General Psychiatry

Partial least squares regression and principal component analysis: similarity and differences between two popular variable reduction approaches

Chenyu Liu,[1] Xinlian Zhang,[1] Tanya T Nguyen,[2] Jinyuan Liu,[1] Tsungchin Wu,[1] Ellen Lee,[2] Xin M Tu[1]

To cite: Liu C, Zhang X, Nguyen TT, *et al*. Partial least squares regression and principal component analysis: similarity and differences between two popular variable reduction approaches. *General Psychiatry* 2022;**35**:e100662. doi:10.1136/gpsych-2021-100662

Received 03 November 2021
Accepted 29 December 2021

[1]Division of Biostatistics and Bioinformatics, Herbert Wertheim School of Public Health and Human Longevity Science, UC San Diego, La Jolla, California, USA
[2]Department of Psychiatry, Stein Institute for Research on Aging, UC San Diego, La Jolla, California, USA

Correspondence to
Dr Chenyu Liu;
chl056@health.ucsd.edu

SUMMARY

In many statistical applications, composite variables are constructed to reduce the number of variables and improve the performances of statistical analyses of these variables, especially when some of the variables are highly correlated. Principal component analysis (PCA) and factor analysis (FA) are generally used for such purposes. If the variables are used as explanatory or independent variables in linear regression analysis, partial least squares (PLS) regression is a better alternative. Unlike PCA and FA, PLS creates composite variables by also taking into account the response, or dependent variable, so that they have higher correlations with the response than composites from their PCA and FA counterparts. In this report, we provide an introduction to this useful approach and illustrate it with data from a real study.

INTRODUCTION

Composite variables are widely used to summarise information from a set of outcomes in statistical analysis. In some studies, composite variables are used to create domain scales or subscales, such as the SF-36 (the MOS item short-form health survey) instrument, while in some other studies, composite variables are used to deal with limitations of data. For example, in regression analysis, we may need to create composite variables if the number of explanatory or independent variables is larger than the sample size. This statistical issue arises when modelling high-throughput data such as in fitting regression models to determine associations of brain functions with behavioural and health outcomes of interest due to large numbers of braining imaging variables and limited study sample sizes in most studies. Principal component analysis (PCA) and factor analysis (FA) are generally used for creating composite variables. In this report, we describe another less-known approach called partial least squares (PLS) regression, to create composite variables and discuss scenarios where this approach is more effective than PCA and FA. We illustrate this approach with a real-life application to research data.

PARTIAL LEAST SQUARES REGRESSION

As noted earlier, PCA and FA are two popular approaches for creating composite variables. Under PCA, a set of ordered composite variables are created to represent the original set of outcomes. Each composite variable is a linear combination, or a weighted sum, of the original outcomes. The coefficients, or weights, of the linear combination in each compositive variable, are called loadings, and their signs and magnitudes indicate the directions and contributions of the corresponding variables. Unlike the original outcomes, the composite variables are orthogonal to each other. Moreover, the first composite variable has the largest variance, followed by the second and so on. FA also creates a set of composite variables. However, unlike PCA, FA composite variables are not ordered in the sense of PCA composites and are not orthogonal to each other. Instead, loadings of the FA composites can be used to group the original variables to create subscale, or domain scales, for different constructs such as the domains of Physical Functioning and Emotional Wellbeing in the SF-36.[1]

PCA and FA create composite variables for general purposes. When composite variables are used as explanatory or independent variables in regression analysis involving a response or dependent variable, a more effective approach is PLS. Like PCA, PLS composite variables are also ordered. However, unlike PCA, PLS composite variables are ordered by their correlations with the response in the regression model; the first composite variable

has the maximum correlation with the response, followed by the second and so on. If interest lies in finding a subset of the original explanatory variables in the linear model that explains the most variability in the response, PLS composite variables are more effective than PCA.

