International Journal of Development Mathematics Vol 2 Issue 4 Page No. 245 - 253

INTERNATIONAL JOURNAL OF DEVELOPMENT MATHEMATICS

ISSN: 3026-8656 (Print) | 3026-8699 (Online)

journal homepage: https://ijdm.org.ng/index.php/Journals

Best Model Selection Method for the Best Latent Variables Determination When Solving Multicollinearity with Partial Least Squares

Olusegun O. Alabi[a], Gbemisola W. Ogunmefun[a], Toba T. Bamidele[a], Rasaki Y. Akinbo[b] and Olusesan T. Akintola[c]

[a]Department of Statistics, Federal University of Technology, Akure, Nigeria
[b]Department of Mathematics and Statistics, Federal Polytechnic, Ilaro, Nigeria
[c]Department of Mathematics and Statistics, Joseph Ayo Babalola University, Ikeji-Arakeji, Nigeria

ARTICLE INFO

Article history:
Received 12 September 2025
Received in revised form 20 November 2025
Accepted 30 November 2025

Keywords:
Multicollinearity, Partial Least Squares, Latent Variables, and Total Mean Squared Error

MSC 2020 Subject classification:
62J99

ABSTRACT

Violating the assumption of independence among explanatory variables in the linear regression model leads to multicollinearity. In the presence of multicollinearity, the Ordinary Least Squares (OLS) estimator yields inefficient parameter estimates, whereas Partial Least Squares (PLS) estimates are more robust. Moreover, in PLS, weights must be assigned to each explanatory variable before the latent variables are extracted. Two significant challenges associated with the PLS method are the choice of the weight scheme and the selection of latent variables (LVs) to obtain an efficient estimate of the model parameters. Two methods of weight allocation are considered in this study: equal weight allocation and the variance of the regressors, while the two commonly known methods of model selection are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). AIC and BIC were used to select the best model for determining the optimal latent variables. Consequently, the study compared the performance of PLS results when the two weight attachment schemes and the two commonly used methods of model selection were used to determine the best latent variables. Efficient validation of PLS was performed using Total Mean Squared Error (TMSE) results for all model parameters obtained by the PLS estimator across different scenarios: varying sample sizes, multicollinearity levels, variability values, weight assignments, and model selection methods. Hence, the study concluded that the BIC method of model selection is the best for determining the optimal latent variables to use when employing PLS methods of estimation to handle multicollinearity in a Linear Regression Model.

1. Introduction

Partial Least Squares (PLS) estimation methods have become a valuable statistical tool for handling complex and high-dimensional data, particularly in fields where multicollinearity and limited sample sizes can complicate traditional regression analyses. PLS methods offer a solution that balances predictive accuracy and interpretability, especially when the predictor variables are highly correlated or when data sets are "wide," meaning they have more predictors than observations (Wold *et al.*, 2016). The origins of PLS lie in econometrics, with subsequent developments enhancing its flexibility and applications across various fields. The PLS approach is designed to build predictive models by finding linear combinations of predictors (latent variables) that maximize the covariance with the response variable (Höskuldsson, 2015). This property makes it a preferred technique for situations where conventional least-squares methods may suffer from overfitting or instability due to multicollinearity.

PLS methods are robust to noisy or missing data. They are increasingly used in high-dimensional predictive modeling tasks in bioinformatics, where genetic and molecular data often exhibit extreme multicollinearity and sparsity

* Corresponding author. Tel.: +2348035807226
E-mail address: alabioo@futa.edu.ng (Alabi O. Olusegun.)
https://doi.org/10.62054/ijdm/020417

(Westerhuis *et al.*, 2016). The robustness of PLS methods in handling such challenges makes them a suitable choice for predictive tasks in personalized medicine, where datasets are often complex and multi-layered (Boulesteix & Strimmer, 2015). In Partial Least Squares (PLS) estimation, weight selection is a critical component that directly impacts the quality, interpretability, and reliability of the resulting model. In PLS, weights are used to construct linear combinations of predictor variables, called latent variables, that maximize the covariance with the response variable. These latent variables are central to the PLS approach, as they serve as reduced-dimensional representations of the original predictors, capturing the most relevant information for predicting the outcome. The choice of weights, therefore, affects not only the relationships modeled but also the overall predictive accuracy of the PLS model. However, selecting appropriate weights is inherently challenging due to the complexity of PLS, which combines elements of principal component analysis and multiple regression to balance data dimensionality reduction and prediction. Consequently, weight selection is often susceptible to multicollinearity and the structure of high-dimensional datasets, making it difficult to determine optimal weights that provide a robust model across various sample distributions and applications.

