The Korean Communications
in Statistics Vol. 11 No. 3, 2004
pp. 539-551

Multivariate Decision Tree for High-dimensional Response Vector with Its Application

Seong Keon Lee[1]

Abstract

Multiple responses are often observed in many application fields, such as customer's time-of-day pattern for using internet. Some decision trees for multiple responses have been constructed by many researchers. However, if the response is a high-dimensional vector that can be thought of as a discretized function, then fitting a multivariate decision tree may be unsuccessful. Yu and Lambert (1999) suggested spline tree and principal component tree to analyze high dimensional response vector by using dimension reduction techniques. In this paper, we shall propose *factor tree* which would be more interpretable and competitive. Furthermore, using Korean internet company data, we will analyze time-of-day patterns for internet user.

Keywords : Factor tree, High-dimensional response, Multivariate decision tree, Principal component tree, Spline tree.

1 Introduction

Decision tree, one of many data mining techniques, is a popular approach for segmentation, classification and prediction by applying series of simple rules. It has an advantage that researchers can easily understand and explain the results because it is expressed by a tree structured diagram as a final output.

The landmark work of decision tree is the methodology of Breiman, Friedman, Olshen, and Stone (1984), who introduced classification tree for a univariate discrete/continuous response. There are various competing approaches to the work of Breiman et al. (1984), such as that of Hawkins and Kass (1982) and Quinlan (1992). These approaches are focused on the single response.

Recently, some decision trees for multiple responses have been constructed by Segal (1992) and Zhang (1998). Segal (1992) suggested a tree that can analyze continuous longitudinal

[1] Department of Mathematics, Chuo University, 1-13-27, Kasuga, Bunkyo-Ku, Tokyo, 112-8551, Japan.
E-mail : sklee@grad.math.chuo-u.ac.jp

response using Mahalanobis distance for within node homogeneity measures. Zhang (1998) suggested a tree that can analyze multiple binary response using generalized entropy criterion which is proportional to maximum likelihood of joint distribution of multiple binary responses (Cox, 1972; Zhao and Prentice, 1990).

Furthermore, in real world application, responses which have many variables can be often observed such as functional data. However, naively applying multivariate decision trees to "long vector responses" is not successful. The examples of multivariate decision trees that give unreasonable results to analyst are shown in the research of Yu and Lambert (1999). Yu and Lambert proposed new tree methodologies, spline tree and principal component tree for analyzing high dimensional response, applying two step procedures that reduce the dimension of the responses and then constructing a tree to lower dimensional responses. Spline tree represents each response vector as a linear combination of spline basis functions and then fits a multivariate tree to the estimated coefficient vectors. Principal component tree uses the first several principal component scores as the response vector.

In this paper, factor analysis will be used to reduce the high dimension responses to low dimensions that have several independent explainable factors by using factor rotation. In next section, we shall introduce and review the two step multivariate decision trees, i.e., spline tree and principal component tree. Then, we shall propose a factor tree that has advantages in the view of interpretation. Finally, using a Korean internet company data set which consists of internet site member's demographic profiles and hourly internet using pattern, we will investigate and compare the performance and results of the two step tree procedures, i.e., spline tree, principal component tree, factor tree.

2 Spline tree, Principal component tree and Factor tree

Sometimes, in particular functional response data, it is reasonable to treat high dimensional response vector as a curve. For these data, Yu and Lambert (1999) suggested the spline tree and principal component tree. Mahalanobis distance was used as an impurity measure in the node.

2.1 Spline tree

Yu and Lambert denoted that the response for subject i by $Y_i(t)$, where $t=(t_1,\cdots,t_m)$ are the midpoints of the time intervals. If $Y(t)$ is smooth, then it can be approximated by a linear combination of basis functions $\{\beta_1,\cdots,\beta_q\}$, and the coefficients of the linear combination for each individual can be used as the response for a multivariate tree. If a roughness penalty is imposed on the approximation, then each response is approximated by only a few basis functions, and the response vector is low dimensional. Generally, the lower the dimension of the response vector, the faster the multivariate trees can be fit.

2.2 Principal component tree

Instead of reducing the dimension of the response by treating it as a curve, we now reduce the dimension by treating it as a vector and applying principal component analysis, retaining only the first several principal components. That is, take

$$\delta_i = \sum_{j=1}^{m} \beta_j Y_i(t_j),$$

where β_j is the weighting coefficient, and the principal component scores δ_i are the un-correlated linear combinations $Y(t)$ of the response with variances that are as large as possible. See Anderson (1984) for more details.