To describe in detail how to compute PLS composite variables, consider a linear regression with a continuous response of interest, Y, and a set of p explanatory variables, $X_1, X_2, \ldots, X_p$. We are interested in modelling the relationship of Y with $X_1, X_2, \ldots, X_p$. Given a sample of n subjects, the classic linear regression relating Y to $X_1, X_2, \ldots, X_p$ is given by:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad 1 \le i \le n, \tag{1}$$

where i indexes the subjects, $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ denote the regression parameters, ε_i is the error term, and $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2 . The first part of the linear regression,

$$\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}, \quad 1 \le i \le n, \tag{2}$$

is called the conditional (population) mean of Y_i given the explanatory variables $X_{i1}, X_{i2}, \ldots, X_{ip}$. On estimating the regression parameters $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$, this conditional mean describes the association of Y_i with each of the explanatory variables.

When the explanatory variables $X_{i1}, X_{i2}, \ldots, X_{ip}$ are highly correlated, estimates of $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ may not be reliable due to multicollinearity using the standard least squares (LS) or maximum likelihood (ML) method. In studies of high-throughput data, the number of explanatory variables exceeds sample sizes, in which case LS method will not apply. In both cases, we need to reduce the number of the variables, $X_{i1}, X_{i2}, \ldots, X_{ip}$, or the dimension p. There are different approaches to address this issue. One may use the least absolute shrinkage and selection operator (LASSO) to determine a subset of $X_{i1}, X_{i2}, \ldots, X_{ip}$ that provides reliable associations with Y_i. Alternatively, one may create composite variables $Z_{i1}, Z_{i2}, \ldots, Z_{ip}$ from $X_{i1}, X_{i2}, \ldots, X_{ip}$ and use a subset of the composite variables to predict Y_i. The latter composite variable approach is preferred if some or all $X_{i1}, X_{i2}, \ldots, X_{ip}$ need to work together to explain the variability in Y_i. For example, if one wants to predict areas of rectangles, lengths or widths alone will not provide reliable predictions since a rectangle with a very large length can still have a small area if it has a small width. LASSO is most effective to deal with high-throughput data as dimension is the primary problem in this case. In the presence of multicollinearity, it is likely more meaningful to aggregate information in correlated variables using a subset of composite variables, rather than to select a subset of the original variables. In this case, correlated variables may all contribute to explaining the variability in the response Y_i and composite variables will account for all such contributions.

The composite variables $Z_{i1}, Z_{i2}, \ldots, Z_{ip}$ for PLS are obtained by solving an optimisation problem.[2] Unlike PCA composite variables, PLS finds directions of the composite variables $Z_{i1}, Z_{i2}, \ldots, Z_{ip}$ that have both high variance and high correlation with the response Y_i. Specifically, let Z_{il} denote the lth composite variable:

$$Z_{il} = X_{i1}\alpha_{l1} + X_{i2}\alpha_{l2} + \cdots + X_{ip}\alpha_{lp}, \quad 1 \le i \le n, \quad 1 \le l \le p, \tag{3}$$

where $\alpha_1, \alpha_2, \ldots, \alpha_p$ denote weights, or loadings, of the composite Z_l. We can also express (3) equivalently in a vector form

$$\widetilde{Z}_l = \widetilde{X}_1\alpha_{l1} + \widetilde{X}_2\alpha_{l2} + \cdots + \widetilde{X}_p\alpha_{lp}, \quad 1 \le i \le n, \quad 1 \le l \le p, \tag{4}$$

or in a matrix form:

$$\widetilde{Z}_l = \boldsymbol{X}\widetilde{\alpha}_l, \qquad 1 \le l \le p,$$

where $\widetilde{Z}_l = (Z_{1l}, Z_{2l}, \cdots, Z_{nl})^T$, $\widetilde{X}_l = (X_{1l}, X_{2l}, \cdots, X_{nl})^T$ and $\widetilde{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_p)^T$ denote $p \times 1$ column vectors, and $X = \left(\widetilde{X}_1, \widetilde{X}_2, \cdots, \widetilde{X}_n\right)$ denotes an $n \times p$ matrix. The loadings are determined by the following optimisation procedure:

$$max_{\alpha_l} Corr^2\left(\widetilde{Y}, X\widetilde{\alpha}_l\right) Var\left(X\widetilde{\alpha}_l\right)$$

$$\text{subject to: } \|\widetilde{\alpha}_l\| = \sqrt{\alpha_{l1}^2 + \alpha_{l2}^2 + \cdots + \alpha_{lp}^2} = 1, \widetilde{\alpha}^T{}_l S\widetilde{\alpha}_m = 0, m = 1, \cdots, l-1.$$

where $\widetilde{Y} = (Y_1, Y_2, \cdots, Y_n)^T$ is a $n \times 1$ column vector, S is the sample covariance matrix of $X_{i1}, X_{i2}, \ldots, X_{ip}$, $Corr^2\left(\widetilde{Y}, X\widetilde{\alpha}_l\right)$ denotes the squared (Pearson) correlation matrix between $\widetilde{Y}$ and $X\widetilde{\alpha}_l$, and $Var\left(X\widetilde{\alpha}_l\right)$ denotes the sample variance of $X\widetilde{\alpha}_l$. The condition $\widetilde{\alpha}_l^T S\widetilde{\alpha}_m = 0$ ensures that the lth composite $\widetilde{Z}_l = X\widetilde{\alpha}_l$ is uncorrelated with all previous composite variables $\widetilde{Z}_m = X\widetilde{\alpha}_m$ $(1 \le m < l \le p)$.

In practice, we can use the following procedure to find the PLS composite variables.[3] We start by standardising each of the original explanatory variables $X_{i1}, X_{i2}, \ldots, X_{ip}$ to have mean 0 and variance 1. Set $\widetilde{Y}^{(0)} = \bar{Y}1_n$ and $\widetilde{X}_l^{(0)} = \widetilde{X}_l$ $(1 \le l \le p)$, where $\bar{Y} = \frac{1}{n}\sum_{i=1}^n (Y_1 + Y_2 + \cdots + Y_n)$ denotes the sample mean of Y_i $(1 \le i \le n)$ and $\widetilde{1}_n = (1, 1, \cdots, 1)^T$ denotes a $n \times 1$ volume vector of 1. For $1 \le l \le p$, we perform the following steps:

(a) $\widetilde{Z}_l = \widetilde{X}_1^{(l-1)}\alpha_{l1} + \widetilde{X}_2^{(l-1)}\alpha_{l2} + \cdots + \widetilde{X}_p^{(l-1)}\alpha_{lp}$, where $\alpha_{lj} = \left\langle \widetilde{X}_j^{(l-1)}, \widetilde{Y} \right\rangle$, where $\langle a, b \rangle$ denotes the inner product between two vectors a and b;

(b) $\theta_l = \dfrac{\left\langle \widetilde{Z}_l, \widetilde{Y} \right\rangle}{\left\langle \widetilde{Z}_l, \widetilde{Z}_l \right\rangle}$;

(c) $\widetilde{Y}^{(l)} = \widetilde{Y}^{(l-1)} + \theta_l \widetilde{Z}_l$;

(d) Orthogonalise each $\widetilde{X}_j^{(l-1)}$ with respect to $\widetilde{Z}_l$:

$$\widetilde{X}_j^{(l)} = \widetilde{X}_j^{(l-1)} - \frac{\left\langle \widetilde{Z}_l, \widetilde{X}_j^{(l-1)} \right\rangle}{\left\langle \widetilde{Z}_l, \widetilde{Z}_l \right\rangle} \widetilde{Z}_l, \qquad 1 \le j \le p.$$

To illustrate the difference between PSL and PCA, here is the procedure to compute composite variables under PCA:

We start with the $n \times p$ data matrix X, which is formed by the column vectors X_1, X_2,…, X_p, that is, $X = [X_1, X_2, \ldots, X_k, \ldots, X_p]$. Then we perform the following steps:

1. Average over all the columns of X: $\bar{X} = \frac{1}{p}\sum_{k=1}^{p} X_k$;
2. Centre the matrix X at this average $\bar{X}$ by subtracting $\bar{X}$ from each column vector X_k of X, denote as: $Z = [X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_p - \bar{X}]$;
3. Compute the sample variance–covariance matrix: $\Sigma = Z^T Z$;
4. Compute the eigenvalues $\{\lambda_k\}_{k=1,2\ldots,p}$ and corresponding eigenvectors $\{U_k\}_{k=1,2\ldots,p}$ of Σ with $\lambda_1 > \lambda_2 > \ldots > \lambda_p$;
5. The top m ($\leq p$) principal components, or composite variables, $\{U_k\}_{k=1,2\ldots,m}$, are then used as independent variables in linear regression models with Y as the dependent variable, where m is generally determined by the magnitude of the sum of the top m eigenvalues relative to the sum of all p eigen values, $\frac{\sum_{k=1}^{m}\lambda_k}{\sum_{l=1}^{p}\lambda_l}$, which has the interpretation of being the percent of the variability of X explained by the top m eigenvectors $\{U_k\}_{k=1,2\ldots,m}$.

By comparing the two procedures, we can see that PCA creates composite variables without using any information in the dependent variable Y as PLS does in creating its composite variables. If the goal is to find composite variables of X that are most predictive of Y, PLS is more preferrable than PCA. On the other hand, if the goal is to find composite variables that maximally explain the variability of the data matrix X, then PCA is more preferrable.

REAL STUDY EXAMPLE

We illustrate PLS with data from a recent study on the association of loneliness and wisdom with gut microbial diversity and composition.[4] Loneliness and wisdom have opposite effects on health and well-being. Loneliness is a serious public health problem associated with increased morbidity and mortality. Wisdom is associated with better health and well-being. Nguyen *et al*[4] successfully applied PLS to demonstrate relationships between the association of loneliness and wisdom with alpha-diversity. We use this study to illustrate the advantages of PLS over standard linear regression. More details about the study population, measures of loneliness, wisdom, gut microbial diversity and other outcomes, and additional findings can be found in the paper.

The study included 184 community-dwelling adults (28–97 years). Participants completed validated scales of loneliness (UCLA Loneliness Scale),[5] wisdom (including cognitive, affective and reflective dimensions; Three-Dimensional Wisdom Scale),[6] compassion (Santa Clara Brief Compassion Scale),[7] social support (Emotional Support Scale)[8] and social engagement (Cognitively Stimulating Questionnaire).[9] These variables are interrelated; loneliness and wisdom have strong inverse correlations; social support, social engagement and loneliness are often inversely correlated, but they are distinct concepts. Faecal samples were obtained from participants using at-home self-collection kits and returned via mail. Alpha- diversity is the ecological diversity (ie, richness, evenness, compositional complexity) of a single sample and was quantified using Faith's Phylogenetic Diversity (PD) based on the DNAs extracted from the faecal samples. It measures the total length of branches in a reference phylogenetic tree for all species in a given sample.[10]

Table 1 Results from linear regression for association of alpha-diversity (Faith's Phylogenetic Diversity) with loneliness and wisdom outcomes, controlling for covariates

Predictors/covariates	β	t	P value
Intercept	11.492	1.52	0.131
Loneliness	0.022	0.40	0.686
Social support	1.314	1.37	0.172
Wisdom components			
Affective	0.869	0.73	0.473
Cognitive	−1.164	−1.10	0.266
Reflective	1.350	1.03	0.302
Compassion	0.270	0.67	0.499
Social engagement	0.312	0.43	0.671
Age	−0.013	−0.45	0.652
BMI	−0.111	−1.37	0.168

BMI, body mass index.

We first fit a standard linear regression to model the association of alpha-diversity with individual loneliness, wisdom, compassion, social support and social engagement outcomes as predictors, controlling for age and body mass index (BMI). Shown in table 1 were estimated regression coefficients (β) for the predictors and covariates, along with associated t-statistics (t) and p values . As seen, none of the predictors were significant.