In a related context, Bamidele *et al.* (2024) used a dimension-reduction procedure to address concerns about high collinearity among variables in simultaneous equation models. Their findings demonstrated that this procedure effectively addressed multicollinearity while preserving the information content of the original dataset, suggesting that similar strategies could enhance the stability and accuracy of PLS weight estimation in complex modeling environments. Given these challenges, finding robust, standardized approaches to weight selection in PLS estimation remains a crucial yet unresolved problem. Without reliable methods for determining optimal weights, PLS's effectiveness is compromised, particularly in applications requiring high predictive accuracy and generalizability. Addressing this problem is therefore essential for advancing the practical application of PLS, ultimately improving its utility in predictive modeling and multivariate data analysis across diverse research fields. This study aims to determine the optimal weight attachment in partial least squares estimation for a linear regression model with multicollinearity. This study proposes and develops new techniques for determining weights in partial least squares (PLS) models, investigates the impact of weight selection on the predictive accuracy and interpretability of PLS models, and determines the optimal weight or overall best weight attachment for addressing multicollinearity with the PLS method of estimation. The study contributes by addressing the practical challenge of selecting the optimal number of latent variables in PLS by demonstrating that the Information Criterion provides a consistent and effective model selection rule.

2. Methodology

2.1 Model Specification

Given: $$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_p X_p + u \quad (1)$$

Y_i is the endogenous variables, $i = 1, 2, 3, \ldots, n$,

X are the exogenous variables, with dimension n by p,

$\beta_0, \beta_1 \beta_2, \ldots, \beta_6$ are the parameters of the model; and

u is the error term that is normally distributed with mean 0 and variance σ^2

2.2 Simulation Procedure

Consider the regression model from equation 1, a Monte Carlo experiment is performed 1000 times for five sample sizes (n=20, 30, 50, 100, and 250) and five levels of multicollinearity ($\rho = 0.8, 0.9, 0.95, 0.99, 0.999$) that exist among the exogenous variables under three different variability(variance) values (v= 1, 25 and 100). The correlation levels ρ were selected to represent a spectrum of strong to near-perfect multicollinearity, ranging from conditions commonly observed in applied studies to extreme cases that severely undermine the stability of OLS estimation. These values allow for a systematic evaluation of estimator robustness as multicollinearity intensifies. Similarly, the variance levels v capture low, moderate, and high error variability, reflecting increasingly noisy data-generating processes encountered in practice. The joint consideration of these parameter settings enables a comprehensive stress test of estimator performance, providing insights into how multicollinearity and error variance interact to affect estimation accuracy and latent-variable selection in PLS regression, thereby enhancing the practical relevance of the study's findings.

The procedure for simulation is as follows;

i. **Simulation of Error Term u_i**

u_i ~ N(0,σ^2), Since the standardized normal Z_i can be expressed as

$$Z_i = \frac{X_i - u_i}{\sigma_i};\ Z_i \sim N(0, 1) \quad (2)$$

ii. **Generation of Independent Variables x_i**

Suppose that, $X_i \sim N(u_i, \sigma^2)$, if these variables are correlated, then X_1, X_2, X_3, X_4, X_5 *and* X_6 can be generated with the equation

$$X_i = ((1-\rho^2)^{0.5}) * u_i + \rho * u_0 \quad (3)$$

where,

ρ is the correlation (multicollinearity) level that exists between the exogenous variables $-1 \leq \rho \leq 1$, $i = 1, 2, 3, \ldots, n$ (Ayinde K, 2007).

u_i and u_0 are independent standard normal random variables,

u_0 is a common latent factor shared by all regressors.

iii. **Generation of Dependent Variable Y_i**

Define Y as a linear combination of the parameters with predictors plus the random noise.