After reducing the dimension, a multivariate decision tree will be constructed using principal component score as responses.

2.3 Factor tree

We would propose a factor tree by using factor analysis as dimension reduction method. As similar as constructing a principal component tree, we can construct a tree by considering factor scores as responses. Principal component analysis is focused on the maximum variance and maximum simultaneous resemblance motivations. In contrast factor analysis variables are assembled from two major components, common "factors" and unique "factors".

$$X = m + Lf + u,$$

where X is a matrix of data, m is the (vector) mean of the variables, L is a $p \times k$ matrix of factor loadings, f and u are random vectors representing the underlying common and unique factors.

The practical difference between principal component and factor analysis lies mainly in the decision whether or not rotating the principal components to emphasize the "simple structure" of the component loadings to make easier interpretation. Therefore, factor tree could be proposed to improve the interpretation of other two-step tree results.

The procedures of two-step trees, i.e., spline tree, principal component tree and factor tree, are as follows.

i) Dimension reduction on high dimensional response vector

spline method : $Y_1, Y_2, \cdots, Y_m \rightarrow S_1, S_2, \cdots, S_q, q < m$

principal component method : $Y_1, Y_2, \cdots, Y_m \rightarrow P_1, P_2, \cdots, P_q, q < m$

factor method : $Y_1, Y_2, \cdots, Y_m \rightarrow F_1, F_2, \cdots, F_q, q < m$

ii) Find the best split using Mahalanobis distance criterion as impurity measure proposed by Segal (1992)

$$SS(t) = \sum (y_i - y(t))' \psi^{-1}(\theta_t)(y_i - y(t))$$

iii) Create sub-tree with pruning procedure proposed by Breiman (1984)

iv) Select the best sub-tree using cost-complexity proposed by Breiman (1984)

3 Application

In this section, an application with responses of high dimensional vector will be shown. A C program was used to construct the trees.

3.1 Data

The problem we face is to predict a customer's time-of-day internet usage pattern. For example, business customers who use internet in daytime are more frequent, while students use internet more frequently at night time. Therefore classification of the time of day patterns would give useful information, such as the optimal time bands of shaver/cosmetic banner advertisements in a web page. So, it is a good strategy to advertise goods in the internet by considering customer's profiles and their surfing time patterns. The data which is used in the application consist of internet usage records of some internet site composed of 771 members.

Table 1 and table 2 show the data profile that have high dimension response vectors. Response variables are partitioned by 30 minutes, so they consist of 48 time intervals as responses. Explanatory variables consist of 5 categorical and 6 continuous variables, such as gender, age, job, etc. As we can see, naively applying multivariate decision trees to "long vector responses" like table 1 may not be successful. So, dimension reduction techniques mentioned above should be used at first and then, we will construct and compare the tree results.

To construct efficient trees, the following conditions were set; the number of subject in parent node should be greater than 30, the reduction of diversity measure should be greater than 0.1 and the number of subject in child node should be greater than 5. Also, to get a reasonable number of terminal nodes, 10-fold cross validation and cost-complexity method suggested by Breiman (1984) were used.

<Table 1> Response variables

Variable	Description
y1	monthly # of visit to the Internet site between 0:00 a.m ~ 0:30 a.m.
y2	monthly # of visit to the Internet site between 0:30 a.m ~ 1:00 a.m.
⋮	⋮
y48	monthly # of visit to the Internet site between 11:30 p.m ~ 0:00 a.m.

<Table 2> Explanatory variables

Variable	Description
# of connecting internet/month	# of count of using internet per month
Gender	Male, Female
Birth Year	Year of birth
Job	Student, General office worker, Financial worker, government officer,⋯, etc
Marital status	Married, Unmarried
Salary/month	0 Won ~ 10 million Won
Education	Under middle school, Under high school, Under college, Graduate school
Connecting Place	Usual place of using internet : Home, Company, Internet cafe, School, etc
Period of using internet	under 1 year, 1 ~ 2 years, 2~3years, 3~ 4 years, 4~ 5 years, over 5 years
Surfing time of internet/week	Under 0.5 hour, 0.5~1 hour, 1~2 hours, 2~3 hours, ⋯, over 50 hours
City	Living province : Seoul, Incheon, Busan, Daegu, ⋯, Jeju

