Communications for Statistical Applications and Methods
2023, Vol. 30, No. 1, 65–73

https://doi.org/10.29220/CSAM.2023.30.1.065
Print ISSN 2287-7843 / Online ISSN 2383-4757

Modified partial least squares method implementing mixed-effect model

Kyunga Kim[a,b], Shin-Jae Lee[c], Soo-Heang Eo[d], HyungJun Cho[e], Jae Won Lee[1,e]

[a]Biomedical Statistics Center, Research Institute for Future Medicine, Samsung Medical Center, Korea;
[b]Department of Digital Health, Samsung Advanced Institute for Health Sciences & Technology, Sungkyunkwan University, Korea;
[c]Seoul National University School of Dentistry & Dental Research Institute, Korea;
[d]GreenLabs Inc., Korea; [e]Department of Statistics, Korea University, Korea

Abstract

Contemporary biomedical data often involve an ill-posed problem owing to small sample size and large number of multi-collinear variables. Partial least squares (PLS) method could be a plausible alternative to an ill-conditioned ordinary least squares. However, in the case of a PLS model that includes a random-effect, how to deal with a random-effect or mixed effects remains a widely open question worth further investigation. In the present study, we propose a modified multivariate PLS method implementing mixed-effect model (PLSM). The advantage of PLSM is its versatility in handling serial longitudinal data or its ability for taking a random-effect into account. We conduct simulations to investigate statistical properties of PLSM, and showcase its real clinical application to predict treatment outcome of esthetic surgical procedures of human faces. The proposed PLSM seemed to be particularly beneficial 1) when random-effect is conspicuous; 2) the number of predictors is relatively large compared to the sample size; 3) the multicollinearity is weak or moderate; and/or 4) the random error is considerable.

Keywords: partial least squares, random-effect, multivariate linear mixed-effects model

1. Introduction

If a data set includes a large number of variables with relatively small sample size, the situation can be called a high-dimensionality low sample size (HDLSS) problem. It may be also common to see a data set that demonstrates a high correlation structure between and/or within variables. An example for this situation might be face change data after cosmetic surgical procedures. In the past, accurate prediction of a treatment outcome after facial cosmetic surgery had been challenging. This was primarily due to the fact that facial soft-tissue responses were influenced by a number of factors that may reach approximately 300 variables including soft-tissue landmarks of a patient, skeletal bone landmarks, the amount of surgical skeletal repositioning of the bone both in vertical and horizontal direction, and demographic characteristics (Suh *et al*., 2019; Hwang *et al*., 2021). Added to

This work was partly supported by the National Research Foundation of Korea (NRF grant No. 2020R1A2C1A0100 8262) funded by the Korea government (MSIT).
Shin-Jae Lee and Kyunga Kim contributed equally to this study.
[1] Corresponding author: Department of Statistics, Korea University, 145 Anam-Ro, Sungbuk-Gu, Seoul 02841, Korea. E-mail: jael@korea.ac.kr

Published 31 January 2023 / journal homepage: http://csam.or.kr

the strong correlation between bone and soft-tissue response, significant correlation structures were commonly observed between and within the predictor and response matrices of an individual subject. Furthermore, the horizontal surgical repositioning of a certain skeletal anatomy could induce its vertical repositioning also, and vice versa. Conventional ordinary least squares methods cannot properly deal with these challenges. Consequently, the prediction of treatment outcome after surgery requires a sophisticated methodology that involves multiple predictors and multiple response variables simultaneously. In this situation, the partial least squares (PLS) method could be a plausible alternative to cope with this problem instead of an ill-conditioned ordinary least squares regression model (Lee *et al*., 2010; Suh *et al*., 2012; Lee *et al*., 2014; Suh *et al*., 2019; Fordellone and Vichi, 2020).

However, the conventional PLS method has a limitation when the predictor variables have mixed effects, i.e., fixed and random effects. Contemporary data sets may often include both fixed-and random effects. In the case of a PLS model that includes a random-effect, how to deal with a random-effect or mixed effects remains a wide open question worth further investigation.

The aim of the present study is to propose a modified multivariate PLS method implementing mixed-effect model (PLSM), which can provide more accurate prediction than the conventional PLS algorithm. To evaluate PLSM and to compare it with the conventional PLS method, we perform a simulation study and apply PLSM to a real clinical data set for predicting treatment outcomes of facial cosmetic surgery.

