IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 February 2024, accepted 27 February 2024, date of publication 5 March 2024, date of current version 13 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3373895

RESEARCH ARTICLE

Partial Least Squares Regression Trees for Multivariate Response Data With Multicollinear Predictors

WENXING YU[1], **SHIN-JAE LEE**[2,3], **AND HYUNGJUN CHO**[1]

[1]Department of Statistics, Korea University, Seoul 02841, Republic of Korea
[2]Department of Orthodontics, Seoul National University School of Dentistry, Seoul 03080, Republic of Korea
[3]Dental Research Institute, Seoul National University School of Dentistry, Seoul 03080, Republic of Korea

Corresponding author: Hyungjun Cho (hj4cho@korea.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by Korean Government (MSIT) under Grant 2022M3J6A1063595, and in part by Korea University under Grant K2301531.

ABSTRACT Some problems arise in analyzing massive complex data consisting of multivariate response variables and a large number of multicollinear predictor variables, especially when the sample sizes compared to the number of predictors are small. Rather than ordinary linear regression modeling approaches, latent variable regression modeling approaches such as partial least squares regression can be used to capture the relationship between the response and predictor variables for such cases. However, for complex nonlinear relationships between the predictor and the response variable, the performance of inference and prediction using regression modeling approaches can be deflated. Regression trees can capture such complex relationships. Thus, we develop a partial least squares tree modeling algorithm that detects complex relationships and makes precise predictions by integrating the merits of partial least squares and regression trees. It is shown that it has better predictive performance than other methods through simulation and it is demonstrated that it generates interpretable predictive models with real data of usedcar and orthognathic surgery.

INDEX TERMS Multicollinear, partial least squares, complex nonlinear relationships.

I. INTRODUCTION

In modern biomedical fields, the following data are often observed. Predictors consist of foundation information variables and multicollinear variables, whereas the number of response variables is large and multicollinear with predictor variables. For example, we consider orthognathic surgery case. The surgery aims to correct conditions of the jaw and face associated with orthodontic problems [1]. Various points were chosen to evaluate the shape of the patient before and after the surgery. The landmark value was measured as a two-dimensional coordinate value from the origin to the landmark. The predictors consisted of six external factor variables (age, gender, and so on) and 232 landmark variables composed of 168 skeletal landmarks before and after surgery, and 64 facial landmarks before surgery. The response variables consisted of 64 facial landmarks after surgery. We are attempting to construct an interpretable predictive model for facial changes before and after surgery by interpreting the effects of predictors. However, many landmark variables are highly correlated with each other, as well as many correlated response variables. Since the dimension of the response variables is large, Chew et al. [2] and Kneafse et al. [3] attempted to fit a multivariate linear regression model. However, it cannot work with multicollinearity and can only work if the number of targets exceeds the number of predictors. To solve this problem, Suh et al. [4] adopted a partial least squares (PLS) regression model. However, the performance of inference and prediction using regression modeling approaches can be deflated for

The associate editor coordinating the review of this manuscript and approving it for publication was Sajid Ali.

complex nonlinear relationships between the predictors and the response variables. In entrust, regression trees can capture such complex relationships.

Tree-structured models have advantages in terms of interpretation and data visualization. An advantage of a tree-structured model as a nonparametric method is that it helps to easily interpret the response effects of predictors, whereas it can provide suggestive insights through model interpretation and data visualization. Since the advancement of the automatic interaction detection algorithm by Morgan et al. [5] for univariate responses, decision trees have become very popular in various fields. Breiman et al. [6] proposed a classification and regression tree and the fast algorithm for classification tree by Loh et al. [7] had a powerful effect on the field of decision trees. It is because of pruning techniques and variable selection approaches. Additionally, It has relatively low computational costs and led to numerous subsequent studies. For example, the survival tree by Ahn et al. [8], the piecewise-polynomial regression trees by Chaudhuri et al. [9], the regression impurity tree by Alexander and Grimshaw [10], the Bayesian tree by Chipman et al. [11], the unbiased interaction selection and estimation by Kim and Loh [12], the mixed-effects longitudinal tree by Eo et al. [13], the Seemingly unrelated regression tree [14] and the Unified Noncrossing Multiple Quantile Regressions Tree [15] by Kim et al. The generalized, unbiased interaction detection and estimation by Loh [16]. More recently, there have been papers on modeling analysis based on regression trees approach in many fields. Such as the piecewise symbolic regression tree by Zhang et al. [17], the boosted regression tree (Knierim et al. [18], Said et al. [19], Han et al. [20], and Alnahit et al. [21]), the logistic regression trees by Loh [22], the bayesian additive regression tree (Pan and Bunn [23], Clark et al. [24], Um et al. [25]) and so on.

Our goal is to construct an interpretable predictive model for the data consisting of multivariate response variables and multiple multicollinear predictor variables. Thus, we combine the merits of regression tree modeling and PLS regression modeling, which detect complex relationships and make an accurate prediction. The idea has been tried in several papers. Yeh and Spiegelman [26] and Reddy et al. [27] sequentially fit PLS regression and regression tree, whereas Hao et al. [28] combined PLS regression and regression tree, and additionally combined PLS regression and random forests. All predictors were used for splitting and fitting in the latter, and only univariate response variables were considered in the studies. We can also use all predictor variables simultaneously for splitting and fitting. However, depending on the characteristics of the data, we divided the predictor variables into two parts. One is for fitting with high collinearity, and the other is for splitting with interpretation. If all predictors are used to construct latent variables, interpretation of predictors in PLS is difficult because the effects of external factor variables, which clinical researchers want to know can be hidden by complex latent structures. It is unsuitable to extract all predictors as latent variables because some external factor variables may be indirectly associated with the response variables. Therefore, we need a model that provides a direct interpretation of external factors concerning personal characteristics whereas maintaining the prediction of performance. For this, we develop a PLSRT (PLS Regression Trees) modeling algorithm by dividing predictors into two parts: external factors for splitting and landmark factors for fitting.

The remainder of this paper is structured as follows. Section II introduces the proposed PLSRT model. Section III presents the simulation studies. The performance of the proposed method is further compared with that of several other models. This is demonstrated using two real data sets in Section IV. Finally, Section V presents concluding remarks.

II. PROPOSED METHOD

This section describes a new model, PLSRT, with basic model settings, impurity functions, split rules, and tree size determination.

A. BASIC MODEL

Let $y_1, y_2, \ldots, y_m$ and $x_1, x_2, \ldots, x_{p+q}$ be the response and predictor variables with n subjects. We consider the multiple response general linear model for regressing $\boldsymbol{y}$ on $\boldsymbol{x}$, as follows:

$$\boldsymbol{y} = B_0 + \boldsymbol{x}B + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_m)$, $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{p+q})$, B_0 is the intercept term, $\mathbf{B}$ is the $(p+q) \times m$ matrix for the regression coefficients, $\boldsymbol{\epsilon}$ is an error term. Some of the predictor variables are correlated with each other. To construct PLSRT, we divide the $p+q$ predictor variables into two parts and assign different roles. For convenience, it is assumed that the first p variables are used for fitting and the others q variables are used for splitting. That is, let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ be fitting variables and $\boldsymbol{x}_{p+1}, \boldsymbol{x}_{p+2}, \ldots, \boldsymbol{x}_{p+q}$ be split variables. To distinguish easily, we use $z_1, z_2, \ldots, z_q$ instead of $\boldsymbol{x}_{p+1}, \boldsymbol{x}_{p+2}, \ldots, \boldsymbol{x}_{p+q}$.