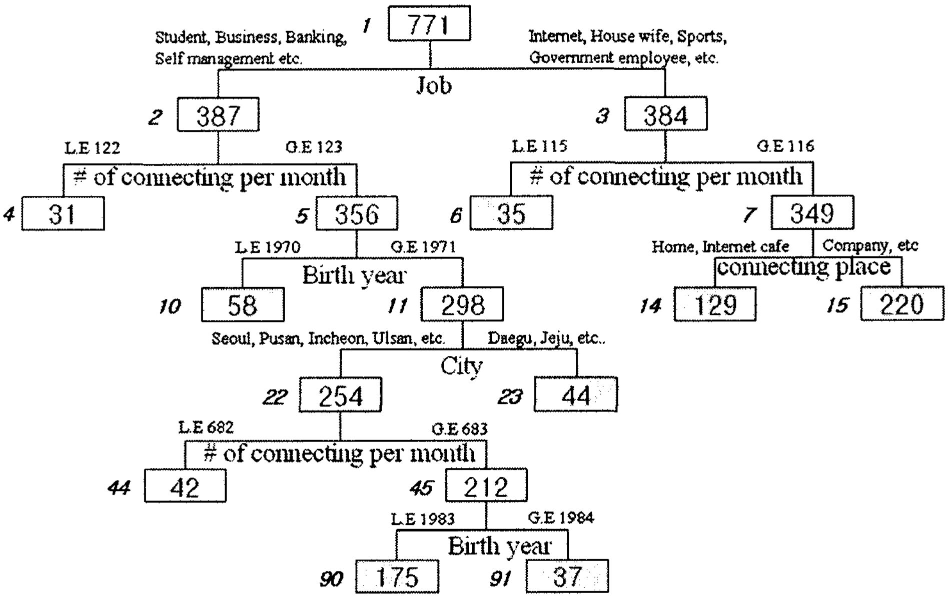
3.2 Spline Tree

The knots for B-spline can be placed on a uniform grid or another fixed grid based on information about the behavior of the response curve. In this application, most customers do not use internet between midnight and 9 a.m., and start to use internet after 9 a.m., so we assign a knot at 9 a.m. and no knots before 9 a.m. in our data. And we set the degree of the piecewise polynomial fit of base function to 3 and the number of derivatives evaluated for the roughness penalty to 2. Then the estimated coefficients of the base function are used as responses in a multivariate decision tree.

As a result, we get the tree structure which has 54 terminal nodes and through the pruning procedure using the cost-complexity measure we get the 43 sub-trees. Finally we could choose the tree, as in figure 1, which has 9 terminal nodes by applying 10-fold cross validation. The average internet using-time profiles at terminal node are shown in figure 2.

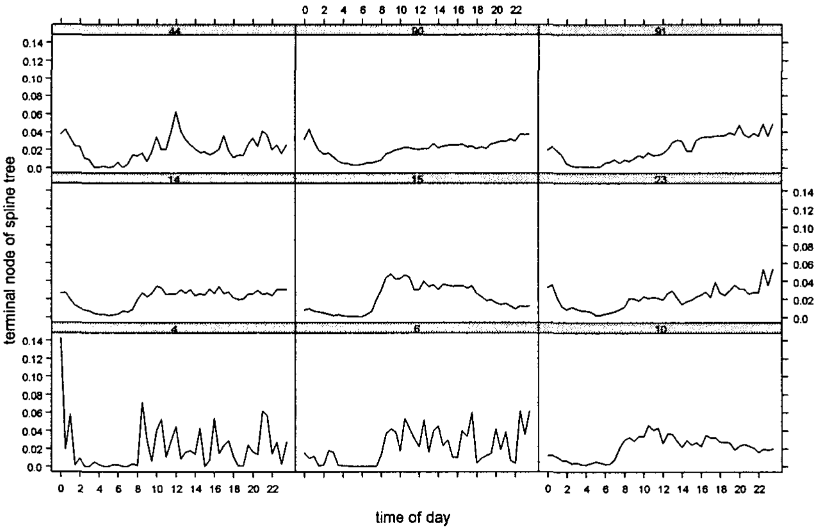
3.3 Principal Component Tree

In principal component tree, we have to consider what the proper number of principal components are. In this application, we take eight principal components by considering the proportion of explanation and the size of eigenvalue. We choose the number of principal components to retain by using scree plot. The first eight principal components together explain 45% of the total variance. Figure 3 gives the first eight principal component loadings of the time-of-day fractions for our data.