2. Review on PLS and the linear mixed-effect model

The PLS regression finds principal components that explain response variables $\mathbf{Y}$ as well as predictors $\mathbf{X}$. In other words, it identifies a set of latent variables, $\mathbf{T}$ and $\mathbf{U}$, that contains most of the variation in $\mathbf{X}$ and models $\mathbf{Y}$ best at the same time,

$$\mathbf{Y} = \mathbf{XB} \text{ with } \mathbf{X} = \mathbf{TP}^\top + \mathbf{E} \quad \text{and} \quad \mathbf{Y} = \mathbf{UQ}^\top + \mathbf{F}, \tag{2.1}$$

where $\mathbf{Y}_{n\times k}$, $\mathbf{X}_{n\times p}$ and $\mathbf{B}_{p\times k}$ are matrices of response, predictor and regression coefficient, respectively; $\mathbf{T}_{n\times a}$, $\mathbf{P}_{p\times a}$, and $\mathbf{E}_{n\times p}$ are score, loading and residual matrices for $\mathbf{X}$, respectively; $\mathbf{U}_{n\times a}$, $\mathbf{Q}_{k\times a}$, and $\mathbf{F}_{n\times k}$ are score, loading and residual matrices for Y, respectively; n, k, p and a are the numbers of observations, response variables, predictor variables and latent variables, respectively. Basic ideas of PLS were developed in chemometrics (Wehrens, 2011). In chemometrics, the number of predictor variables (p) often exceeds the number of observations (n), a typical HDLSS situation. In addition, when the variables are highly collinear or correlated, the conventional ordinary least squares model is not suitable for producing robust results (Zhou *et al*., 2005). Since PLS can eliminates multi-collinearity and reduce the dimensionality to improve the prediction accuracy, it has become a viable tool for predictive purposes in many scientific and technological applications including image analysis, biostatistics and bioinformatics (Dai *et al*., 2006; Mevik and Wehrens, 2007; Hastie *et al*., 2009; Martins *et al*., 2010; Krishnan *et al*., 2011).

A mixed-effect model contains both predictor variables with a fixed-effect and those with a random-effect,

$$\mathbf{Y} = \mathbf{XB} + \mathbf{ZA} + \mathbf{E}, \tag{2.2}$$

where $\mathbf{Y}_{n\times k}$ and $\mathbf{E}_{n\times k}$ are response and residual matrices; $\mathbf{X}_{n\times p}$ and $\mathbf{B}_{p\times k}$ are design and parameter matrices for predictors with fixed effects; $\mathbf{Z}_{n\times q}$ and $\mathbf{A}_{q\times k}$ are design and parameter matrices for predictors with random effects; n, k, p and q are the numbers of observations, response, fixed-effect predictor and random-effect predictor variables, respectively.

The linear mixed model has some statistical advantages compared to the linear model with only fixed effects. The mixed model does not require observations to be independent and with equal variance, and can test random effects providing individual variability. Therefore, it can be useful when measurements are made on the same or related statistical units, such as repeated measurements over time in longitudinal growth data sets (Laird and Ware, 1982).

3. Joint modelling approach for the PLS algorithm implementing mixed effects: A modified PLS method implementing mixed-effect model (PLSM)

Although the PLS method is successful to solve a multi-collinearity problem with HDLSS data ($n << p$) in a linear model only with fixed effects, it remains limited when there exist random effects. In order to handle multi-collinearity and HDLSS in a linear mixed-effect model, $\mathbf{Y} = \mathbf{XB} + \mathbf{ZA} + \mathbf{E}$, we propose a three-step approach that employs the PLS algorithm and takes random effects into account. In the first step, random effects are regressed out using the random-effect only model:

- **Step 1** : $\mathbf{Y} = \mathbf{ZA}^* + \mathbf{E}^*$, and residuals $\mathbf{Y}^* = \mathbf{Y} - \hat{\mathbf{Y}}$ are computed and stored for the next step.

  Then we use the residuals $\mathbf{Y}^*$ as response variables when constructing PLS components $\mathbf{T}$ from the fixed-effect only model based on outer relations below:

- **Step 2** : $\mathbf{Y}^* = \mathbf{XB}^* + \mathbf{E}^{**}$ with $\mathbf{X} = \mathbf{T}^\top\mathbf{P}$ and $\mathbf{Y}^* = \mathbf{UQ}^\top$.

  In the final step, a linear mixed-effect model with $a(< p)$ dominant PLS components $\mathbf{T}_a$ and the random effects is considered to predict $\mathbf{Y}$. At this step, the optimal number of PLS components is determined by the cross-validation technique:

- **Step 3** : $\mathbf{Y} = \mathbf{T}_a\mathbf{B}_a + \mathbf{ZA} + \mathbf{E}$.

  It shows how to combine the linear mixed-effect model and the conventional PLS to implement this method computationally. To restate, aforementioned joint modelling can be called a simple modification upon the conventional PLS algorithm by replacing the first PLS component with random effects at the first step.

4. Simulation studies

Empirical studies using real data and simulations have an important role in investigating and validating a new method. Simulations can be performed to investigate statistical properties of the new method as to whether it satisfies the assumptions, for what range of parameter values it performs well before it will be applied to real data (Garthwaite, 1994). In this section, we investigated the performance of PLSM and compared it with the conventional PLS method. The prediction performance was measured by the root mean squared error of prediction (RMSEP).