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_6 X_6 + u_i$$

where,

$\beta_0 = 0.5;\ \beta_1 = 1.2;\ \beta_2 = 0.9;\ \beta_3 = 0.8;\ \beta_4 = 2.1;\ \beta_5 = 1.8;\ \beta_6 = 3.2;$

$i = 1, 2, 3, \ldots, n;$

$j = 1, 2, 3, \ldots, 6$

2.3 Partial Least Squares (PLS) Method of Estimation

The PLS Method of estimation can be done as follows:

Obtain the Latent Variable Matrix:

$$T = XW \quad (4)$$

where,

T is an n by k matrix of latent variables,

W is a p by k matrix of weights.

Regress Y on Latent Variable:

$$Y = T\beta + \varepsilon \quad (5)$$

where,

β is the vector of regression coefficients for the latent variables T,

ε is the error term. $\varepsilon \sim N(0, \sigma^2 I)$

2.4 Existing Weight Attachment (W)

i. Equal Weight: The Weight for each predictor will be set to the same value. This implies that no predictor is prioritized over others at the outset.

$$w_i = 1 \quad (6)$$

ii. Variance of Regressors: Weight for each predictor $X_i$will be proportional to the variance of that predictor,

$$w_i = Var(x_i) \quad (7)$$

iii.

2.5 Extraction of Principal Components

Principal components were extracted with Non-Linear Iterated Partial Least Squares Algorithm (NIPALS), which was first proposed by Wold (1966) and further developed by Wold (1975). In the former, the NIPALS algorithm was used to extract principal components. Then, deflation techniques were used to remove the extracted component, and the algorithm continued extracting the second component and so on. The NIPALS algorithm, modified by Wold (1975) to account for responses, then yielded PLS regression orthogonal scores presented by Wold. The general NIPALS algorithm for PLS is given as follows:

The algorithm carries out the bilinear decomposition given by

$$Y = TQ + E \quad (8)$$

$$X = t_1 p_1^{'} + \ldots + t_k p_k^{'} + F \quad (9)$$

$$Y = u_1 q_1^{'} + \ldots + u_k q_k^{'} + E \quad (10)$$

With E and F corresponding to residual terms. The decomposition in (9) and (10) is extensively analyzed in Martens and Naes (1989). Indeed, the bilinear decomposition links the matrix Y to the matrix by using latent vectors $t_1, t_2, \ldots, t_k$. The expressions (9) and (10) are successfully used together with scores (t_k and u_k) and weight vectors (w_k and q_k) as defined in this algorithm:

Input:$(X_0 \leftarrow X; Y_0 \leftarrow Y)$

For k= 1, 2, …,p and $u_k$is a column of Y.

Continue with this until convergence is met, then do the following:

Step1. $w_k \alpha X_{k-1}^{'} u_k C(X)$

Step2. $t_k \alpha X_{k-1} w_k R(X)$

Step3. $q_k \alpha Y_{k-1}^{'} t_k C(Y)$

Step4. $u_k \alpha Y_{k-1} q_k R(X)$

If convergence is met, compute the loading vector:

$$p_k = \frac{X_{k-1}^{'} t_k}{t_k^{'} t_k} \quad (11)$$

Store: $T[k] \leftarrow t_k; U[k] \leftarrow u_k; P[k] \leftarrow p_k; Q[k] \leftarrow q_k$

Take as new data

$X_k = X_{k-1} - t_k p_k'$ and $Y_k = Y_{k-1} - t_k q_k^{'}$

And go to step1, with $X = X_k$ and $Y = Y_k$.

αemphasizes the fact that different normalizations can be used. For example, normalizing the extracted vectors by using the normalization constants $w_k' w_k, t_k' t_k$ etc., allows the whole procedure to be formulated as a sequence of simple regressions. The derived vectors then correspond to the slope in simple univariate regressions (Bastien et. al., 2005). This approach permits the straightforward, direct handling of missing data (Tenenhaus, 1988). The notation C and R indicate the column and row space of matrices X and Y, respectively. This is also true for all $2 \leq k \leq p$ with data being deflated according to the above expression. The deflated data can be expressed in terms of the original data. X_0 and Y_0 at any k, according to the equations:

$$X_0 = (1 - P_{Tk}) X_0 \quad (12)$$

$$Y_0 = (1 - P_{Tk}) Y_0 \quad (13)$$

where,

P_{Tk} denotes the projection or the hat score matrix $T_k (T_k^{'} T_k)^{-1} T_k^{'}$ (Kondylis, 2006).