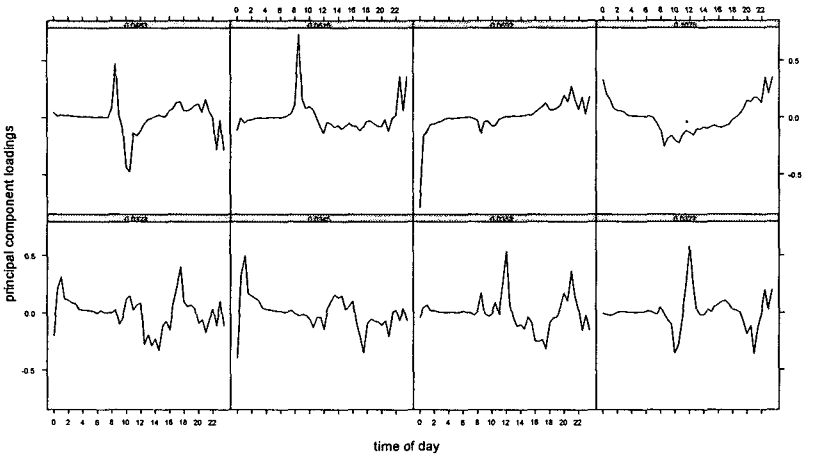


<Figure 1> Constructing Spline Tree

The number written in the node express the number of subjects and the number appeared besides of node is the node identification number.



<Figure 2> Average internet using-time profiles at terminal node of Spline tree: The number written at the top of each profile express the node identification number.



<Figure 3> Principal component loading

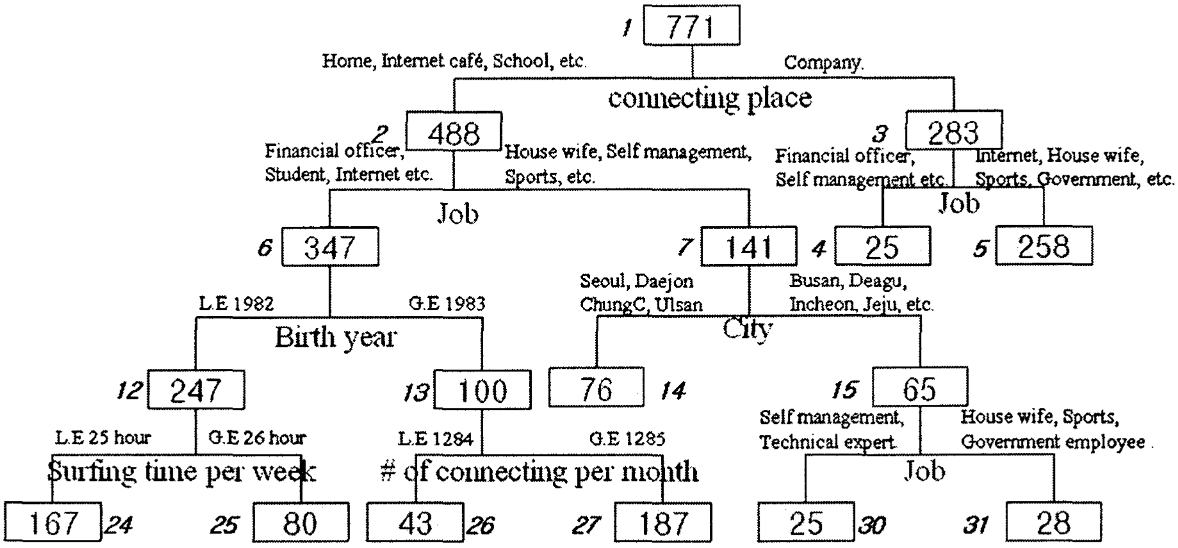
The number written at the top of each principal component profile expresses the proportion of explanation. The first principal component is the top of right side one.

Figure 4 shows the tree fit to the first eight principal component scores with the time-of-day fractions as responses. Seeing figure 4 and figure 1, we can find that the figures have similar results and shapes but have some differences on the split variables. Figure 1, spline tree, suggests that "job"is the first split factor for time of day using the pattern of internet. But figure 4, principal component tree, suggests that "connecting place" is the first split factor. And "internet using time", which is not the split variable of spline tree, is the spit variable of principal component tree.

Figure 4 expresses the tree diagram of principal component tree and figure 5 expresses the leaf node profiles. Interpreting a part of results, for example, we can see some facts from figure 4 that the customers in terminal node 5 are customers who are connect the internet at their work place and whose "job" is one of house wife, internet, sports, or government officer, etc. Furthermore, from figure 5, we can see also that the customers in terminal node 5 usually use internet during 8 a.m. ~ 6 p.m., i.e. their working time. But, the customers in terminal node 4 usually use internet all day long from 8 a.m.