4.1. Simulation settings

Inspired by real clinical datasets to which we applied our PLSM later in Section 5, we formulated simulation datasets with a sample size of $n = 100$ based on the linear mixed-effect model (2) with various multi-collinearity structures:

1. Multi-collinear predictor variables of $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ with fixed effects.

2. A single random-effect $\alpha \sim N(0, \sigma_a^2)$ which is independent of $\mathbf{X}$.

Table 1: Where the random error, $\sigma_\epsilon = 1$, comparisons of the root mean squared error of prediction (RMSEP) according to the five simulation scenarios at varying extents of the random-effect

Scenario	Random-effect	$n = 100 > p = 30$		$n = 100 = p = 100$		$n = 100 < p = 200$		$n = 100 << p = 500$	
(ρ_1, ρ_2)	σ_a	PLS	PLSM	PLS	PLSM	PLS	PLSM	PLS	PLSM
(0.00, 0.00)		1.27 (0.01)	1.39 (0.01)	1.60 (0.01)	1.66 (0.01)	1.77 (0.01)	1.18 (0.01)	1.90 (0.01)	1.88 (0.01)
(0.25, 0.25)		1.27 (0.01)	1.42 (0.01)	1.60 (0.01)	1.68 (0.01)	1.77 (0.01)	1.79 (0.01)	1.89 (0.01)	1.88 (0.01)
(0.80, 0.50)	1.0	1.34 (0.01)	1.51 (0.01)	1.50 (0.01)	1.64 (0.01)	1.64 (0.01)	1.17 (0.01)	1.80 (0.01)	1.81 (0.01)
(0.80, 0.80)		1.33 (0.01)	1.52 (0.01)	1.51 (0.01)	1.70 (0.01)	1.64 (0.01)	1.78 (0.01)	1.86 (0.01)	1.88 (0.01)
(0.87, 0.82)		1.33 (0.01)	1.50 (0.01)	1.48 (0.01)	1.67 (0.01)	1.60 (0.01)	1.15 (0.01)	1.78 (0.01)	1.84 (0.01)
(0.00, 0.00)		1.55 (0.01)	1.60 (0.01)	1.86 (0.01)	1.85 (0.01)	2.00 (0.01)	1.94 (0.01)	2.07 (0.01)	2.02 (0.01)
(0.25, 0.25)		1.56 (0.01)	1.63 (0.01)	1.85 (0.01)	1.85 (0.01)	2.00 (0.01)	1.95 (0.01)	2.07 (0.01)	2.02 (0.01)
(0.80, 0.50)	1.5	1.56 (0.01)	1.69 (0.01)	1.73 (0.01)	1.82 (0.01)	1.87 (0.01)	1.90 (0.01)	2.00 (0.01)	1.97 (0.01)
(0.80, 0.80)		1.55 (0.01)	1.70 (0.01)	1.74 (0.01)	1.87 (0.01)	1.89 (0.01)	1.96 (0.01)	2.08 (0.01)	2.04 (0.01)
(0.87, 0.82)		1.53 (0.01)	1.68 (0.01)	1.69 (0.01)	1.84 (0.01)	1.81 (0.01)	1.90 (0.01)	1.99 (0.01)	1.98 (0.01)
(0.00, 0.00)		1.88 (0.02)	1.85 (0.01)	2.16 (0.02)	2.06 (0.01)	2.25 (0.02)	2.14 (0.01)	2.27 (0.01)	2.20 (0.01)
(0.25, 0.25)		1.86 (0.02)	1.87 (0.01)	2.16 (0.02)	2.07 (0.01)	2.29 (0.02)	2.16 (0.01)	2.31 (0.01)	2.23 (0.01)
(0.80, 0.50)	2.0	1.85 (0.02)	1.93 (0.01)	2.00 (0.02)	2.03 (0.01)	2.13 (0.02)	2.09 (0.01)	2.26 (0.02)	2.17 (0.01)
(0.80, 0.80)		1.85 (0.02)	1.96 (0.01)	1.96 (0.01)	2.04 (0.01)	2.18 (0.02)	2.18 (0.01)	2.31 (0.02)	2.22 (0.01)
(0.87, 0.82)		1.85 (0.02)	1.95 (0.01)	1.96 (0.02)	2.05 (0.01)	2.09 (0.02)	2.12 (0.01)	2.26 (0.02)	2.20 (0.01)
(0.00, 0.00)		3.98 (0.05)	3.64 (0.04)	4.39 (0.06)	3.79 (0.04)	4.29 (0.05)	3.79 (0.04)	4.13 (0.04)	3.85 (0.04)
(0.25, 0.25)		3.96 (0.05)	3.67 (0.04)	4.31 (0.05)	3.75 (0.04)	4.35 (0.05)	3.82 (0.04)	4.19 (0.04)	3.90 (0.04)
(0.80, 0.50)	5.0	3.79 (0.05)	3.67 (0.04)	4.08 (0.05)	3.77 (0.04)	4.18 (0.05)	3.76 (0.04)	4.21 (0.05)	3.82 (0.04)
(0.80, 0.80)		3.83 (0.05)	3.76 (0.04)	4.02 (0.05)	3.77 (0.04)	4.33 (0.05)	3.90 (0.04)	4.40 (0.05)	3.90 (0.04)
(0.87, 0.82)		3.82 (0.05)	3.75 (0.04)	3.97 (0.05)	3.77 (0.04)	4.24 (0.05)	3.89 (0.04)	4.34 (0.05)	3.84 (0.04)