We find appropriate estimates of the regression coefficients $\mathbf{B}$ to interpret and predict the response values $\boldsymbol{Y}$ of individuals at each partitioned node. To do this, we employ partial least squares (PLS) to solve the seriously correlated regression problem. For tree-based PLS regression modeling, a basic model with individuals at an arbitrary node t is defined as follows:

$$y_{ki} = \beta_{k0}^{(t)} + \sum_{j=1}^{p} \beta_{kj}^{(t)} x_{ji} + \epsilon_{ki}, \quad k = 1, 2, \ldots, m, \quad i \in t \tag{2}$$

where y_{ki}, x_{ji}, $\beta_{kj}^{(t)}$ and ϵ_{ki} are the elements of the response variables $\boldsymbol{y}$, predictor variables $\boldsymbol{x}$, regression coefficients $\mathbf{B}$ and error terms $\boldsymbol{\epsilon}$. The regression coefficients are estimated using PLS, with individuals at node t. Therefore, the response variables y_{ki}, $(k = 1, 2, \ldots, m, i \in t)$ are obtained from the PLS estimates $\hat{\beta}_{k0}^{(t)}, \hat{\beta}_{k1}^{(t)}, \ldots, \hat{\beta}_{kp}^{(t)}$ as follows:

$$\hat{y}_{ki} = \hat{\beta}_{k0}^{(t)} + \sum_{j=1}^{p} \hat{\beta}_{kj}^{(t)} x_{ji}, \quad k = 1, 2, \ldots, m, \quad i \in t \tag{3}$$

B. IMPURITY FUNCTION

An impurity function is a fundamental element in the construction of a tree. We consider the sum of squared errors as an impurity function at node t as follows:

$$i(t) = \sum_{k=1}^{m} \sum_{i \in t} (y_{ki} - \hat{y}_{ki})^2 = \sum_{k=1}^{m} \sum_{i \in t} (y_{ki} - \hat{\beta}_{k0}^{(t)} - \sum_{j=1}^{p} \hat{\beta}_{kj}^{(t)} x_{ji})^2 \tag{4}$$

The impurity function measures the prediction errors which are the differences between y_{ki} and $\hat{y}_{ki}$ at each node.

C. SPLIT RULE SELECTION

To find a good split, we can evaluate the following reduction of impurities for all possible splits.

$$\Delta(t) = i(t) - [i(t_L) + i(t_R)] \tag{5}$$

where $i(t)$, $i(t_L)$ and $i(t_R)$ are the impurities at node t, its left and right subnodes, t_L and t_R, respectively. Among all possible splits, the best split generates the greatest reduction in impurity. We refer this routine to as the exhaustive search (ES) approach. The ES approach has some problems, such as variable selection bias and considerable computational cost, as indicated by Loh [16], [29] in classification and regression tree problems. Therefore, we can use the residual analysis (RA) approach to solve these problems. In this paper, we apply the RA approach first to define the residuals at node t, as follows:

$$r_{ki} = y_{ki} - \hat{y}_{ki}, \quad k = 1, 2, \ldots, m, \quad i \in t \tag{6}$$

These are the differences between the response values y_{ki} and predicted values $\hat{y}_{ki}$. The signs of the residuals are formed as follows:

$$\text{sign}(r_{ki}) = \begin{cases} -1, & \text{if} \quad r_{ki} \leq 0 \\ +1, & \text{if} \quad r_{ki} > 0 \end{cases} \tag{7}$$

We further constructed a contingency table with the signs of the residuals for each split variable. If a split variable $z_l(l = 1, 2, \ldots, q)$ is numerical, we divide it into four quantile categories (C_1, C_2, C_3, C_4) to construct a 2×4 contingency table. If a split variable $z_l(l = 1, 2, \ldots, q)$ is categorical with g categories $(C_1, C_2, \ldots, C_g)$, we construct a $2 \times g$ contingency table, as shown in Table 1. We further obtain the Pearson chi-squared statistic χ_ν^2 for each contingency table. Because the degrees of freedom differ for each predictor variable, we equalize them into one by employing the Wilson-Hilferty approximation [30], $W_k(z_l) = \max\left(0, \left[\frac{7}{9} + \sqrt{\nu}\left\{\left(\frac{\chi_\nu{}^2}{\nu}\right)^{\frac{1}{3}} - 1 + \frac{2}{9\nu}\right\}\right]\right)$, $(k = 1, 2, \ldots, m, l = 1, 2, \ldots, q)$. The transformed statistics for each predictor variable are compared, and the predictor variables with the largest statistics are selected as the split variable. The split point (or set) of the selected split variable can be determined by evaluating the reduction of impurities. The ES and RA approaches are summarized as follows:

TABLE 1. Contingency table.

Numerical	$sign(r_{ik})$	C_1	C_2	C_3	C_4
	-1	n_{11}	n_{12}	n_{13}	n_{14}
	+1	n_{21}	n_{22}	n_{23}	n_{24}
Categorical	$sign(r_{ik})$	C_1	C_2	$\ldots$	C_g
	-1	n_{11}	n_{12}	$\ldots$	n_{1g}
	+1	n_{21}	n_{22}	$\ldots$	n_{2g}

PLSRT-ES algorithm for split rule selection:

1. Fit the basic model (2) with the fitting variables to the data.
2. Calculate impurity (4).
3. Evaluate the reduction of impurity (5) for all possible splits of the split variables.
4. Choose the best split variable and split point (or set) to maximize the reduction of impurity (5).

PLSRT-RA algorithm for split rule selection:

1. Obtain the residuals (6) after fitting the basic model (2) with the fitting variables to the data.
2. Obtain the statistic χ_1^2 for each categorical split variable.
   a) From a $2 \times g$ contingency table, obtain the chi-squared statistic χ_ν^2 to test independence, where ν is the degree of freedom of the contingency table $(\nu = g - 1)$.
   b) Use the Wilson-Hilferty approximation to convert χ_ν^2 to 1-d.f.chi-squared χ_1^2.
3. Obtain the statistic χ_1^2 for each non-categorical split variable.
   a) Divide the non-categorical variables into four quantile categories C_1, C_2, C_3, C_4.
   b) Use the Wilson-Hilferty approximation to convert χ_ν^2 to 1-d.f.chi-squared χ_1^2.
4. Find the best split variable with the largest χ_1^2.
5. Choose the best split point (or set) to maximize the reduction of impurity (5) for all possible split points (or sets) of the selected split variable.

The PLSRT-ES or PLSRT-RA algorithm is applied recursively to partition the data until each node has fewer than a pre-specified number of observations. Thus, a tree-structured model is obtained.