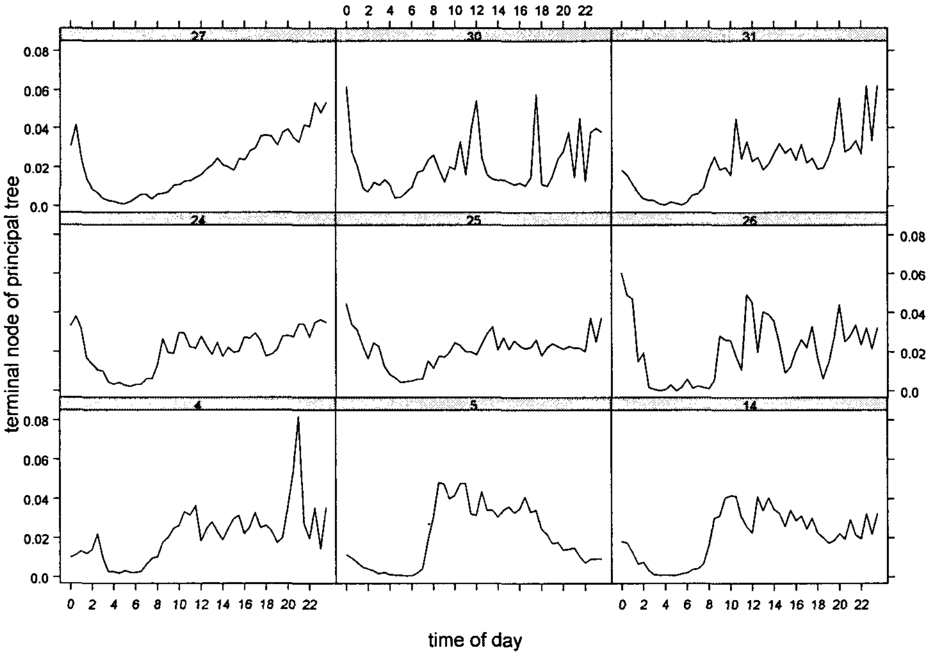
3.4 Factor Tree

We use iterative principal factor analysis to reduce the dimension of responses. As a result, we get eight factors that easily interpret the meaning by using varimax rotation method. We choose the number of factors same as the number of principal components of Section 3.3 in order to compare with the principal component tree in same dimensions of responses. Full tree has 45 terminal nodes and through the pruning procedure, the tree that has 9 terminal nodes is selected as figure 7.


1 771
Home, Internet café, School, etc.
Company.
connecting place
2 488
3 283
Financial officer, Student, Internet etc.
House wife, Self management, Sports, etc.
Financial officer, Self management etc.
Internet, House wife, Sports, Government, etc.
Job
Job
6 347
7 141
4 25
5 258
Seoul, Daejon ChungC, Ulsan
Busan, Deagu, Incheon, Jeju, etc.
City
L.E 1982
G.E 1983
Birth year
12 247
13 100
76 14
15 65
L.E 25 hour
G.E 26 hour
Surfing time per week
L.E 1284
G.E 1285
# of connecting per month
Self management, Technical expert
House wife, Sports, Government employee
Job
167 24
25 80
43 26
27 187
25 30
31 28

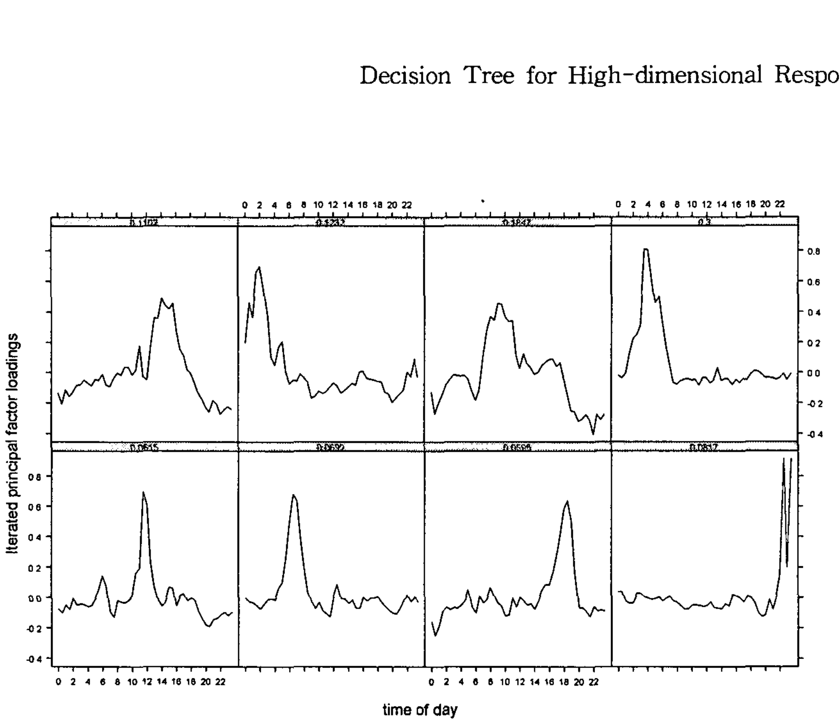
<Figure 4> Constructing Principal Component Tree