(ρ_1, ρ_2), correlation coefficients for two blocks with AR(1)-Type variance-covariance structures, respectively;
p,the number of predictor variables;
n, the sample size;
PLS, partial least squares regression;
PLSM, modified partial least squares implementing mixed-effect;
The values in the parentheses were the standard errors.
When the PLSM method showed a significantly better result (p-value < 0.0001), red-faced type was applied.

3. Univariate response y is considered with random error $\epsilon \sim N(0, \sigma_\epsilon^2)$ for the ease of calculation and interpretation.

For a sample size of $n = 100$, four different numbers of predictor variables p were considered: $n > p = 30; n = p = 100; n < p = 200$ and $n << p = 500$.

For simulating multi-collinearity structure, we mimicked the correlation structure captured in predictor variables of real clinical data (see Section 5). The predictor variables of this real clinical data are the Cartesian x- and y-coordinates of 71 anthropological landmarks detected on lateral human face. These coordinates values show the AR(1)-Type correlation structures within x-coordinates and within y-coordinates, but no correlation between x-coordinates and y-coordinates.

Based on this observation, we develop a block-diagonal variance-covariance structure Λ with two blocks, each of which has the AR(1) variance-covariance structure with auto-correlation coefficient ρ. Then, predictors are generated from a multivariate normal distribution MVN$(0, \Lambda)$. Note that the use of the AR(1) variance-covariance structure was inspired by the simulation study of Chun and Keles (2010). We consider a variety of multi-collinearity with the following five simulation scenarios:

1. No collinearity with $(\rho_1, \rho_2) = (0.00, 0.00)$.
2. Moderate collinearity with $(\rho_1, \rho_2) = (0.25, 0.25)$.
3. High and moderate collinearity with $(\rho_1, \rho_2) = (0.80, 0.50)$.
4. High collinearity with $(\rho_1, \rho_2) = (0.80, 0.80)$.
5. Very strong collinearity with $(\rho_1, \rho_2) = (0.87, 0.82)$, similar to the sample covariance of the real data.

Table 2: Where the random error, $\sigma_\epsilon = 2$, comparisons of the root mean squared error of prediction (RMSEP) according to the five simulation scenarios at varying extents of the random-effect