We then applied PLS to construct composite variables from all the predictors and included the extracted composite variables and the covariates to build the linear regression to predict alpha-diversity by examining the contribution of each composite component added in terms of the amount of explained variability in the outcome of alpha-diversity.[11] We settled on the first two composite variables because adding component 3 led to a decreased adjusted R squared. Shown in table 2 are estimated regression coefficients (β) for the predictors and covariates, along with t-statistics (t) and p values. The model revealed that the effect of component 1 was significantly positively associated with alpha-diversity (p=0.008), whereas component 2 was not (p=0.217).

Table 2 Coefficients from linear regression model of partial least squares (PLS) composite variables predicting alpha-diversity (Faith's Phylogenetic Diversity), controlling for age and BMI

	β	t	P value
Intercept	21.911	8.52	<0.001
Component 1	0.717	2.71	0.008
Component 2	0.545	1.24	0.217
Age	−0.015	−0.57	0.569
BMI	−0.103	−1.32	0.188

BMI, body mass index.

Table 4 Coefficients from linear regression model of principal component analysis (PCA) composite variables predicting alpha-diversity (Faith's Phylogenetic Diversity), controlling for age and BMI

	β	t	P value
Intercept	21.940	8.48	<0.001
Component 1	0.639	2.47	0.015
Component 2	0.500	1.32	0.190
Age	−0.021	−0.78	0.439
BMI	−0.092	−1.18	0.240

BMI, body mass index.

When applying PLS, it is important to determine directions of effects for the original predictors of interest (loneliness, wisdom, compassion, social support and social engagement) when a composite variable is used as a predictor in the final model. Shown in table 3 were loadings of individual predictors on the first two composite variables. The sign of the loading of a predictor on the composite variable indicates the direction of association of the predictor with the composite variable. Except for loneliness, all the predictors had positive loadings on the first composite variable, indicating that wisdom, compassion, social support and social engagement had positive associations with alpha-diversity. Loneliness had a negative association with alpha-diversity because of its negative sign. The first composite variable also accounted for 40% of the total variability of the psychosocial variables.

To illustrate the differences between PLS and PCA, we also applied PCA to construct composite variables and use them as explanatory variables in modelling the association of alpha-diversity with the psychosocial variables. To be consistent with the PLS, we used the first two eigenvectors as the composite variables and controlled for age and BMI. Shown in table 4 were estimated regression coefficients (β) for the predictors and covariates, along with t-statistics (t) and p values (component 1: p=0.015; component 2: p=0.190). The results were similar to their PLS counterparts. A notable difference is the slightly weaker association between the first composite variable and alpha-diversity. Both PCA and PLS yielded the same conclusion regarding the association of composite variables with alpha-diversity.

Shown in table 5 were loadings of individual predictors on the first two PCA composite variables. The signs of the loadings are consistent with their PLS counterparts. The wisdom-cognitive subscore had less loading under PLS than PCA, while compassion, social support and social engagement had higher loadings under PLS than PCA.

DISCUSSION

In this report, we described the partial least squares (PLS) regression, discussed its relationship with a closely related alternative, the principal component analysis (PCA), and illustrated the PLS with a real study example. Although both aim to reduce explanatory variables (predictors), PLS and PCA work quite differently in developing composite variables. While PCA constructs the composite variables to explain the maximum variability in all the original predictors, or the explanatory variables of interest, PLS creates its composite variables to explain the maximum variability in the response within the context of linear regression.

