Tree-structured Partial Least Squares with an Application to Othodontics Data

Soo-Heang Eo[a] , Shin-Jae Lee[b] , Sungwan Bang[c] * and HyungJun Cho[a] †‡

[a] Department of Statistics, Korea University, Seoul, South Korea
[b] Department of Orthodontics, Seoul National University School of Dentistry, Seoul, South Korea
[c] Department of Mathematics, Korea Military Academy, Seoul, South Korea

Keywords: PLS, Model-based recursive partitioning, GUIDE, Orthodontics data

Introduction

Partial least squares (PLS) regression is an alternative to classical regression for handling multicollinearity by combining the merits of principal component analysis and multiple linear regression. The aim of PLS is to predict a set of responses from a set of predictors by finding latent variables that capture the variability in both responses and predictors. Latent variables used in PLS are defined as the linear combinations of original variables. Among predictors, some variables can come from an extremely different nature or some can be indirectly related to responses. In that case, taking all predictors into account for extracting latent variables is not suitable from the model selection point of view. More latent variables can be selected for model fitting. If the variable from other natures has a different role rather then a predictor in PLS, it can be possible to improve the prediction performance by using the less number of latent variables. A disadvantage of PLS would be the deficiency in data visualization and model interpretation among others, which may be as important as building an optimal predictive model.

A recursive partitioning algorithm emerges as one of the solutions capable of achieving these purposes. We here propose new piecewise regression by combining the model-based recursive partitioning and PLS in order to improve prediction performance, and provide a visually interpretable model. Recursive partitioning, also known as tree-structured modeling, has been widely used because they allow us to give easy data visualization and interpretation. The objective of our proposed method is to select the most relevant predictors recursively, and provide more accurate prediction performance. Instead of fitting a global PLS to the whole data, one might fit a collection of local PLS to subsets of the data so that a better fit and higher predictive accuracy are obtained. We alleviate variable selection bias using the merits of the residual analysis approach of [1] and conditional inference of [2]. Our developed software program can be obtained from the authors upon request.

Proposed Method

Let $(\mathbf{Y},\mathbf{X},\mathbf{Z})$ be matrices of the response, predictor, and partitioning variables where $\mathbf{Y}$ and $\mathbf{X}$ are the q- and p-dimensional spaces, and $\mathbf{Z}=(\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_r)$ with r-dimensional space, respectively. As a basic model, we employ

*Corresponding author. E-mail: wan1365@hanmail.net
†Corresponding author. E-mail: hj4cho@korea.ac.kr.
‡This research was supported by the National Research Foundation of Korea (NRF) funded by the Korean Government (2012-0007545) for S.-J. Lee, funded by the Ministry of Science, ICT and Future Planning (NRF-2013R1A1A1007536) for S. Bang, and funded by the Ministry of Education, Science and Technology (2010-0007936) for H. Cho.

the following PLS regression

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\mathbf{P}^T + \mathbf{E}, \\ \mathbf{Y} &= \mathbf{U}\mathbf{Q}^T + \mathbf{R}, \end{aligned} \tag{1}$$

where T and U are $n \times K$ matrices that are projections of $\mathbf{X}$ and $\mathbf{Y}$, *i.e.*, the score matrices of $\mathbf{X}$ and $\mathbf{Y}$. The $\mathbf{P}$ and $\mathbf{Q}$ are $p \times K$ and $q \times K$ orthogonal loading matrices, respectively. It is assumed that the error matrices $\mathbf{E}$ and $\mathbf{R}$ are independent and identically distributed random normal. Simply, the NIPALS algorithm is used to obtain parameters [3]. The basic model is fitted to have estimates that are different at each node so that terminal nodes might have different estimates. As an impurity function at node t, $R(t)$, mean squared errors are used.

For the partitioning, the partitioning variables $\mathbf{Z}$ are used to divide the data set into two subregions in which the basic model is fitted. Split rules are utilized to find the best split variable and point by evaluating the impurities between parent and child nodes over all possible splits. We here propose new split rule algorithms to provide the unbiased variable selection and visually interpretable model. The approach selects split rules by investigating the randomness of residuals after fitting the basic model to the data at each node, resulting in fast and ignorable-bias selection of split rules [1, 2]. This selection rule consists of univariate and bivariate algorithms. For the determination of tree size, the permutation test with multi-step stopping rule [4] is utilized.

Results and Conclusions

We applied the proposed method and the standard PLS regression to the craniofacial X-ray data set used in [5, 6]. The data sample consists of a lateral cephalogram, an X-ray image of craniofacial area on which the variables are gathered by the two-dimensional coordinates. The response variable $\mathbf{Y}$ is the coordinates on the post-surgery facial landmarks. The coordinates on the skeletal structure before and after surgery and pre-existing facial landmarks are used for predictors $\mathbf{X}$. Moreover, demographic variables, such as age, gender, interval time after surgery, operation type, and so on are used to partitioning variables $\mathbf{Z}$.

For comparison, the mean squared error of prediction (MSEP) is calculated by 10-fold cross validation. The overall MSEP of the proposed method is 3.70 with standard deviation 0.82, compared to those of the standard PLS are 4.51 (mean) and 1.02 (SD). When it comes to the errors of each response, our proposed method also outperforms the standard PLS. Our proposed method provides easy interpretation of the interrelation between terminal nodes by data visualization since the structure of the tree has a biplot per each terminal node. Throughout data analysis, we see that our proposed method attains better performance in terms of the prediction and the data visualization compared to the standard PLS.

References

[1] W.-Y. Loh and W. Zheng, "Regression trees for longitudinal and multiresponse data," *The Annals of Applied Statistics* **7**(1), pp. 495–522, 2013.

[2] T. Hothorn, K. Hornik, and A. Zeileis, "Unbiased recursive partitioning: A conditional inference framework," *Journal of Computational and Graphical Statistics* **15**(3), 2006.

[3] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems* **2**(1), pp. 37–52, 1987.

[4] S.-H. Eo and H. Cho, "Tree-structured mixed-effects regression modeling for longitudinal data," *To appear in Journal of Computational and Graphical Statistics* , 2013.

[5] H.-Y. Suh, S.-J. Lee, Y.-S. Lee, R. E. Donatelli, T. T. Wheeler, S.-H. Kim, S.-H. Eo, and B.-M. Seo, "A more accurate method of predicting soft tissue changes after mandibular setback surgery," *Journal of Oral and Maxillofacial Surgery* **70**(10), pp. e553–e562, 2012.

[6] H.-J. Lee, H.-Y. Suh, Y.-S. Lee, S.-J. Lee, R. E. Donatelli, C. Dolce, and T. T. Wheeler, "A better statistical method of predicting postsurgery soft tissue response in class II patients," *To appear in Angle Orthodontist* , 2013.