Scenario	random-effect	$n = 100 > p = 30$		$n = 100 = p = 100$		$n = 100 < p = 200$		$n = 100 << p = 500$	
(ρ_1, ρ_2)	σ_a	PLS	PLSM	PLS	PLSM	PLS	PLSM	PLS	PLSM
(0.00, 0.00)		2.03 (0.01)	2.04 (0.01)	2.34 (0.01)	2.28 (0.01)	2.43 (0.01)	2.34 (0.01)	2.43 (0.01)	2.39 (0.01)
(0.25, 0.25)		2.00 (0.01)	2.04 (0.01)	2.32 (0.01)	2.28 (0.01)	2.42 (0.01)	2.35 (0.01)	2.45 (0.01)	2.39 (0.01)
(0.80, 0.50)	1.0	1.98 (0.01)	2.08 (0.01)	2.17 (0.01)	2.24 (0.01)	2.29 (0.01)	2.31 (0.01)	2.40 (0.01)	2.36 (0.01)
(0.80, 0.80)		1.97 (0.01)	2.10 (0.01)	2.15 (0.01)	2.28 (0.01)	2.31 (0.01)	2.39 (0.01)	2.49 (0.01)	2.45 (0.01)
(0.87, 0.82)		1.95 (0.01)	2.09 (0.01)	2.11 (0.01)	2.25 (0.01)	2.25 (0.01)	2.35 (0.01)	2.43 (0.01)	2.42 (0.01)
(0.00, 0.00)		2.21 (0.01)	2.18 (0.01)	2.52 (0.01)	2.40 (0.01)	2.59 (0.01)	2.46 (0.01)	2.57 (0.01)	2.51 (0.01)
(0.25, 0.25)		2.20 (0.01)	2.20 (0.01)	2.51 (0.01)	2.41 (0.01)	2.59 (0.01)	2.48 (0.01)	2.59 (0.01)	2.51 (0.01)
(0.80, 0.50)	1.5	2.14 (0.01)	2.21 (0.01)	2.34 (0.01)	2.36 (0.01)	2.45 (0.01)	2.42 (0.01)	2.56 (0.01)	2.49 (0.01)
(0.80, 0.80)		2.13 (0.01)	2.24 (0.01)	2.31 (0.01)	2.41 (0.01)	2.48 (0.01)	2.51 (0.01)	2.64 (0.01)	2.57 (0.01)
(0.87, 0.82)		2.11 (0.01)	2.23 (0.01)	2.27 (0.01)	2.39 (0.01)	2.41 (0.01)	2.47 (0.01)	2.59 (0.01)	2.54 (0.01)
(0.00, 0.00)		2.45 (0.01)	2.37 (0.01)	2.77 (0.02)	2.58 (0.01)	2.80 (0.01)	2.62 (0.01)	2.76 (0.01)	2.67 (0.01)
(0.25, 0.25)		2.41 (0.01)	2.37 (0.01)	2.73 (0.02)	2.57 (0.01)	2.80 (0.01)	2.62 (0.01)	2.76 (0.01)	2.66 (0.01)
(0.80, 0.50)	2.0	2.37 (0.01)	2.41 (0.01)	2.59 (0.01)	2.56 (0.01)	2.70 (0.01)	2.61 (0.01)	2.78 (0.01)	2.65 (0.01)
(0.80, 0.80)		2.33 (0.01)	2.42 (0.01)	2.54 (0.01)	2.59 (0.01)	2.71 (0.01)	2.67 (0.01)	2.87 (0.02)	2.73 (0.01)
(0.87, 0.82)		2.32 (0.01)	2.41 (0.01)	2.49 (0.01)	2.56 (0.01)	2.61 (0.01)	2.62 (0.01)	2.80 (0.01)	2.69 (0.01)
(0.00, 0.00)		4.30 (0.05)	3.93 (0.04)	4.69 (0.05)	4.06 (0.04)	4.68 (0.05)	4.14 (0.04)	4.38 (0.04)	4.11 (0.04)
(0.25, 0.25)		4.30 (0.05)	3.97 (0.04)	4.64 (0.05)	4.05 (0.04)	4.60 (0.05)	4.08 (0.04)	4.42 (0.04)	4.12 (0.04)
(0.80, 0.50)	5.0	4.19 (0.04)	4.05 (0.04)	4.44 (0.05)	4.09 (0.04)	4.53 (0.05)	4.09 (0.04)	4.58 (0.05)	4.15 (0.04)
(0.80, 0.80)		4.04 (0.04)	3.97 (0.04)	4.40 (0.05)	4.15 (0.04)	4.58 (0.05)	4.15 (0.04)	4.68 (0.05)	4.14 (0.04)
(0.87, 0.82)		4.02 (0.04)	3.97 (0.04)	4.24 (0.04)	4.06 (0.04)	4.54 (0.05)	4.17 (0.04)	4.74 (0.05)	4.20 (0.04)

(ρ_1, ρ_2), correlation coefficients for two blocks with AR(1)-Type variance-covariance structures, respectively;
p,the number of predictor variables;
n, the sample size;
PLS, partial least squares regression;
PLSM, modified partial least squares implementing mixed-effect;
The values in the parentheses were the standard errors.
When the PLSM method showed a significantly better result (p-value < 0.0001), red-faced type was applied.

For simplicity, we assumed that only two predictors have significant effects (namely coefficients, 2 and 0.5), each of which was affiliated to each block individually. For the rest of predictors, coefficients were set at 0. A random-effect with four levels was generated for different values of σ_a = 1, 1.5, 2, and 5. Finally, we generated responses $\mathbf{y}$ under the linear mixed-effect model, $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\alpha + \epsilon$, where σ_ϵ = 1 or 2.

For each simulation setting, we performed 1,000 runs of simulations, and computed the root mean squared error of prediction (RMSEP) as a measure for the prediction performance. We also used the RMSEP to compare the proposed PLSM method against the conventional PLS method. In simulation studies, two PLS components were identified for both PLS and PLSM. The R packages **mnormt** and **pls** were used for generation and analysis of simulation data in the R environment (Azzalini *et al.*, 2021; Liland *et al.*, 2021; R Development Core Team, 2021).

4.2. Results and discussion

In the simulation results (Tables 1 and 2), we first observed that PLSM showed significantly better performance than the conventional PLS as the random-effect becomes large. However, this pattern weakened when multi-collinearity becomes strong, and reversed with small random-effect and strong multi-collinearity. Our conjecture for this case was that the control of severe multi-collinearity might be more crucial to prediction than the handling of a random-effect. Thus, PLSM seems less beneficial than PLS when multi-collinearity is very strong and random-effect is not large. We cautiously expect that PLSM may perform better than the conventional PLS if the model includes a larger number of random effects.