D. TREE SIZE DETERMINATION

One of the important problems to consider whereas constructing a tree model is determining tree size. It is crucial to determine an appropriate tree size because trees that are too small may miss significant splits, whereas trees that are too large may cause an overfitting problem. Appropriate stopping and pruning rules can be used to solve these problems. We employ the cost-complexity pruning technique by Breiman et al. [6], which splits the data recursively until the sample sizes at each node are reasonably small and then prunes off insignificant branches of nodes. The minimal cost-complexity pruning technique finds an optimal tree T by minimizing the cost-complexity: $R_\alpha(T) = R(T) + \alpha\|\tilde{T}\|$, where $R(T)$ is the sum of the resubstitution loss functions over the terminal nodes of a tree T, α is a complexity parameter, and $\|\tilde{T}\|$ is the number of terminal nodes, $\tilde{T}$, in T. The algorithm first obtains a series of candidate trees, which are evaluated to select a final value for α, and hence a final tree.

III. SIMULATION STUDY

In this section, we investigate the performance of the proposed PLSRT modeling algorithm using simulated data. They are compared with several other methods for split variable selection and prediction accuracy under several situations for univariate and multivariate response cases. In both cases, we simulate the data using the following fitting and split variables. The fitting variables $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_5)$ are assumed to follow $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where:

$$\boldsymbol{\mu} = \begin{bmatrix} 5 \\ 5 \\ 5 \\ 5 \\ 5 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.9 & 0.7 & 0.5 & 0.3 \\ 0.9 & 1 & 0.9 & 0.7 & 0.5 \\ 0.7 & 0.9 & 1 & 0.9 & 0.7 \\ 0.5 & 0.7 & 0.5 & 1 & 0.9 \\ 0.3 & 0.5 & 0.7 & 0.9 & 1 \end{bmatrix} \tag{8}$$

The three split variables are numerical: $z_1 \sim$ N(0, 1), $z_2 \sim$ U(0, 1), $z_3 \sim$ Exp(1), and the two split variables are categorical: $z_4 \sim$ Binomial(1/2, 1/2) and $z_5 \sim$ Multinomial(1/12, ..., 1/12), which have two and twelve categories, respectively, with the same probabilities.

A. UNIVARIATE CASE

1) MODEL SETTING

We considered five different models to evaluate PLSRT.

$$\boldsymbol{y} = 1 + \boldsymbol{x}_1 + \epsilon \tag{9}$$

$$\boldsymbol{y} = \begin{cases} 1 + \boldsymbol{x}_1 + \epsilon & \text{if } z_1 \leq 0 \\ (1+\beta_1) + (1+\beta_2)\boldsymbol{x}_1 + \epsilon & \text{otherwise} \end{cases} \tag{10}$$

$$\boldsymbol{y} = \begin{cases} 1 + \boldsymbol{x}_1 + \boldsymbol{x}_2 + \epsilon & \text{if } z_1 \leq 0 \\ (1+\beta_1) + (1+\beta_2)\boldsymbol{x}_1 + \boldsymbol{x}_2 + \epsilon & \text{otherwise} \end{cases} \tag{11}$$

$$\boldsymbol{y} = \begin{cases} 1 + \boldsymbol{x}_1 + \boldsymbol{x}_2 + \boldsymbol{x}_3 + \boldsymbol{x}_4 + \boldsymbol{x}_5 + \epsilon & \text{if } z_1 \leq 0 \\ (1+\beta_1) + (1+\beta_2)\boldsymbol{x}_1 + \boldsymbol{x}_2 + \boldsymbol{x}_3 & \text{otherwise} \\ +\boldsymbol{x}_4 + \boldsymbol{x}_5 + \epsilon & \end{cases} \tag{12}$$

$$\boldsymbol{y} = \begin{cases} 1 + \boldsymbol{x}_1 + \epsilon & \text{if } z_1 \leq 0 \text{ and } z_4 \in \{0\} \\ (1+\beta_1) + (1+\beta_2)\boldsymbol{x}_1 + \epsilon & \text{otherwise} \end{cases} \tag{13}$$

where the random error ϵ follows a standard normal distribution $N(0, 1)$. From each model, 200 observations are generated for training, and 200 are independent for testing.

TABLE 2. Variable selection probabilities with models (9), (10), (11), (12), and (13).

Model	(β_1, β_2)	Approach	z_1	z_2	z_3	z_4	z_5	z_1 or z_4
Model (9)		ES	0.120	0.095	0.070	0.000	0.715	
		RA	0.260	0.225	0.190	0.185	0.140	
Model (10)	(0.4, 0)	ES	0.340	0.075	0.045	0.000	0.540	
		RA	0.625	0.135	0.085	0.085	0.070	
	(0, 0.13)	ES	0.760	0.025	0.020	0.000	0.195	
		RA	0.880	0.045	0.035	0.025	0.015	
	(0.2, 0.1)	ES	0.815	0.020	0.015	0.000	0.150	
		RA	0.905	0.045	0.020	0.015	0.015	
Model (11)	(0.6, 0)	ES	0.765	0.025	0.025	0.005	0.180	
		RA	0.855	0.030	0.040	0.050	0.025	
	(0, 0.1)	ES	0.620	0.055	0.050	0.005	0.270	
		RA	0.735	0.035	0.080	0.070	0.080	
	(0.3, 0.05)	ES	0.700	0.045	0.030	0.005	0.220	
		RA	0.800	0.035	0.055	0.070	0.040	
Model (12)	(0.5, 0)	ES	0.615	0.055	0.065	0.005	0.260	
		RA	0.765	0.035	0.075	0.060	0.065	
	(0, 0.1)	ES	0.645	0.055	0.050	0.000	0.250	
		RA	0.720	0.045	0.090	0.085	0.060	
	(0.1, 0.08)	ES	0.635	0.055	0.050	0.000	0.260	
		RA	0.735	0.055	0.080	0.075	0.055	
Model (13)	(1, 0)	ES	0.370	0.045	0.025	0.175	0.385	0.545
		RA	0.370	0.030	0.015	0.575	0.010	0.945
	(0, 0.1)	ES	0.170	0.105	0.070	0.020	0.635	0.190
		RA	0.330	0.120	0.100	0.385	0.065	0.715
	(1, 0.1)	ES	0.560	0.010	0.005	0.330	0.095	0.890
		RA	0.370	0.005	0.005	0.620	0.000	0.990

2) SPLIT VARIABLE SELECTION

We further investigated the split variable selection of PLSRT. The estimated probabilities of selecting each predictor are recorded for 200 iterations, as listed in Table 2.

The response variable for the null model (9) is independent of all the split variables $z_l (l = 1, 2, \ldots, 5)$. Therefore, the five split variables should have the same selection probability of 0.2. PLSRT-RA selects each split variable with similar probabilities. However, using PLSRT-ES, the binary variable z_4 is not selected, and the 12-category variable z_5 is often selected with a probability of 0.715. It tends to choose z_5 more frequently than the other split variables because z_5 has a larger number of possible splits. For the other models (10), (11) and (12), the response variable $\boldsymbol{y}$ depends on the split variable z_1. Therefore, it is expected that split variable z_1 is selected. By PLSRT-RA, z_1 is mostly selected, but z_1 is less selected, and z_5 is selected by PLSRT-ES. For model (13), the response variable $\boldsymbol{y}$ depends on the split variables z_1 and z_4. Therefore, it is expected that split variables z_1 or z_4 is selected. By PLSRT-RA, the split variables z_1 and z_4 are mostly selected, however, the variable z_1 and z_4 are less selected, and z_5 is often selected by PLSRT-ES. This result implies that the ES approach is vulnerable to selection bias towards variables with larger possible splits. The RA approach selects the correct split variables fairly accurately.

