using a dataset about Pasific Ocean surface temperature, and showed that S-ANN-PCA was better. Türe et. al. (12) compared PCA, generalized PCA, linear and non-linear PCA based on artificial neural networks in dimension reduction of questions measuring patient satisfaction, and found that explained variance ratio of S-ANN-PCA was higher than other methods.

This study is aimed in this study to develop supervised S-ANN-PCA in multidimensional gene dataset and compare to S-PCA. A simulation programme was built in R language for this purpose. After the simulation study, we applied the methods that we suggested on a real data set. In this study, an alternative new approach was proposed, which can determine the effect of genes on survival time of researchers working with genes.

2. Material and Methods

2.1. Principal Component Analysis with Supervised Artificial Neural Networks

The S-ANN-PCA is a generic feature extraction algorithm that includes features containing as much information as possible from the original data set. If there are non-linear correlations between the variables and there is sufficient data to supply the formulation between the more complex mapping functions, S-ANN-PCA performance becomes better than the data PCA (10). S-ANN-PCA method, presented by Kramer, based on auto associative neural networks that are trained by backpropagation (4). S-ANN-PCA means to reduce the dimension of the input variables and each layer completely connected to the next. It uses five-layer feed-forward network with a bottleneck layer of nodes in order to do this (4, 13). The activation functions of the second and fourth layers of the network are sigmoidal, so layers 1, 2, and 3 and layers 3, 4, and 5 model nonlinear functions while the activation functions of the third and fifth layers are linear. The input (first) and output (fifth) layers have p units (the number of variables in the data set). The number of the nodes in the third layer is fewer (m<p) than the first or fifth. The values of the output nodes in layer 5 are trained in order to approximate the inputs. After the network has been trained, bottleneck node activation values in layer 3 give a lower dimensional representation of the inputs (6, 14). The network aims to minimize the error term (e). It is then trained to try and reproduce the input pattern at the output layer by using an error term calculated as summation of squared difference between the network prediction (X_i) and input pattern (X'_i);

$$\mathrm{e} = \sum_{i=1}^{p} \left(\mathrm{X}_i - \mathrm{X}'_i\right)^2$$

(i=1,2,...,p).

S-ANN-PCA makes a fit a curve through the data and then reduces the dimension of the inputs. The first three layers of the network creates the projection of the original data onto the curve and the activation values of the bottleneck layer, called scores, give its location. The last three layers describe the curve (10, 11, 15, 16).

2.2. Supervised Principal Component Analysis

S-PCA is used for supervised dimensionality reduction, particularly in solution of regression problems with high dimensions. S-PCA is a method which is developed for the analysis of gene expression data. Since it counts on the values of dependent variable in the step of gene selection, principal components are predicted by selected subset of genes which are correlated to dependent variable (1, 2). Assume we have a set of n data points (2, 17) i=1 each consisting of p features, stacked in the p ×n matrix, and denote the jth feature (jth row of X) by X_j. The S-PCA procedure is summarized as follows:

Compute standard regression coefficients for each feature.

Reduce the data matrix X to contain only the features whose coefficients exceed a threshold in absolute value. Calculate the first few principal components of the reduced data matrix. Use the principal components which are calculated in step 3 in a regression model or a classification algorithm to predict the outcome.

2.3. Wald test

The Wald test is used to compute from the ratio of the estimated model coefficient to its estimated standard error for significance of the coefficient:

$$W_j = \frac{\hat{\beta}_j}{\overset{\wedge}{SH}\left(\hat{\beta}_j\right)} \quad \text{(j=1,2,...,p)}$$

The square of the Wald statistic follows a chi-square distribution with one degree-of-freedom (17-19).