The number written in the node express the number of subjects and the number appeared besides of node is the node identification number.



<Figure 5> Average internet using-time profiles at terminal node of Principal component tree

The number written at the top of each profile express the node identification number.



<Figure 6> Iterative principal factor loading (first eight factors)

The number written at the top of each factor profile expresses the proportion of explanation. The top of right side one expresses the first factor profile.

<Table 3> Interpretation of Factors

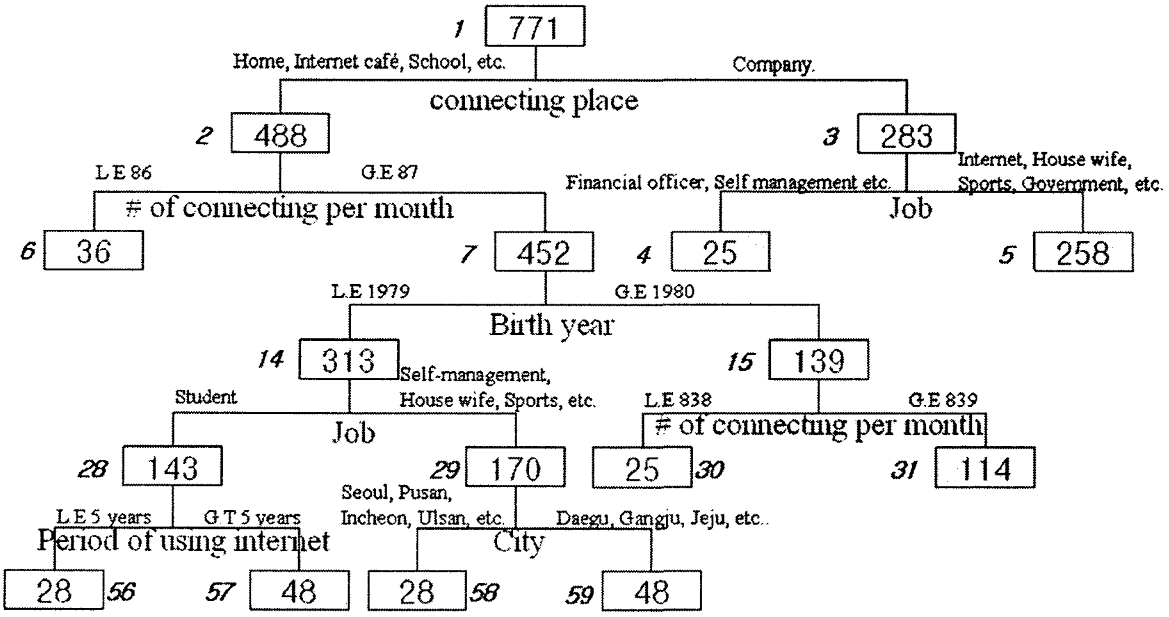
Factor number	Factor loading time	(%)
1	2 a.m. ~ 7 a.m.	30.00
2	7 a.m. ~ 11 a.m., 6 a.m.~2 a.m.	18.47
3	0 a.m. ~ 4 a.m.	12.32
4	1 p.m. ~ 5 p.m., 8 p.m ~ 1 a.m.	11.02
5	10 p.m ~ 12 p.m.	8.17
6	5 p.m. ~ 8 p.m.	6.95
7	5 a.m. ~ 8 a.m.	6.92
8	11 a.m. ~ 1 p.m.	6.15

Seeing the factor loading, in figure 6, factor profiles are more easily interpreted than principal component in figure 3. Figure 6 represents that each factor has from 4 hour to 6 hour intervals which have higher factor loading. For example, the first factor, right upper side in figure 6, has an interval which has higher loading between 2 a.m. and 7 a.m. So, the first factor could distinguish the customers who usually use internet during midnight.