3) PREDICTION ACCURACY

The proposed PLSRT method requires a stopping rule when performing the split, which sets the terminal node size to 10 or more. To evaluate the prediction accuracy, the correlation (Corr) between the observed and predicted response values y_i and $\hat{y}_i$ and the mean squared errors (MSE) are calculated

TABLE 3. Mean and standard error of Corr & MSE with models (9), (10), and (11).

Model	(β_1, β_2)	Methods	Corr(mean)	Corr(s.e)	MSE(mean)	MSE(s.e)
Model (9)		OLS	0.668	0.0030	1.112	0.0085
		PCR	0.670	0.0038	1.095	0.0096
		Tree	0.561	0.0042	1.514	0.0136
		RF	0.636	0.0032	1.221	0.0095
		PLS	0.681	0.0030	1.066	0.0083
		PLSRT-ES	0.690	0.0029	1.043	0.0077
		PLSRT-RA	0.690	0.0029	1.043	0.0077
Model (10)	(0.4, 0)	OLS	0.671	0.0030	1.130	0.0085
		PCR	0.673	0.0038	1.109	0.0096
		Tree	0.555	0.0041	1.569	0.0142
		RF	0.640	0.0032	1.242	0.0096
		PLS	0.683	0.0029	1.085	0.0083
		PLSRT-ES	0.683	0.0029	1.084	0.0080
		PLSRT-RA	0.684	0.0029	1.081	0.0079
	(0, 0.13)	OLS	0.696	0.0027	1.164	0.0088
		PCR	0.703	0.0031	1.132	0.0093
		Tree	0.590	0.0042	1.606	0.0159
		RF	0.672	0.0030	1.286	0.0101
		PLS	0.706	0.0027	1.123	0.0087
		PLSRT-ES	0.706	0.0029	1.120	0.0086
		PLSRT-RA	0.708	0.0028	1.115	0.0084
	(0.2, 0.1)	OLS	0.692	0.0028	1.169	0.0088
		PCR	0.698	0.0033	1.139	0.0097
		Tree	0.586	0.0042	1.610	0.0153
		RF	0.669	0.0030	1.283	0.0099
		PLS	0.702	0.0028	1.126	0.0087
		PLSRT-ES	0.704	0.0029	1.120	0.0085
		PLSRT-RA	0.705	0.0029	1.115	0.0083
Model (11)	(0.6, 0)	OLS	0.875	0.0017	1.151	0.0082
		PCR	0.880	0.0011	1.102	0.0083
		Tree	0.821	0.0018	1.630	0.0144
		RF	0.864	0.0012	1.284	0.0098
		PLS	0.883	0.0010	1.079	0.0074
		PLSRT-ES	0.883	0.0011	1.075	0.0073
		PLSRT-RA	0.884	0.0010	1.065	0.0071
	(0, 0.1)	OLS	0.880	0.0011	1.143	0.0081
		PCR	0.885	0.0011	1.097	0.0082
		Tree	0.829	0.0017	1.611	0.0134
		RF	0.869	0.0012	1.280	0.0099
		PLS	0.887	0.0010	1.074	0.0074
		PLSRT-ES	0.888	0.0010	1.072	0.0073
		PLSRT-RA	0.888	0.0010	1.066	0.0071
	(0.3, 0.05)	OLS	0.878	0.0011	1.146	0.0081
		PCR	0.882	0.0011	1.100	0.0084
		Tree	0.825	0.0017	1.622	0.0141
		RF	0.866	0.0012	1.284	0.0098
		PLS	0.885	0.0010	1.075	0.0074
		PLSRT-ES	0.885	0.0010	1.075	0.0074
		PLSRT-RA	0.886	0.0010	1.066	0.0070

TABLE 4. Mean and standard error of Corr & MSE with models (12) and (13).

Model	(β_1, β_2)	Methods	Corr(mean)	Corr(s.e)	MSE(mean)	MSE(s.e)
Model (12)	(0.5, 0)	OLS	0.971	0.0003	1.140	0.0081
		PCR	0.972	0.0003	1.095	0.0075
		Tree	0.907	0.0009	3.582	0.0297
		RF	0.963	0.0004	1.661	0.0172
		PLS	0.973	0.0003	1.074	0.0076
		PLSRT-ES	0.973	0.0003	1.058	0.0073
		PLSRT-RA	0.973	0.0003	1.052	0.0071
	(0, 0.1)	OLS	0.972	0.0003	1.143	0.0081
		PCR	0.973	0.0003	1.100	0.0081
		Tree	0.908	0.0009	3.636	0.0309
		RF	0.963	0.0004	1.678	0.0176
		PLS	0.973	0.0003	1.080	0.0077
		PLSRT-ES	0.974	0.0003	1.060	0.0074
		PLSRT-RA	0.974	0.0003	1.056	0.0072
	(0.1, 0.08)	OLS	0.972	0.0003	1.142	0.0081
		PCR	0.973	0.0003	1.097	0.0080
		Tree	0.908	0.0009	3.626	0.0311
		RF	0.963	0.0004	1.675	0.0176
		PLS	0.973	0.0003	1.078	0.0076
		PLSRT-ES	0.974	0.0003	1.059	0.0073
		PLSRT-RA	0.974	0.0003	1.055	0.0072
Model (13)	(1, 0)	OLS	0.673	0.0028	1.209	0.0090
		PCR	0.665	0.0035	1.223	0.0102
		Tree	0.543	0.0039	1.738	0.0150
		RF	0.643	0.0031	1.337	0.0102
		PLS	0.674	0.0029	1.195	0.0090
		PLSRT-ES	0.670	0.0030	1.209	0.0088
		PLSRT-RA	0.685	0.0029	1.162	0.0084
	(0, 0.1)	OLS	0.697	0.0027	1.140	0.0087
		PCR	0.699	0.0033	1.123	0.0095
		Tree	0.591	0.0040	1.568	0.0144
		RF	0.664	0.0029	1.279	0.0100
		PLS	0.705	0.0027	1.106	0.0086
		PLSRT-ES	0.708	0.0027	1.097	0.0081
		PLSRT-RA	0.709	0.0027	1.093	0.0080
	(1, 0.1)	OLS	0.700	0.0026	1.332	0.0099
		PCR	0.692	0.0033	1.351	0.0119
		Tree	0.601	0.0041	1.807	0.0177
		RF	0.689	0.0029	1.454	0.0114
		PLS	0.689	0.0028	1.362	0.0101
		PLSRT-ES	0.725	0.0031	1.227	0.0125
		PLSRT-RA	0.738	0.0027	1.178	0.0096

as follows:

$$\mathrm{Corr} = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)(\hat{y}_i - \bar{\hat{y}}_i)}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}}_i)^2}} \tag{14}$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{15}$$

where $\bar{y}_i$ and $\bar{\hat{y}}_i$ are the sample means for y_i and $\hat{y}_i$. They are evaluated by independent n=200 test observations. Several other methods such as ordinary least squares (OLS), principal component regression (PCR), ordinary regression tree (Tree), random forest (RF), and partial least squares (PLS) are included for comparison. The Corr and MSE values are calculated for 200 iterations.