2.4. Random Survival Forests

RSF is used for the analysis of right censored survival data and is a survival based on tree method. It is based on a splitting rule and bootstrap samples. In RSF, randomization is done by two steps. In the first step, it uses a randomly drawn bootstrap sample of the data to grow the tree. In the second step, the tree learner is grown by splitting nodes on randomly selected predictors. While at first sight Random Forest might seem an unusual procedure, remarkable empirical evidence has shown it to be very effective. In standard analyses, there are always some limitations on assumptions such as proportional hazards. Also, with such methods there is always the concern whether associations between predictors and hazards have been modelled appropriately, and whether or not non-linear effects or higher order interactions for predictors should be included. These problems are overcome smoothly and automatically in the RSF (20, 21). We use the log-rank splitting rule for the survival splitting. The log-rank splitting rule for a split at the value c for predictor x is

$$L(x,c)=\frac{\sum_{i=1}^{N}\left(d_{i,1}-Y_{i,1}\frac{d_i}{Y_i}\right)}{\sqrt{\sum_{i=1}^{N}\frac{Y_{i,1}}{Y_i}\left(1-\frac{Y_{i,1}}{Y_i}\right)\left(\frac{Y_i-d_i}{Y_i-1}\right)d_i}}$$

where n is the number of individuals within h node, $d_{i,j}$ is the number of deaths at time t_i in the daughter nodes j=1,2, $Y_{i,j}$ is the number of individuals at risk at time t_i in the daughter nodes and $t_1 < t_2 < \cdots < t_N$ is the ordered survival time. The value $|L(x,c)|$ is the measure of node separation. The fact that getting larger of $|L(x,c)|$ brings about larger value for $|L(x,c)|$, greater difference between the two groups and much better split (20).

2.5 Simulation

In this study, codes were written in R package program using *superpc*, *rpart*, *pcaMethods*, *survival*, *random Survival Forest* and *stats* packages.

In the simulation application, the following steps were carried out respectively:

1. According to multivariate normal distribution and different correlation levels in which correlation coefficient varied between r=-1 and(r=+1, 5000 genes for 100 units and survival time and status variables correlated to this gene dataset were generated. Our simulation algorithm was performed based on algorithm created by Bender, Augustin, & Blettner 2005 (22).

2. Two random sets of 70% training and 30% validity were created.

3. A selection method based on the Wald statistic was used to determine the subset of important genes for taining set.

4. Dimension reduction was done by S-PCA and S-ANN-PCA methods.

5. The dimension reduction performances the methods were compared according to the total explained variance ratios of the components obtained by methods.

6. 1000 repetitive-simulation application were performed in evaluating the performances of the methods.

2.6. Real Dataset Application

In application, 7399 gene expression data were used for 240 individuals from the study of Rosenwald et al. (23) in patients with diffuse large B-cell lymphoma (DLBCL- diffuse large B-cell lymphoma) (23). The dataset contains 7399 gene expression value, survival time (year) of the patients. Of the 240 patients; 138 (57.5%) are dead, 102 (42.5%) are censored observations. Two random sets of 70% training and 30% validity were created in DLBCL dataset.

S-PCA and S-ANN-PCA were used in dimension reduction step. In the S-PCA and S-ANN-PCA, components were obtained by generating a reduced size data matrix based on the Wald statistic.

3. Results

3.1. Simulation

In the simulation application, S-ANN-PCA and S-PCA methods were compared using program codes. Simulation was carried out with 1000 repetitions.

After 1000-times repetitive simulation, median explanatory rates of the methods were obtained.

As a result of the simulations, the median (25-75th percentiles) recorded for the test set was 0.119 (0.084-0.167) for S-PCA and 0.056 (0.031-0.092) for S-ANN-PCA. As a result of statistical comparison by Mann Whitney U test of the two methods, it was obtained that the total explained variance ratios of S-PCA was significantly different from S-ANN-PCA (p<0.001) (Figure 1).

It was found that the error rate for the S-PCA was 28.82% and 35.4% for the S-ANN-PCA as a result of RSF. The importance value of S-PCA method was found to be higher and its error rate was found to be lower than the other method.

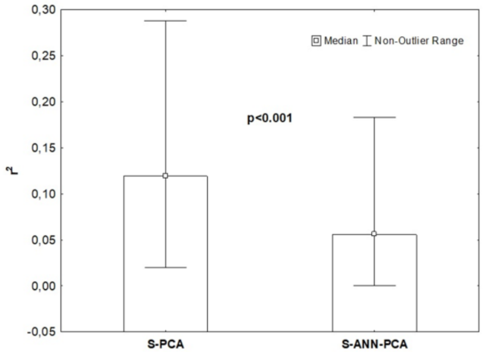

Figure 1. *Explanatory rates of S-PCA and S-ANN-PCA for test set*

3.2. DLBCL Data

The Wald statistic values for the 7399 gene were calculated using the training set for 168 patients. 134 genes were determined as statistically significant according to Wald statistic and reduced data matrix was obtained.