2.6 Procedures to determine the Latent Variables (LVs):

Considering univariate PLS, that is to say $i = 1$, so that Y is scalar. This is the typical multiple regression setting. An attempt was made to find the models of the form.

$$\hat{Y} = \hat{\beta}_1 T_1 + \ldots + \hat{\beta}_p T_p \quad (14)$$

where T_p is the linear combination of the $X's$

Assuming that all variables have been centered to have a mean of 0. This means that our intercept terms will always be zero. Here is the algorithm for determining the $T_i's$.

i. Regress Y on each X_i in turn to get b_{1i}.

ii. Form $T_i = \sum_{i=1}^{m} W_{1i} b_{1i} X_{1i}$ where the weights w_{1i} sum to one.

iii. Regress Y on $\mathbf{T}_1$ and each x_i on $\mathbf{T}_1$.The residuals from these regressions have the effect of T_i removed. Replace Y and each X_i by the residuals of each corresponding regression.

iv. Go back to step one, updating the index.

v. Continue in this manner till all the components or latent variables are derived.

2.7 Determination of the best model that contained the best latent variables (LVs)

Latent Variables are the hidden or unobserved variables that cannot be directly measured, but their presence can be inferred from other variables. In this work, variables will be divided into several components called latent variables, Since all the variable components or the latent variables cannot be used to have a good result in PLS method, we need to select the best latent variables, selection of these good latent variables is one of the problems in partial least square, to avoid these we adopt the usage of the commonly used methods of model selection to select the best model

that contains the best latent variables, selections of the best latent variables to be used with the weight attachment was done by the two standard model selection methods are: Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

i. Akaike Information Criterion (AIC): Adjusting for the number of parameters, this criterion assesses the quality of fit between different models. In other words, the model with the lowest AIC score is chosen. AIC is defined as:

$$\begin{aligned} \text{AIC} &= e^{\frac{2k}{n}} \sum_1^n \frac{u_i^2}{n} \\ &= e^{\frac{2k}{n}} \frac{RSS}{n} \end{aligned} \tag{15}$$

where k is the number of regressors (including the intercept) and n is the number of observations

$$\ln(\text{AIC}) = \frac{2k}{n} + ln(\frac{RSS}{n}) \tag{16}$$

where ln(AIC)= natural log of AIC and $\frac{2k}{n}$is the penalty factor.

ii. Bayesian Information Criterion (BIC): This criterion is quite similar to AIC, except it penalises the usage of more parameters more severely. It is recommended to choose the model with the smallest BIC value.

$$\begin{aligned} \text{BIC} &= n^{\frac{k}{n}} \sum_1^n \frac{u_i^2}{n} \\ &= n^{\frac{k}{n}} (\frac{RSS}{n}) \end{aligned} \tag{17}$$

In log-form

$$\ln(\text{BIC}) = \frac{k}{n} ln(n) + ln(\frac{RSS}{n}) \tag{18}$$

where, $\left[\frac{k}{n} ln(n)\right]$is the penalty factor.

The model with the lowest AIC and BIC values is often chosen as the best. The model selected as best by AIC or BIC, with the best latent variables, is to be used. The variable component in the best-selected model is the set of latent variables selected for PLS estimation. The model with the lowest AIC or BIC is best and requires the fewest LVs. LVs can be selected with either AIC or BIC. The next step is to determine the best LV from the two selected by AIC or BIC.