Table 3 and Table 4 show the mean and standard error (s.e) of Corr and MSE values, respectively. Corr(mean) and Corr(s.e) are the mean and standard error of Corr and MSE(mean) and MSE(s.e) are the mean and standard error of MSE. For the OLS, PCR, and PLS methods, all the predictor variables are used for fitting, whereas for the tree and random forest methods, all the predictor variables are used for splitting. It is shown that PLSRT-ES and PLSRT-RA had higher correlations and lower MSEs than the other methods. Additionally, PLSRT-RA shows slightly better performance than PLSRT-ES.

B. MULTIVARIATE CASE

1) MODEL SETTING

For the multivariate response case, we first assume that $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2)$ is a 2-dimensional response variable. We considered four different models to evaluate PLSRT.

$$\begin{aligned} \boldsymbol{y}_1 &= 1 + \boldsymbol{x}_1 + \epsilon_1 \\ \boldsymbol{y}_2 &= 1 + \boldsymbol{x}_2 + \epsilon_2 \end{aligned} \tag{16}$$

$$\boldsymbol{y}_1 = \begin{cases} 1 + \boldsymbol{x}_1 + \epsilon_1 & \text{if } z_1 \leq 0 \text{ and } z_4 \in \{0\} \\ (1+\beta_1) + (1+\beta_2)\boldsymbol{x}_1 & \text{otherwise} \\ +\epsilon_1 & \end{cases}$$

TABLE 5. Variable selection probabilities with models (16), (17), (18) and (19).

Model	(β_1, β_2)	Approach	z_1	z_2	z_3	z_4	z_5	z_1 or z_4
Model (16)		ES	0.070	0.110	0.105	0.000	0.715	
		RA	0.265	0.200	0.195	0.150	0.190	
Model (17)	(1.1, 0)	ES	0.600	0.010	0.000	0.260	0.130	0.860
		RA	0.370	0.005	0.005	0.615	0.005	0.985
	(0, 0.2)	ES	0.540	0.025	0.020	0.230	0.185	0.770
		RA	0.395	0.015	0.005	0.570	0.015	0.965
	(1, 0.05)	ES	0.635	0.000	0.000	0.295	0.070	0.930
		RA	0.375	0.005	0.005	0.615	0.000	0.990
Model (18)	(1.5, 0)	ES	0.555	0.005	0.005	0.255	0.180	0.810
		RA	0.370	0.000	0.000	0.630	0.000	1.000
	(0, 0.3)	ES	0.590	0.010	0.000	0.225	0.175	0.815
		RA	0.355	0.005	0.000	0.640	0.000	0.995
	(1, 0.1)	ES	0.575	0.005	0.000	0.250	0.170	0.825
		RA	0.385	0.000	0.000	0.615	0.000	1.000
Model (19)	(1.5, 0)	ES	0.520	0.000	0.005	0.350	0.125	0.870
		RA	0.240	0.000	0.000	0.760	0.000	1.000
	(0, 0.25)	ES	0.535	0.000	0.010	0.300	0.155	0.835
		RA	0.290	0.000	0.000	0.710	0.000	1.000
	(0.9, 0.1)	ES	0.560	0.000	0.000	0.290	0.150	0.850
		RA	0.285	0.000	0.000	0.715	0.000	1.000

$$\boldsymbol{y}_2 = \begin{cases} 1 + \boldsymbol{x}_2 + \epsilon_2 & \text{if } z_1 \leq 0 \text{ and } z_4 \in \{0\} \\ (1+\beta_1) + (1+\beta_2)\boldsymbol{x}_2 & \text{otherwise} \\ +\epsilon_2 & \end{cases} \tag{17}$$

$$\boldsymbol{y}_1 = \begin{cases} 1 + \boldsymbol{x}_1 + \epsilon_1 & \text{if } z_1 \leq 0 \text{ and } z_4 \in \{0\} \\ (1+\beta_1) + (1+\beta_2)\boldsymbol{x}_1 & \text{otherwise} \\ +\epsilon_1 & \end{cases}$$

$$\boldsymbol{y}_2 = 1 + \boldsymbol{x}_2 + \epsilon_2 \tag{18}$$

$$\boldsymbol{y}_1 = \begin{cases} 1 + \boldsymbol{x}_1 + \boldsymbol{x}_3 + \boldsymbol{x}_5 + \epsilon_1 & \text{if } z_1 \leq 0 \text{ and } z_4 \in \{0\} \\ (1+\beta_1) + (1+\beta_2)\boldsymbol{x}_1 & \text{otherwise} \\ +\boldsymbol{x}_3 + \boldsymbol{x}_5 + \epsilon_1 & \end{cases}$$

$$\boldsymbol{y}_2 = \begin{cases} 1 + \boldsymbol{x}_2 + \boldsymbol{x}_4 + \boldsymbol{x}_5 + \epsilon_2 & \text{if } z_1 \leq 0 \text{ and } z_4 \in \{0\} \\ (1+\beta_1) + (1+\beta_2)\boldsymbol{x}_2 & \text{otherwise} \\ +\boldsymbol{x}_4 + \boldsymbol{x}_5 + \epsilon_2 & \end{cases} \tag{19}$$

where the random errors (ϵ_1, ϵ_2) follow a standard normal distribution $N(0, 1)$. From each model, 200 observations are generated for training, and 200 are independent for testing.

2) SPLIT VARIABLE SELECTION

We investigated the split variable selection of PLSRT. The estimated probabilities of selecting each predictor are recorded for 200 iterations, as listed in Table 5.

The response variables $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2)$ for the null model (16) are independent of all split variables $z_l (l = 1, 2, \ldots, 5)$. Therefore, the five split variables should have a similar selection probability of 0.2. PLSRT-RA selects each split variable with similar probabilities. However, the binary variable z_4 is not selected, whereas the 12-category variable z_5 is often selected with a probability of 0.715 by PLSRT-ES. For the other models (17), (18) and (19), the response variables $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ depend on the split variables z_1 and z_4. Therefore, it is expected that split variable z_1 or z_4 is selected. The split variables z_1 and z_4 are mostly selected by PLSRT-RA. However, by PLSRT-ES, the z_1 and z_4 are less selected and z_5 is often selected. This is because it is the same as that in univariate cases. Therefore, the ES approach is vulnerable to selection bias toward variables with larger possible splits in both univariate and multivariate response cases. The RA approach selects the correct split variables fairly accurately.