After the singular value decomposition of this matrix, 3 supervised principal components (S-PCA-1, S-PCA-2, S-PCA-3) for S-PCA method were obtained. Besides, gene expression values for 134 important genes were taken as input variables in S-ANN-PCA and then S-ANN-PCA was applied to these data. Three components (S-ANN-PCA-1, S-ANN-PCA-2 and S-ANN-PCA-3) were attained as a result of the analysis.

Prediction performances of S-PCA and S-ANN-PCA methods were compared by examining the effects of genes on survival time by using the components after random survival forest method. Number of deaths and trees values were set to 138 and 1000 respectively; log-rank was used as splitting rule for test set. Importance values obtained for each component after RSF were given in Table 1.

It was found that the error rate for the S-PCA was 36.78% and 43% for the S-ANN-PCA as a result of RSF. The importance value of S-PCA method was found to be higher and its error rate was found to be lower than the other method.

Table 1.
Importance values be obtained by random survival forests method for test set
Importance Value

	S-PCA	S-ANN-PCA
Component 1	0.0505	0.0209
Component 2	0.0117	0.0036
Component 3	0.0017	0.0007

4. Discussion and Conclusion

Gene expression data are characterized by multidimensionality. Dimensional reduction and Cox regression analyzes are often used to develop a model for predicting survival times, taking into consideration the gene profile and the survival time of patients from such data (1, 17, 24-26). When the studies using the ANN method are examined in the analysis of gene expression data, Liu, B et al. (27) analyzed genes identified by three different gene selection methods with a ANN containing a hidden layer. O’Neil and Song (28) analyzed lenfoma data by three layered ANN using sigmoidal activation function. The study of Khan et.al. (29), contains the gene selection by filtration, dimension reduction of the genes filtered, and prediction using an ANN which has no hidden layer. Since there is no dependent variable in Khan et.al’s study (29), the correlation between genes and dependent variable couldn’t be examined. In our study, it is aimed to solve the multidimensionality problem of gene data by proposing a new dimension reduction approach in which S-ANN-PCA is used as an alternative to S-PCA and to compare the performances of S-PCA and S-ANN-PCA based on Wald statistic for determining the genes correlated to survival in gene expression data. When the approaches developed for the analysis of gene expression data are examined, Zhao et al. (26), using hierarchical clustering and S-PCA, identified genes that accurately predict renal cell carcinoma patients’ survival from their survival-associated gene expression profiling. Quackenbush (30), stated that PCA, when used with other classification techniques, is a strong technique on gene expression data analysis. Zhang et al. (31) reported that the methods used in the analysis of gene expression data are uncontrolled methods, such as clustering analysis, and that the primary goal of clustering is to collect genes with similar characteristics, which is a disadvantage in diagnosing diseases. Dudoit et al. (32) performed a discriminant analysis, nearest neighbor, and classification and regression trees on existing data sets in their work comparing performance of different separation methods for classifying tumors from gene expression data. In our work, an alternative approach to size reduction-based survival estimation was developed by using data from 7399 gene expressions from 240 individuals with common B-cell lymphoma from Rosenwald et.al. (23). In this approach, which determines important genes based on Wald’s statistic, components with explained variance ratio close to that of S-PCA were obtained. When RSF was applied for the test set using the principal components of S-PCA and S-ANN-PCA from the gene expression data subset, it was determined that the importance values of the components found to be important in both models were close to each other. In our simulation, it was seen that the explanatory power of the S-PCA method was significantly higher than that of S-ANN-PCA.

*This research was supported by Adnan Menderes University Research Fund.