Second, the PLSM method demonstrated a considerable benefit compared to PLS when the num-

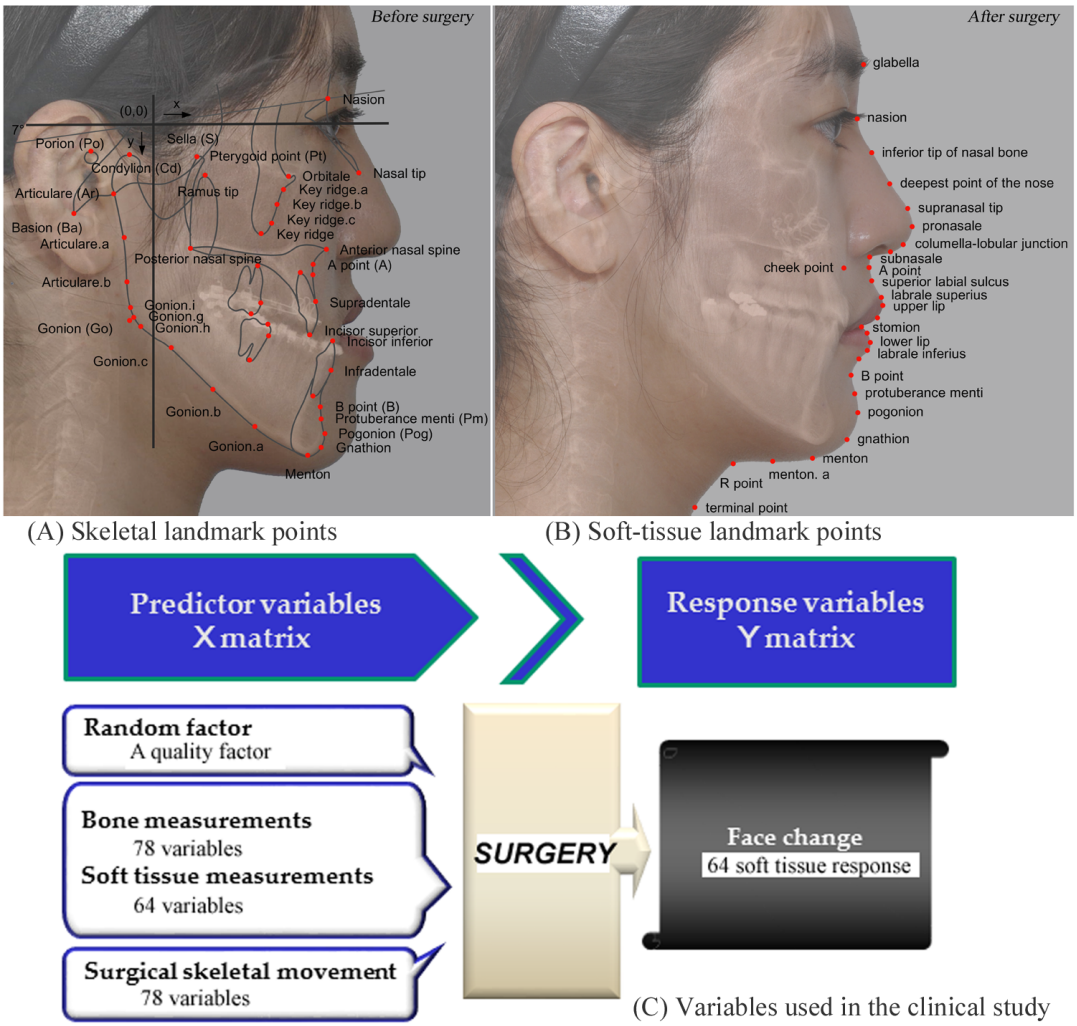

Figure 1: *The bone (A) and soft-tissue landmark points (B). There were 220 predictor- and 64 response variables. A random factor had two levels of a quality factor (C).*

ber of predictor variables (p) increased. Especially with $p = 500$, PLSM showed significantly lower prediction errors than PLS did. Third, compared to PLS, PLSM showed more accurate prediction when the random error was large, namely $\sigma_\epsilon = 2$ than when $\sigma_\epsilon = 1$.

In summary, the PLSM results showed better predictive performance than the conventional PLS method did 1) when the random-effect was greater, 2) the number of predictors is larger, especially exceeds the sample size, 3) the multi-collinearity is not so strong, and 4) the random noise is considerable.

As a limitation, the simulation of the current study set up a univariate response variable. It should be noted that results for multivariate response variables could have been different.