2.8 Criterion for Determination of the Best Latent Variables

The criterion used for the determination of the best latent variables is Total Mean Squared Error (TMSE), which was obtained from Mean Squared Error of each of the parameters in the model by adding the MSE value of every parameter in the model; the formulae for MSE and TMSE are as follows.

$$MSE(\hat{\beta}_{ij}) = \frac{1}{R} \sum_i \sum_j (\hat{\beta}_{ij} - \beta_{ij})^2 \tag{19}$$

where $i = 1, 2, 3, \dots, n$; $j = 0, 1, 2, 3, \dots, 6$

$$TMSE(\hat{\beta}_{ij}) = \sum_1^n \sum_{j=0}^6 MSE(\hat{\beta}_{ij}) \tag{20}$$

The method of model selection (AIC or BIC methods) was used to select the LVs; the selected LVs was used with weight attachment method(s) under the study for PLS estimation to obtained MSE for each of the model parameters, performance of the PLS estimator was determined by Compared its TMSE value with that of TMSE of OLS, these comparison was done in a particular sample size with a particular level of multicollinearity, a particular variability value with different weight attachment and different methods of model selection(AIC and BIC). Then, the model selection method (AIC or BIC) that yielded the minimum TMSE for the parameters was determined to be the best for selecting the latent variables in that scenario.

3. Results and Findings

Determination of the best-performed PLS Estimators after weight was attached, and AIC and BIC were computed to determine the best latent variables, with the procedure of selecting the model that contained the minimum AIC or BIC value first as the model with the best LVs. Then, the selected best LVs with the lowest AIC were used with weight one (w1) to obtain estimators for each parameter in the model using the PLS method of estimation. The MSE was calculated for each estimated parameter. After all the above procedures, the MSE values for each parameter were summed to obtain the TMSE for a given scenario. As the procedures was follow under weight one (w1) with AIC

as the method of method of determining the best LVs, it was done under Weight two(w2) for AIC, the same procedure was also followed under weight one (w1) with BIC as the method of method of determining the best LVs, likewise the procedure was followed under Weight two(w2) with BIC as the method of method of determining the best LVs, then, the TMSE value was obtained when OLS method of estimation was also used for each of the scenario. That is, when using weight one(w1) or weight two (w2) as the weight attachment, with the best latent variables selection (LVs)that was determined by either AIC or BIC, at a given sample size, a level of multicollinearity and a particular variance value, the Total Mean Squared Error (TMSE) value obtained for each Scenario is given in Table 1.

Table 1: Total Mean Squared Error values for OLS and PLS estimators under different weight schemes and model selection criteria.

Sample size	Multicollinearity levels	Variability values(variance)	TMSE values under two different weight attachments when AIC and BIC was used to select the best model that contained the best latent variables				OLS Total Mean Squared Error (TMSE)
N	ρ	V/Weights	Equal weight (W1)		Variance as weight(W2)		
		V	AIC	BIC	AIC	BIC	
20	0.8	1	1.261	1.263	1.270	1.263	1.750
		25	28.367	25.477	29.960	25.477	43.400
		100	113.288	97.454	117.506	97.454	174.800
	0.9	1	2.811	2.285	2.825	2.285	3.540
		25	64.789	51.705	59.652	51.705	88.620
		100	227.760	202.480	238.010	238.010	354.480
	0.95	1	4.535	4.464	4.858	4.464	7.340
		25	117.392	107.636	122.479	107.636	181.590
		100	468.872	424.215	488.301	424.215	734.370
	0.99	1	24.786	23.608	26.166	23.608	40.070
		25	634.789	588.876	661.630	588.876	1001.630
		100	2527.812	2359.731	2644.491	2359.731	4006.510
	0.999	1	272.464	240.530	278.687	240.530	430.830
		25	6748.463	6108.792	6905.683	6108.792	10070.650
		100	27007.437	24480.073	27639.680	24480.073	43082.610
30	0.8	1	1.041	1.196	1.070	1.196	1.060
		25	19.125	20.371	22.480	20.371	26.550
		100	71.533	72.499	85.357	72.499	106.180
	0.9	1	2.125	2.360	2.239	2.360	2.180
		25	36.072	39.016	44.103	39.016	54.562
		100	138.943	147.260	172.965	147.260	218.250
	0.95	1	4.059	4.150	4.239	4.150	4.580
		25	72.310	76.468	88.013	76.468	114.500
		100	281.765	299.540	352.130	299.540	457.990
	0.99	1	17.100	16.753	18.750	16.753	25.520
		25	391.436	373.907	445.058	373.907	638.080
		100	1546.870	1491.876	1780.689	1491.876	2552.330
	0.999	1	172.340	148.818	180.602	148.818	278.380
		25	4269.644	3658.225	4454.342	3658.225	6959.450
		100	17048.615	14638.851	17839.292	14638.851	27837.820
	0.8	1	0.451	0.587	0.469	0.587	0.450
		25	10.039	10.217	10.715	10.217	11.160
		100	38.886	37.396	44.069	37.396	44.660
	0.9	1	0.087	0.870	0.886	0.870	0.900
		25	19.344	20.241	22.355	20.241	22.570