Factor tree has more similar tree results to that of principal component tree than that of spline tree. The first split variable at the root node is also "connecting place", i.e., same as that of principal component tree. Many split variables which are chosen in principal tree are also chosen in factor tree, but they appear in different level of tree depth. For example "connecting frequency" and "birth year"are selected at upper level of depth in the factor tree, but the principal component tree suggests that "job", "city" and "birth year" are split variables at upper level of depth.

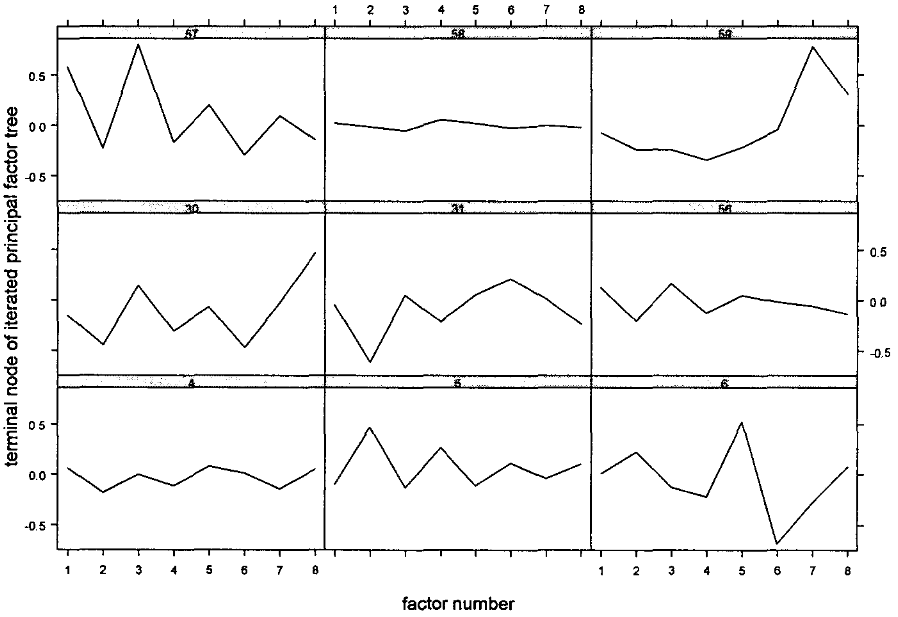
Figure 8 represents the factor mean score at each terminal node of the factor tree. It is difficult to explain the meaning of principal component and spline coefficients, but factor can be easily explained by factor rotation. Therefore, using the factor mean score at each terminal node is helpful for understanding results.

Figure 7, factor tree, suggests that customers in terminal node 56 and 58, who were born before 1980 and use internet more than the average using time, would connect to the internet similarly with the pattern of average of over-all customers. This is because most of the factor means in the terminal nodes are close to 0. Therefore they usually use internet at daytime and evening. Interpreting a part of results in detail, customers in node 59 have similar profiles as node 56 and 58, but they usually use internet during 5 a.m. ~ 7 a.m. and 11 a.m. ~ 1 p.m., since the scores of factor 7 and 8 are high. Customers in node 57, who are more than 23 years old and whose period of using internet is over 5 years, usually use internet from midnight to early morning since the score of factor 1 and factor 3 which have high factor loadings at the time interval 2 a.m. ~ 7 a.m. and 0 a.m. ~ 4 a.m. are high. And customers in node 5, who use internet at their office and are not financial officer and self management, usually use internet during daytime since the score of factor 2 and factor 4 are high.


1 771
Home, Internet café, School, etc.
Company.
connecting place
2 488
3 283
L.E 86
G.E 87
Financial officer, Self management etc.
Internet, House wife, Sports, Government, etc.
Job
# of connecting per month
6 36
7 452
4 25
5 258
L.E 1979
G.E 1980
Birth year
14 313
Self-management, House wife, Sports, etc.
15 139
Student
L.E 838
G.E 839
Job
# of connecting per month
28 143
29 170
25 30
31 114
Seoul, Pusan, Incheon, Ulsan, etc.
Daegu, Gangju, Jeju, etc.
L.E 5 years
G.T 5 years
Period of using internet
City
28 56
57 48
28 58
59 48