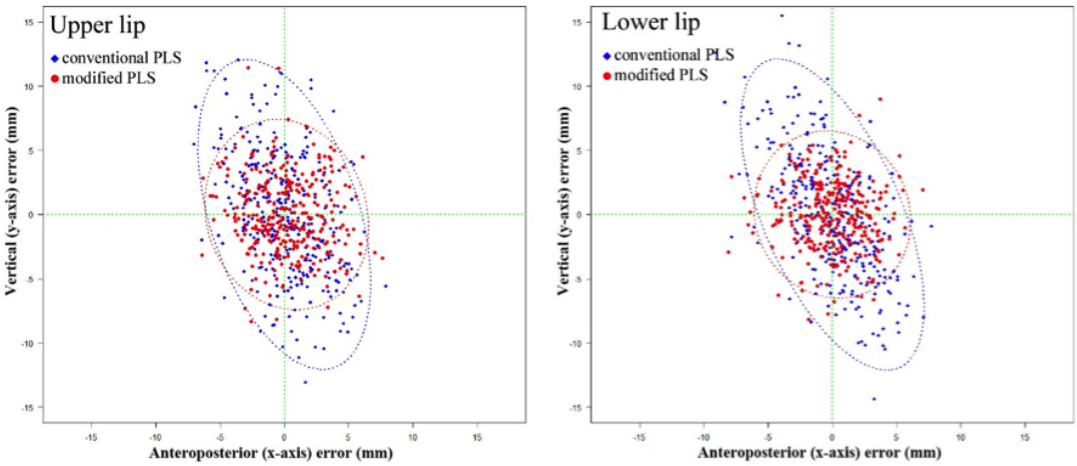

Figure 2: *Scattergrams and 95% confidence ellipses for the prediction errors in the upper- and lower lip position. The errors were obtained from the PLS (blue) and PLSM (red) methods. The plots indicated that the PLS method demonstrated larger prediction errors than the PLSM prediction did.*

5. Clinical data application

In order to illustrate our new proposal, we applied PLSM to a real clinical dataset, to which the conventional PLS method was previously applied (Suh *et al*., 2012; Lee *et al*., 2014; Suh *et al*., 2019). We compared the prediction accuracy between PLSM and PLS. Some benefits of PLSM over the conventional PLS methods are also discussed.

5.1. Characteristics of the data structure and the formulation of the prediction study

The dataset consisted of the cephalometric measurements for n = 318 subjects who underwent facial cosmetic surgeries. Figure 1(A) and (B) illustrate the anthropological landmarks detected on lateral human face images for bones (39 landmark points) and soft tissues (32 landmark points), respectively. All variables indicate the cartesian coordinates in the horizontal-(x-axis) and vertical (y-axis) directions. The number of predictor variables was 220: 142 pre-existing characteristics of bones and soft-tissues; and 78 variables indicating the amount surgical skeletal repositioning. The face change after surgery was characterized with 64 soft-tissue response variables.

We considered a quality factor with two categorical levels, which can accommodate the surgeon-to-surgeon variability with a 2-surgeon case, the procedure-to-procedure variability with two types of minute surgical procedures, or the surgeons' preference to two surgical procedures (Figure 1(C)). This factor was assumed to have a random-effect because the effect of its particular levels is not of as much interest to us as is the amount of variation in the response that can be attributed to its different levels. Note that the optimal number of PLS components was determined by the leave-one-out cross-validation method (LOOCV) and the PLS model with 30 components was selected. The test error was also calculated using LOOCV.

5.2. Results and discussion

When predicting the face shape after surgery, it is often of interest to represent the pattern of errors in a graphic fashion (Donatelli and Lee, 2013). For display purpose, we reported the prediction errors

for each patient only at the upper- and lower lip position (Figure 2). A negative value of the error indicates that the predicted value is more posterior in the x-axis or more superior in the y-axis when compared to the actual/true value. The contours stand for the 95% confidence ellipse of the prediction errors, each of which satisfies $(\mathbf{x}-\mu)^\top\Sigma^{-1}(\mathbf{x}-\mu) \leq \chi^2_{0.05}(2)$ where $\mathbf{x}$ is the 2-dimensional (x- and y-axes) vectors for the errors; μ is the mean vector for $\mathbf{x}$; $\Sigma = (\sigma_1, \sigma_2; \rho_{12})$ is the covariance matrix; and $\chi^2_\alpha(df)$ is the upper $100(1-\alpha)^{th}$ percentile of the chi-square distribution with df degrees of freedom (Johnson and Wichern, 2007).

In the Figure 2, scattergrams and 95% confidence ellipses demonstrated that the pattern of prediction error is more favorable for PLSM than for PLS. The size of the 95% confidence ellipse for the PLS method was greater than that for PLSM. We also observed that the shape of the confidence ellipse was closer to a circle for PLSM ($\hat{\rho}_{12} = -0.12$ and -0.08 at upper and lower lip positions, respectively) than for PLS ($\hat{\rho}_{12} = -0.44$ and -0.59) while both ellipses were located almost at the origin ($\hat{\mu} \approx \mathbf{0}$). It might indicate that PLSM deals with the correlation structure between x- and y-axes better than PLS although a thorough investigation is further needed.