In practice, if the goal is to develop a set of composite variables for use as explanatory variables in regression models for multiple responses, PCA may be preferred since, unlike PLS, it will create a common set of composite variables for regression across all the responses. On the other hand, if the objective is to develop a set of composite variables to explain the maximum variability for a given

Table 3 Loadings for PLS composite variables

	Composite variable 1	Composite variable 2
Loneliness	−0.419	0.272
Wisdom-cognitive	0.233	−0.836
Wisdom-reflective	0.462	−0.370
Wisdom-affective	0.454	−0.140
Compassion	0.417	0.421
Social support	0.316	0.187
Social engagement	0.358	0.233

Table 5 Loadings for the first PCA composite variables

	Composite variable 1	Composite variable 2
Loneliness	−0.423	0.071
Wisdom-cognitive	0.317	−0.601
Wisdom-reflective	0.483	−0.253
Wisdom-affective	0.442	−0.037
Compassion	0.342	0.528
Social support	0.280	0.058
Social engagement	0.303	0.536

response, then PLS should be used. When applying PLS to develop composite variables for regression analysis for multiple responses, multiple sets of composite variables will be created with one set for each response and consequently, regression results from composite variables must be interpreted with respect to factor loadings within each set of composite variables.

In the illustrative example, the two approaches yield similar results. In general, results from the two approaches may differ and yield different conclusions. For example, PLS may yield significant associations of its composite variables while PCA does not. If interest lies in finding associations of a response with a set of explanatory variables, PLS should be used.

Contributors CL conceived the initial idea, searched the literature on related topics, performed analyses and assisted in manuscript preparation. XZ participated in the discussion of the statistical problems and helped with technical details of PLS, and helped finalise the manuscript. TTN brought the statistical problem in the real study example, participated in the discussion of the statistical problems, helped in the interpretation of estimates in the real study example and helped with technical details of least squares, and helped finalise the manuscript. JL, TW, XMT researched the statistical issues, directed simulation studies, drafted parts of the manuscript and finalised the manuscript. All authors provided critical feedback and helped shape the research, analysis and manuscript.

Funding This study was funded by National Institutes of Health (UL1TR001442).

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval The studies involving human participants were reviewed and approved by UCSD Human Research Protections Program. The patients/participants provided their written informed consent to participate in this study. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Commissioned; externally peer reviewed.

REFERENCES

1 Hays RD, Sherbourne CD, Mazel RM. The RAND 36-item health survey 1.0. *Health Econ* 1993;2:217–27.
2 Garthwaite PH. An interpretation of partial least squares. *J Am Stat Assoc* 1994;89:122–7.
3 Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning. Springer series in statistics*. New York, NY, USA: Springer New York Inc, 2001.
4 Nguyen TT, Zhang X, Wu T-C, *et al*. Association of loneliness and wisdom with gut microbial diversity and composition: an exploratory study. *Front Psychiatry*;12:648475.
5 Russell DW. UCLA loneliness scale (version 3): reliability, validity, and factor structure. *J Pers Assess* 1996;66:20–40.
6 Ardelt M. Empirical assessment of a three-dimensional wisdom scale. *Res Aging* 2003;25:275–324.
7 Hwang JY, Plante T, Lackey K. The development of the Santa Clara brief compassion scale: an abbreviation of Sprecher and Fehr's compassionate love scale. *Pastoral Psychol* 2008;56:421–8.
8 Seeman TE, Lusignolo TM, Albert M, *et al*. Social relationships, social support, and patterns of cognitive aging in healthy, high-functioning older adults: MacArthur studies of successful aging. *Health Psychol* 2001;20:243–55.
9 Krueger KR, Wilson RS, Kamenetsky JM, *et al*. Social engagement and cognitive function in old age. *Exp Aging Res* 2009;35:45–60.
10 Faith DP. Conservation evaluation and phylogenetic diversity. *Biol Conserv* 1992;61:1–10.
11 Mevik B-H, Wehrens R. The pls package: principal component and partial least squares regression in *R*. *J Stat Softw* 2007;18.

Chenyu Liu is a PhD student in Division of Biostatistics and Bioinformatics, Herbert Wertheim School of Public Health and Human Longevity Science, UC San Diego in USA. She is currently working as a Graduate Student Researcher at UC San Diego. She got her master's degree in statistics from University of Minnesota, Twin Cities in USA. Her main research interests include statistical learning, statistical methods, clinical trial and causal inference.