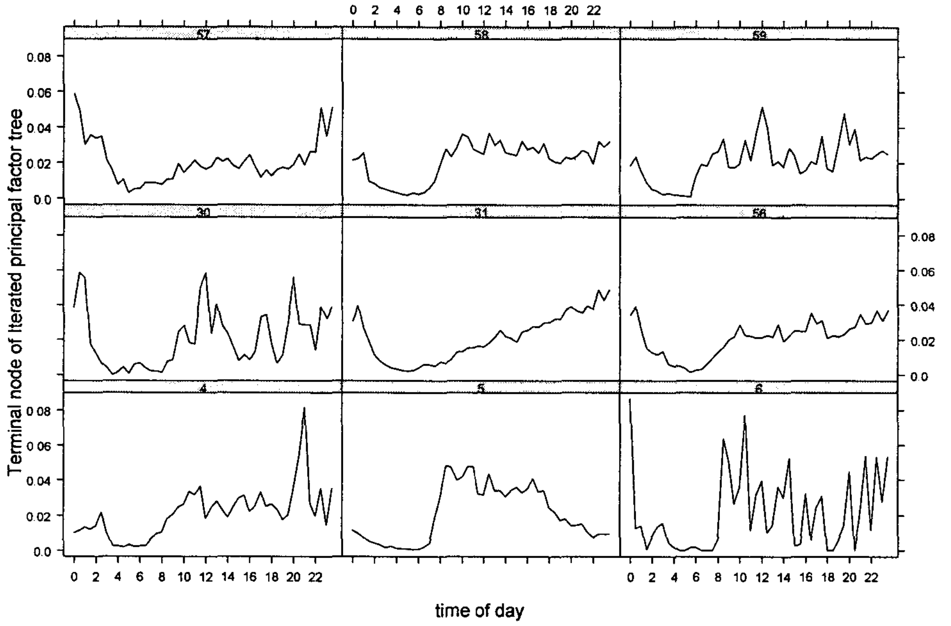
<Figure 7> Constructing Factor Tree

The number written in the node express the number of subjects and the number appeared besides of node is the node identification number.



<Figure 8> mean factor score at terminal node

The number written at the top of each profile express the node identification number.



<Figure 9> Average internet using-time profiles at terminal node of Factor tree

The number written at the top of each profile express the node identification number.

<Table 4> Sum of distances in Terminal nodes of Each Trees

	Mahalanobis distance	Euclidean distance
Spline tree	982794.38	30054.76
Principal component tree	1039586.34	39660.56
Factor tree	1040449.49	36447.42

Finally, Mahalanobis distance and Euclidean distance are used to compare the efficiency of tree methods. Table 4 shows the distances between customers at terminal nodes. Spline tree has the smallest distance compared to other two trees. The distance of principal component tree and factor tree are almost same. But the factor tree has an advantage that a factor can be easily interpreted. Therefore, we can say that the suggested factor tree is competitive.

4 Conclusion and discussion

Multivariate decision trees using spline method and dimension reduction techniques, such as principal component analysis and factor analysis to analyze high dimensional responses were introduced and suggested in previous sections. Also, through the analysis of application data, we could find little difference in their tree structure outputs. Principal component tree and factor tree had similar split variables and tree shape, but not spline tree. It may be because dimension reduction methods of principal analysis and factor analysis are similar.

In the view of homogeneity at terminal nodes, spline tree is the best. It may be because spline tree fits the model using non-linear function at each interval. But in the view of interpretability, factor tree is the best. Also, with regards to the homogeneity, factor tree is not the worst case in these three methods. So, we can say factor tree is competitive but it is hard to discriminate on superiority. Though continuous responses were only considered in this paper, studies on tree for high dimensional discrete responses can be topics for further study.

REFERENCES

[1] Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1984), Classification and Regression Trees, Wadsworth, CA.

[2] Green, P. J., Silverman, B. W. (1999), Nonparametric Regression and generalized Linear Models, Chapman and Hall, London.

[3] Johnson, R. A., Wichern, D. W. (1998), Applied Multivariate Statistical Analysis, Prentice-Hall, London.

[4] Segal, M. R. (1992), Tree-Structured Methods for Longitudinal Data, Journal of the American Statistical Association, 87, 407-418.

[5] Shikin, E. V., Plis, A. I. (1995), Handbook on Splines for the user, CRC Press, FL.

[6] Venables, W. N., Ripley, B. D (1999), Modern Applied Statistics with S-PLUS, Springer-Verlag, New York.

[7] Yu Y., Lambert, D. (1999), Fitting Trees to Functional Data, With an Application to Time-of-Day Patterns, Journal of Computational and graphical Statistics, 8, 749-762.

[8] Zhang, H. (1998). Classification Trees for Multiple Binary Responses, Journal of the American Statistical Association, 93, 180-193.

[9] Zhang, H., Singer, B. (1999), Recursive Partitioning in the Health Sciences, Springer-Verlag, New York.

[Received July 2004, Accepted September 2004]