Even if the sample size in the real clinic dataset was relatively small, the PLSM method appeared suitable for exploring relationships between multiple predictor and response variables with a random factor. We hope that this method would be valuable in numerous situations where variables are highly inter-correlated within an individual subject or each characteristic group.

6. Concluding remarks

A feasible computational method was developed to formulate a modified partial least squares regression by relating it to the mixed-effect analysis. An advantage of the proposed PLSM model might be its ability to handle multivariate responses in a serial longitudinal data consisted of repeated measurements and/or to take the random-effect of a randomized block design into account. The proposed PLSM would be particularly beneficial when a random-effect exists and its effect is conspicuous; the number of predictors is relatively greater than the sample size; the multi-collinearity is weak or moderate; and/or the random error is large.

Acknowledgments

The data presented in the present study were part of a doctoral dissertation (SJL).

References

Azzalini A, Genz A, Miller A, Wichura MJ, Hill GW, and Ge Y (2021). *Mnormt: The multivariate normal and t-distributions, and their truncated versions*, Available from: http://CRAN.R-project.org/package=mnormt

Chun H and Keles S (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**, 3–25.

Dai JJ, Lieu L, and Rocke D (2006). Dimension reduction for classification with gene expression microarray data, *Statistical Applications in Genetics and Molecular Biology*, **5**, 6.

Donatelli RE and Lee SJ (2013). How to report reliability in orthodontic research. Part 2, *American Journal of Orthodontics and Dentofacial Orthopedics*, **144**, 315–318.

Fordellone M and Vichi M (2020). Finding groups in structural equation modeling through the partial least squares algorithm, *Computational Statistics & Data Analysis*, **147**, 106957.

Garthwaite PH (1994). An interpretation of partial least-squares, *Journal of the American Statistical Association*, **89**, 122–127.

Hastie T, Tibshirani R, and Friedman J (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction (2nd ed)*, Springer Verlag, New York.

Hwang HW, Moon JH, Kim MG, Donatelli RE and Lee SJ (2021). Evaluation of automated cephalometric analysis based on the latest deep learning method, *The Angle Orthodontist*, **91**, 329–335.

Johnson RA and Wichern DW (2007). *Applied Multivariate Statistical Analysis*, Pearson Prentice Hall, New Jersey.

Krishnan A, Williams LJ, McIntosh AR, and Abdi H (2011). Partial least squares (PLS) methods for neuroimaging: A tutorial and review, *Neuroimage*, **56**, 455–475.

Laird NM and Ware JH (1982). Random-effects models for longitudinal data, *Biometrics*, **38**, 963–974.

Lee D, Lee W, Lee Y, and Pawitan Y (2010). Super-sparse principal component analyses for high-throughput genomic data, *BMC Bioinformatics*, **11**, 1–10.

Lee YS, Suh HY, Lee SJ, and Donatelli RE (2014). A more accurate soft-tissue prediction model for Class III 2-jaw surgeries, *American Journal of Orthodontics and Dentofacial Orthopedics*, **146**, 724–733.

Liland KH, Mevik BH, Wehrens R, and Hiemstra P (2021). *PLS: partial least squares and principal component regression*. R package version 2.8-0, Available from: http://CRAN.R-project.org/package=pls

Martins JPA, Teofilo RF, and Ferreira MMC (2010). Computational performance and cross-validation error precision of five PLS algorithms using designed and real data sets, *Journal of Chemometrics*, **24**, 320–332.

Mevik BH and Wehrens R (2007). The pls package: Principal component and partial least squares regression in R, *Journal of Statistical Software*, **18**, 1–24.

R Development Core Team (2021). *R: A language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna.

Suh HY, Lee SJ, Lee YS, Donatelli RE, Wheeler TT, Kim SH, Eo SH, and Seo BM (2012). A more accurate method of predicting soft-tissue changes after mandibular setback surgery, *Journal of Oral and Maxillofacial Surgery*, **70**, e553–e562.

Suh HY, Lee HJ, Lee YS, Eo SH, Donatelli RE, and Lee SJ (2019). Predicting soft-tissue changes after orthognathic surgery: The sparse partial least squares method, *The Angle Orthodontist*, **89**, 910–916.

Wehrens R (2011). *Chemometric with R: Multivariate Data Analysis in the Natural Sciences and Life Sciences* (1st ed), Springer, Heidelberg.

Zhou XF, Shao Q, Coburn RA, and Morris ME (2005). Quantitative structure-activity relationship and quantitative structure-pharmacokinetics relationship of 1,4-dihydropyridines and pyridines as multidrug resistance modulators, *Pharmaceutical Research*, **22**, 1989–1996.

Received April 21, 2022; Revised August 16, 2022; Accepted September 25, 2022