3) PREDICTION ACCURACY

For the multivariate cases, the proposed PLSRT method requires a stopping rule when performing the split, which sets the terminal node size to 10 or more as in the univariate case. We calculated Corr and MSE to examine the prediction accuracy, as follows:

$$\text{Corr} = \frac{1}{m}\sum_{k=1}^{m} \frac{\sum_{i=1}^{n}(y_{ki} - \bar{y}_{ki})(\hat{y}_{ki} - \bar{\hat{y}}_{ki})}{\sqrt{\sum_{i=1}^{n}(y_{ki} - \bar{y}_{ki})^2 \sum_{i=1}^{n}(\hat{y}_{ki} - \bar{\hat{y}}_{ki})^2}} \tag{20}$$

$$\text{MSE} = \frac{1}{n \times m}\sum_{k=1}^{m}\sum_{i=1}^{n}(y_{ki} - \hat{y}_{ki})^2 \tag{21}$$

where $m = 2$ is the dimension of $\boldsymbol{Y}$, $\bar{y}_{ki}$ and $\bar{\hat{y}}_{ki}$ are the sample means of y_{ki} and $\hat{y}_{ki}$. They are evaluated using independent 200 test observations. Several methods such as PCR and PLS are further included for comparison. The Corr and MSE values are then calculated for 200 iterations.

Table 6 shows the mean and standard error of Corr and MSE values. For PCR and PLS, all predictor variables are used for fitting. It is shown that PLSRT-ES and PLSRT-RA had higher correlations and lower MSEs than the other methods. Additionally, PLSRT-RA shows slightly better performance than PLSRT-ES.

IV. CASE STUDY

A. UNIVARIATE CASE

We consider univariate response data, referred to as Usedcar data from Kcar Corporation. The data comprised 985 vehicles sold. The response variable, the real sell price, is the vehicle's real sold price. The predictor variables are further used as six spilt variables and six fitting variables. The variables as described in Table 7.

Fig.1 shows the results of the PLSRT-RA and PLSRT-ES for the Usedcar data. By PLSRT-RA, the type of vehicle segment is the primary factor that creates the subgroups, and then the vehicle used years and vehicle mileage divide the vehicle sold situation into four parts. PLS fits for each terminal node are obtained in the following form:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1\boldsymbol{x}_1 + \hat{\beta}_2\boldsymbol{x}_2 + \hat{\beta}_3\boldsymbol{x}_3 + \hat{\beta}_4\boldsymbol{x}_4 + \hat{\beta}_5\boldsymbol{x}_5 + \hat{\beta}_6\boldsymbol{x}_6 \tag{22}$$

Table 8 shows the PLS estimates for each terminal node. The PLS estimates for terminal node 5 significantly differ from those for the other terminal nodes. Terminal node 5 belongs

TABLE 6. Mean and standard error of Corr & MSE with models (16), (17), (18), and (19).

Model	(β_1, β_2)	Methods	Corr(mean)	Corr(s.e)	MSE(mean)	MSE(s.e)
Model (16)		PCR	0.681	0.0027	1.074	0.0076
		PLS	0.691	0.0019	1.050	0.0056
		PLSRT-ES	0.698	0.0019	1.031	0.0054
		PLSRT-RA	0.698	0.0019	1.031	0.0054
Model (17)	(1.1, 0)	PCR	0.693	0.0022	1.166	0.0069
		PLS	0.690	0.0019	1.179	0.0063
		PLSRT-ES	0.691	0.0023	1.179	0.0075
		PLSRT-RA	0.700	0.0021	1.149	0.0065
	(0, 0.2)	PCR	0.736	0.0020	1.159	0.0072
		PLS	0.735	0.0017	1.168	0.0062
		PLSRT-ES	0.732	0.0019	1.181	0.0069
		PLSRT-RA	0.740	0.0018	1.151	0.0065
	(1, 0.05)	PCR	0.706	0.0021	1.197	0.0070
		PLS	0.703	0.0018	1.210	0.0065
		PLSRT-ES	0.715	0.0024	1.174	0.0082
		PLSRT-RA	0.722	0.0022	1.149	0.0072
Model (18)	(1.5, 0)	PCR	0.675	0.0029	1.215	0.0085
		PLS	0.688	0.0019	1.176	0.0064
		PLSRT-ES	0.687	0.0022	1.176	0.0072
		PLSRT-RA	0.697	0.0020	1.141	0.0063
	(0, 0.3)	PCR	0.706	0.0028	1.232	0.0097
		PLS	0.719	0.0018	1.185	0.0067
		PLSRT-ES	0.717	0.0020	1.178	0.0077
		PLSRT-RA	0.726	0.0019	1.137	0.0064
	(1, 0.1)	PCR	0.687	0.0028	1.219	0.0088
		PLS	0.699	0.0019	1.178	0.0065
		PLSRT-ES	0.698	0.0021	1.175	0.0074
		PLSRT-RA	0.707	0.0020	1.140	0.0064
Model (19)	(1.5, 0)	PCR	0.920	0.0007	1.262	0.0066
		PLS	0.920	0.0007	1.268	0.0070
		PLSRT-ES	0.923	0.0009	1.220	0.0105
		PLSRT-RA	0.927	0.0007	1.160	0.0067
	(0, 0.25)	PCR	0.929	0.0006	1.223	0.0064
		PLS	0.929	0.0006	1.227	0.0066
		PLSRT-ES	0.930	0.0007	1.209	0.0089
		PLSRT-RA	0.933	0.0006	1.156	0.0073
	(0.9, 0.1)	PCR	0.924	0.0007	1.242	0.0065
		PLS	0.924	0.0007	1.247	0.0069
		PLSRT-ES	0.926	0.0008	1.220	0.0101
		PLSRT-RA	0.930	0.0007	1.158	0.0077

TABLE 7. Variable description for the Usedcar data.

Role	Variable	Description	Type
Response	y	Real sell price	Numeric
Fitting	x_1 : **price1**	Recommended purchase price	Numeric
	x_2 : **price2**	Real purchase price	Numeric
	x_3 : **price3**	Recommended sell price	Numeric
	x_4 : **price4**	Commercialization price	Numeric
	x_5 : **price5**	Prediction profit	Numeric
	x_6 : **price6**	Warranty	Numeric
Splitting	z_1 : **segment**	Vehicle segment (city, compact, mid-size, large-size, SUV)	Categorical
	z_2 : **color**	Vehicle color (white, pearl, black, silver, dark gray, others)	Categorical
	z_3 : **fuel**	Type of vehicle fuel (gasoline, diesel)	Categorical
	z_4 : **accident**	Type of vehicle accident (accident-free, accident, replacement)	Categorical
	z_5 : **year**	Vehicle used year	Numeric
	z_6 : **mileage**	Vehicle Mileage	Numeric

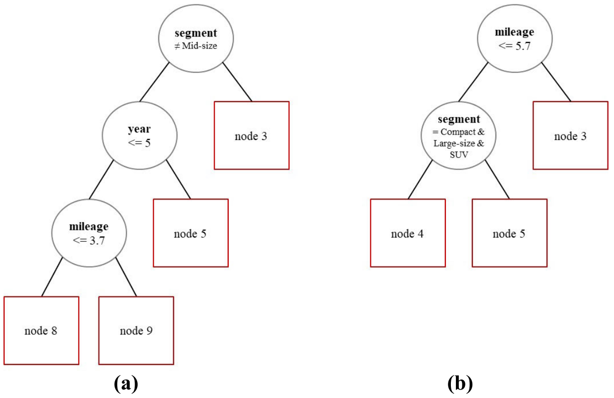

FIGURE 1. PLSRT for the Usedcar data. (a) PLSRT-RA split rule selection approach and the cost-complexity pruning. (b) PLSRT-ES split rule selection approach and the cost-complexity pruning. An observation moves to the left node if the condition is satisfied and moves to the right otherwise.