Project Number: TPF- 12020

REFERENCES

1. Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. PLoS biology. 2004;2(4):e108.
2. Chen X, 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.
3. Beer DG, Kardia SL, Huang C-C, Giordano TJ, Levin AM, Misek DE, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nature medicine. 2002;8(8):816-24.
4. Kramer MA. Nonlinear principal component analysis using autoassociative neural networks. AIChE journal. 1991;37(2):233-43.
5. Hsieh WW. Machine learning methods in the environmental sciences: Neural networks and kernels: Cambridge university press; 2009.
6. Monahan AH. Nonlinear principal component analysis by neural networks: theory and application to the Lorenz system. Journal of Climate. 2000;13(4):821-35.
7. Scholz M, Fraunholz M, Selbig J. Nonlinear principal component analysis: neural network models and applications. Principal manifolds for data visualization and dimension reduction: Springer; 2008. p. 44-67.
8. Dong D, McAvoy TJ. Batch tracking via nonlinear principal component analysis. AIChE Journal. 1996;42(8):2199-208.
9. HAYAT EA, Mevlut T, SENOL S. An Alternative Dimension Reduction Approach to Supervised Principal Components Analysis in High Dimensional Survival Data. Turkiye Klinikleri Journal of Biostatistics. 2016;8(1):21-9.
10. Albanis G, Batchelor R, editors. Assessing the long-term credit standing using dimensionality reduction techniques based on neural networks—an alternative to overfitting. The proceedings of the SCI 99/ISAS 99 conference, Orlando, US; 1999.
11. Hsieh WW. Nonlinear principal component analysis by neural networks. Tellus A: Dynamic Meteorology and Oceanography. 2001;53(5):599-615.
12. Ture M, Kurt I, Akturk Z. Comparison of dimension reduction methods using patient satisfaction data. Expert Systems with Applications. 2007;32(2):422-6.
13. Oja E. Principal components, minor components, and linear neural networks. Neural networks. 1992;5(6):927-35.
14. Fotheringhame D, Baddeley R. Nonlinear principal components analysis of neuronal spike train data. Biological Cybernetics. 1997;77(4):283-8.
15. Daszykowski M, Walczak B, Massart D. A journey into low-dimensional spaces with autoassociative neural networks. Talanta. 2003;59(6):1095-105.
16. Michailidis G, de Leeuw J. Multilevel homogeneity analysis with differential weighting. Computational statistics & data analysis. 2000;32(3):411-42.
17. Zhang H, Yu C-Y, Singer B, Xiong M. Recursive partitioning for tumor classification with gene expression microarray data. Proceedings of the National Academy of Sciences. 2001;98(12):6730-5.
18. Bøvelstad HM, Nygård 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.
19. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology. 1983;148(3):839-43.
20. Breiman L. Random forests. Machine learning. 2001;45(1):5-32.
21. Haykin S. Neural Networks, a comprehensive foundation,2nd ed., Prentice Hall, 842 p. 1999.
22. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. Statistics in medicine. 2005;24(11):1713-23.
23. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. New England Journal of Medicine. 2002;346(25):1937-47.
24. Ishwaran H, Kogalur UB. Random survival forests for R. R News. 2007;7(2):25-31.
25. Nguyen TS, Rojo J. Dimension reduction of microarray data in the presence of a censored survival response: a simulation study. Statistical applications in genetics and molecular biology. 2009;8(1):1-38.
26. Van Wieringen WN, Kun D, Hampel R, Boulesteix A-L. Survival prediction using gene expression data: a review and comparison. Computational statistics & data analysis. 2009;53(5):1590-603.
27. Zhao H, Ljungberg B, Grankvist K, Rasmuson T, Tibshirani R, Brooks JD. Gene expression profiling predicts survival in conventional renal cell carcinoma. PLoS medicine. 2005;3(1):e13.
28. Liu B, Cui Q, Jiang T, Ma S. A combinational feature selection and ensemble neural network

method for classification of gene expression data. BMC bioinformatics. 2004;5(1):136.

29. O'Neill MC, Song L. Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect. BMC bioinformatics. 2003;4(1):13.
30. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature medicine. 2001;7(6):673-9.
31. Quackenbush J. Computational analysis of microarray data. Nature reviews genetics. 2001;2(6):418-27.
32. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American statistical association. 2002;97(457):77-87.

©Copyright 2018 by Osmangazi Tıp Dergisi - Available online at tip.ogu.edu.tr ©Telif Hakkı 2018 ESOGÜ Tıp Fakültesi - Makale metnine dergipark.gov.tr/otd web sayfasından ulaşılabilir.