TABLE 8. Estimates for each terminal node with PLSRT method.

Method	Node	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
PLSRT-RA	node 8	19.78	-0.03	0.82	0.21	0.68	0.83	0.06
	node 9	43.51	0.02	0.80	0.17	0.59	0.87	-0.08
	node 5	47.63	-0.22	1.24	0.01	0.63	0.78	0.25
	node 3	17.88	-0.01	0.88	0.14	0.82	0.95	-0.25
PLSRT-ES	node 4	53.81	-0.10	0.78	0.28	0.56	0.84	-0.10
	node 5	-5.02	-0.15	0.82	0.33	0.68	0.91	0.05
	node 3	48.35	-0.08	1.07	0.02	0.73	0.76	0.17

to a case in which the type of vehicle segment is not mid-size and **year** is more than 5. It has little influence on the recommended sell price ($\hat{\beta}_3 = 0.01$), whereas it has more influence on the recommended purchase price ($\hat{\beta}_1 = -0.22$) and the real purchase price ($\hat{\beta}_2 = 1.24$). According to PLSRT-ES, the mileage factor is the primary factor that creates subgroups, and the type of vehicle segment divides the vehicle sold situation into three parts. The PLS estimate values for terminal node 3 differ from those for the other terminal nodes. Terminal node 3 belongs to the case where the vehicle used year is less than 5.7, and the type of vehicle segment is mid-size or compact. It insignificantly influences the recommended sell price ($\hat{\beta}_3 = 0.02$), but has a greater influence on the real purchase price ($\hat{\beta}_2 = 1.07$). Using both approaches, the segment and mileage variables are more important than the other split variables. Additionally, the vehicle used year divides the data in more detail using the PLSRT-RA.

B. MULTIVARIATE CASE

We consider multivariate response data, referred to as SNUDH data, from the Seoul National University Dental Hospital. The data included 318 patients and is detailed in the Suh et al. [31]. The 64 response variables, Facial landmarks after orthognathic surgery, are the facial landmarks (x-coordinates and y-coordinates) after orthognathic surgery. The predictor variables are divided into six spilt variables (external factors) and 232 fitting variables (landmark factors), as described in Table 9.

Fig.2 shows the results of the PLSRT-RA and PLSRT-ES for the SNUDH data. PLS fits for each terminal node are obtained in the following form:

$$\hat{y}_{k.} = \hat{\beta}_{0,k} + \hat{\beta}_{1,k}\boldsymbol{x}_1 + \ldots + \hat{\beta}_{232,k}\boldsymbol{x}_{232}, \; k = 1, 2, \ldots, 64 \tag{23}$$

TABLE 9. Variable description for SNUDH data.

Role	Variable	Description	Type
Response	$y_1, y_2, \ldots, y_{64}$	Facial landmarks after surgery	Numeric
Fitting	$x_1, x_2, \ldots, x_{64}$	Facial landmarks before surgery	Numeric
	$x_{65}, x_{66}, \ldots, x_{232}$	Skeletal landmarks before and after surgery	Numeric
Splitting	z_1 : **class**	Type of craniofacial class (class 2, class 3)	Categorical
	z_2 : **mx**	Type of jaw (mandibula, maxilla)	Categorical
	z_3 : **genio**	Type of genioplasty (not, genioplasty)	Categorical
	z_4 : **sex**	Gender (female, male)	Categorical
	z_5 : **age**	Age at surgery	Numeric
	z_6 : **asym**	Amount of mandibular asymmetry	Numeric

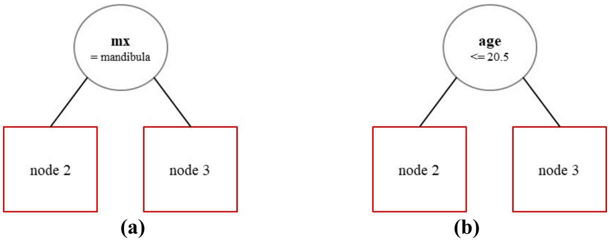

FIGURE 2. PLSRT for SNUDH data. (a) PLSRT-RA split rule selection approach and the cost-complexity pruning. (b) PLSRT-ES split rule selection approach and the cost-complexity pruning.

According to PLSRT-RA, the type of jaw is a factor that creates heterogeneous subgroups. It further divides the surgery situation into two subgroups. Using PLSRT-ES, the age at surgery divides the surgery situation into two subgroups. There are 232 estimated regression coefficients for each PLSRT model. In orthognathic surgery, the type of jaw is more important than the age at surgery. Therefore, PLSRT-RA yields a more realistic result.

V. CONCLUSION

In this study, we propose a new PLSRT modeling algorithm. This can be used when many predictor variables have multicollinear relationships. It integrates the merits of PLS regression modeling and tree-structured modeling; hence, it can solve the multicollinear problem and capture complex non-linear relationships. This provides a visible and interpretable predictive model. Such structures for modern biomedical data are often found. First, we divided the predictor variables into two categories. One is for fitting with high collinearity and the other is for splitting with interpretation. For variable selection, the RA approach solves the problems encountered by the ES approach, such as the undue preference for split variables with more possible splits and considerable computational costs. The PLSRT modeling method can be applied to both univariate and multivariate response data. Through simulation and case studies, we investigated the performance of PLSRT by comparing it with existing methods. From these advantages, we conclude that PLSRT satisfies both prediction accuracy and model interpretation.

REFERENCES

[1] J. H. Phillips, I. Nish, and J. Daskalogiannakis, "Orthognathic surgery in cleft patients," *Plastic Reconstructive Surgery*, vol. 129, no. 3, pp. 535e–548e, Mar. 2012.

[2] M. T. Chew, A. Sandham, and H. B. Wong, "Evaluation of the linearity of soft- to hard-tissue movement after orthognathic surgery," *Amer. J. Orthodontics Dentofacial Orthopedics*, vol. 134, no. 5, pp. 665–670, Nov. 2008.

[3] L. C. Kneafsey, S. J. Cunningham, A. Petrie, and T. J. Hutton, "Prediction of soft-tissue changes after mandibular advancement surgery with an equation developed with multivariable regression," *Amer. J. Orthodontics Dentofacial Orthopedics*, vol. 134, no. 5, pp. 657–664, Nov. 2008.

[4] 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," *J. Oral Maxillofacial Surg.*, vol. 70, no. 10, pp. e553–e562, Oct. 2012.

[5] J. N. Morgan and J. A. Sonquist, "Problems in the analysis of survey data, and a proposal," *J. Amer. Stat. Assoc.*, vol. 58, no. 302, p. 415, Jun. 1963.

[6] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification Regression Trees*. London, U.K.: Chapman & Hall, 1984.

[7] W.-Y. Loh and N. Vanichsetakul, "Tree-structured classification via generalized discriminant analysis," *J. Amer. Stat. Assoc.*, vol. 83, no. 403, p. 715, Sep. 1988.

[8] H. Ahn and W.-Y. Loh, "Tree-structured proportional hazards regression modeling," *Biometrics*, vol. 50, no. 2, p. 471, Jun. 1994.

[9] P. Chaudhuri, M. C. Huang, W. Y. Loh, and R. Yao, "Piecewise-polynomial regression trees," *Statistica Sinica*, vol. 4, pp. 143–167, Jan. 1994.

[10] W. P. Alexander and S. D. Grimshaw, "Treed regression," *J. Comput. Graph. Statist.*, vol. 5, no. 2, pp. 156–175, Jun. 1996.

[11] H. A. Chipman, E. L. George, and R. E. McCulloch, "Bayesian cart model search," *J. Amer. Stat. Assoc.*, vol. 93, pp. 935–948, Sep. 1998.

[12] H. Kim and W.-Y. Loh, "Classification trees with unbiased multiway splits," *J. Amer. Stat. Assoc.*, vol. 96, no. 454, pp. 589–604, Jun. 2001.

[13] S.-H. Eo and H. Cho, "Tree-structured mixed-effects regression modeling for longitudinal data," *J. Comput. Graph. Statist.*, vol. 23, no. 3, pp. 740–760, Jul. 2014.

[14] J. Kim and H. Cho, "Seemingly unrelated regression tree," *J. Appl. Statist.*, vol. 46, no. 7, pp. 1177–1195, May 2019.

[15] J. Kim, H. Cho, and S. Bang, "Unified noncrossing multiple quantile regressions tree," *J. Comput. Graph. Statist.*, vol. 28, no. 2, pp. 454–465, Apr. 2019.

[16] W. Y. Loh, "Regression trees With unbiased variable selection and interaction detection," *Statistica Sinica*, vol. 12, pp. 361–386, Apr. 2002.

[17] H. Zhang, A. Zhou, H. Qian, and H. Zhang, "PS-tree: A piecewise symbolic regression tree," *Swarm Evol. Comput.*, vol. 71, Jun. 2022, Art. no. 101061.

[18] K. J. Knierim, J. A. Kingsbury, C. J. Haugh, and K. M. Ransom, "Using boosted regression tree models to predict salinity in Mississippi embayment aquifers, central United States," *JAWRA J. Amer. Water Resour. Assoc.*, vol. 56, no. 6, pp. 1010–1029, Dec. 2020.

[19] Z. Said, P. Sharma, A. K. Tiwari, V. V. Le, Z. Huang, V. G. Bui, and A. T. Hoang, "Application of novel framework based on ensemble boosted regression trees and Gaussian process regression in modelling thermal performance of small-scale organic Rankine cycle (ORC) using hybrid nanofluid," *J. Cleaner Prod.*, vol. 360, Aug. 2022, Art. no. 132194.

[20] D. Han, H. An, F. Wang, X. Xu, Z. Qiao, M. Wang, X. Sui, S. Liang, X. Hou, H. Cai, and Y. Liu, "Understanding seasonal contributions of urban morphology to thermal environment based on boosted regression tree approach," *Building Environ.*, vol. 226, Dec. 2022, Art. no. 109770.

[21] A. O. Alnahit, A. K. Mishra, and A. A. Khan, "Stream water quality prediction using boosted regression tree and random forest models," *Stochastic Environ. Res. Risk Assessment*, vol. 36, no. 9, pp. 2661–2680, Sep. 2022.

[22] W. Y. Loh, "Logistic regression tree analysis," in *Springer Handbook of Engineering Statistics*. Piscataway, NJ, USA: Springer, 2023, pp. 593–604.

[23] J. Pan, V. Bunn, B. Hupf, and J. Lin, "Bayesian additive regression trees (BART) with covariate adjusted borrowing in subgroup analyses," *J. Biopharmaceutical Statist.*, vol. 32, no. 4, pp. 613–626, Jul. 2022.

[24] T. E. Clark, F. Huber, G. Koop, M. Marcellino, and M. Pfarrhofer, "Tail forecasting with multivariate Bayesian additive regression trees," *Int. Econ. Rev.*, vol. 64, no. 3, pp. 979–1022, Aug. 2023.

[25] S. Um, A. R. Linero, D. Sinha, and D. Bandyopadhyay, ''Bayesian additive regression trees for multivariate skewed responses,'' *Statist. Med.*, vol. 42, no. 3, pp. 246–263, Feb. 2023.

[26] C. H. Yeh and C. H. Spiegelman, ''Partial least squares and classification and regression trees,'' *Chemometric Intell. Lab. Syst.*, vol. 22, no. 1, pp. 17–23, Jan. 1994.

[27] N. Reddy, M. Gebreslasie, and R. Ismail, ''A hybrid partial least squares and random forest approach to modelling forest structural attributes using multispectral remote sensing data,'' *South Afr. J. Geomatics*, vol. 6, no. 3, p. 377, Nov. 2017.

[28] Z. Hao, J. Du, B. Nie, F. Yu, R. Yu, and W. Xiong, ''Random forest regression based on partial least squares,'' in *Proc. Int. Conf. Artif. Intell., Technol. Appl.*, 2006, pp. 1–6.

[29] W.-Y. Loh, ''Improving the precision of classification trees,'' *Ann. Appl. Statist.*, vol. 3, no. 4, pp. 1710–1737, Dec. 2009.

[30] E. B. Wilson and M. M. Hilferty, ''The distribution of chi-square,'' *Proc. Nat. Acad. Sci. USA*, vol. 17, pp. 684–688, Dec. 1931.

[31] H.-Y. Suh, H.-J. Lee, Y.-S. Lee, S.-H. Eo, R. E. Donatelli, and S.-J. Lee, ''Predicting soft tissue changes after orthognathic surgery: The sparse partial least squares method,'' *Angle Orthodontist*, vol. 89, no. 6, pp. 910–916, Nov. 2019.

WENXING YU received the B.S. degree in statistics from Shandong University of Technology, Shandong, China, in June 2010, the M.S. degree in economic statistics from Korea University, Sejong, Republic of Korea, in June 2012, and the Ph.D. degree in statistics from Korea University, Seoul, Republic of Korea, in Feburary 2024.

From March 2013 to February 2019, he was with SK Encar Company, Seoul. He has published three journals related to statistics.

SHIN-JAE LEE received the DDS Diploma from the Seoul National University School of Dentistry in 1991, the M.S. and Ph.D. degrees in orthodontics from Seoul National University, Seoul, South Korea, in 1994 and 1999, respectively, and the Ph.D. degree in statistics from Korea University, in 2012.

Currently, he is a Professor with the Department of Orthodontics, Seoul National University, where he has been working since 2002.

HYUNGJUN CHO received the B.E. and M.S. degrees in statistics from Korea University, Seoul, in 1991 and 1993, respectively, and the Ph.D. degree in statistics from the University of Wisconsin, Madison, in 2002.

He was a Research Associate and an Assistant Professor with the Division of Biostatistics and Epidemiology, Department of Public Health Sciences, University of Virginia, Charlottesville, from 2002 to 2005 and from 2005 to 2006, respectively. Since September 2006, he has been a Professor with the Department of Statistics, Korea University. His research interests include data mining and bioinformatics.