50		100	76.014	76.915	91.794	76.915	90.280
	0.95	1	1.747	1.459	1.780	1.459	1.880
		25	38.873	40.710	46.579	40.710	46.910
		100	154.432	158.503	189.287	158.503	187.660
	0.99	1	8.028	8.362	8.524	8.362	10.360
		25	207.210	202.445	238.890	202.445	258.930
		100	829.121	813.295	959.091	813.295	1035.730
	0.999	1	88.209	80.243	93.029	80.243	112.540
		25	2217.912	1994.023	2349.310	1994.023	2813.510
		100	8877.912	7991.572	9415.276	7991.572	11254.050
100	0.8	1	0.205	0.246	0.210	0.246	0.212
		25	5.296	5.288	5.626	5.288	5.240
		100	18.831	15.781	20.541	15.781	20.170
	0.9	1	0.352	0.557	0.388	0.557	0.407
		25	9.637	9.448	10.856	9.448	10.060
		100	35.353	31.391	41.234	31.391	40.240
	0.95	1	1.016	1.111	0.964	1.111	0.830
		25	18.387	17.590	21.411	17.590	20.700
		100	70.349	63.783	83.245	63.783	82.790
	0.99	1	4.427	4.822	4.726	4.822	4.500
		25	92.985	81.821	104.995	81.821	112.560
		100	92.985	81.821	104.995	81.821	450.230
	0.999	1	40.286	33.803	42.171	33.803	48.480
		25	979.775	785.770	1029.234	785.770	1212.100
		100	3921.208	3124.614	4116.841	3124.614	4848.380
250	0.8	1	0.085	0.086	0.085	0.086	0.086
		25	2.166	2.487	2.368	2.487	2.030
		100	8.206	7.082	8.933	7.082	8.110
	0.9	1	0.208	0.274	0.165	0.274	0.166
		25	4.045	4.471	4.058	4.471	4.048
		100	30.048	13.113	17.654	13.113	16.160
	0.95	1	0.322	0.456	0.334	0.456	0.330
		25	7.866	7.897	9.008	7.897	8.300
		100	32.562	24.588	35.466	24.588	33.210
	0.99	1	1.709	1.932	1.808	1.932	1.800
		25	39.503	31.373	44.502	31.373	45.040
		100	156.383	114.637	176.902	114.637	180.160
	0.999	1	16.928	14.157	17.727	14.157	19.370
		25	413.672	281.094	434.692	281.094	484.130
		100	1650.918	1095.747	1735.652	1095.747	1936.510

Note. OLS = Ordinary Least Squares; PLS = Partial Least Squares; TMSE = Total Mean Squared Error; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; W1 = Equal weight; W2 = Variance weight; n = sample size; ρ = multicollinearity level; V = variability value.
Source: Author's computation from the simulated data.

4. Discussion

From the results presented in Table 1, several key patterns emerge regarding the performance of Partial Least Squares (PLS) estimation under different scenarios of sample size, multicollinearity, variability, weight attachment, and model selection criteria (AIC and BIC):

1. Impact of Multicollinearity: At a fixed sample size and variability, the Total Mean Squared Error (TMSE) increases as the level of multicollinearity (ρ) increases. For instance, at n=20 and V=1, TMSE rises from 1.261 (ρ=0.8) to 272.464 (ρ=0.999) under equal weight (W1) with AIC. This indicates that higher collinearity among predictors amplifies estimation error, reflecting the challenge multicollinearity poses for regression modeling.
2. Effect of Variability: At a fixed sample size and multicollinearity level, TMSE increases as the variability (V) of the errors rises. For example, at n=30 and ρ=0.9, TMSE for W1 under AIC increases from 2.125 (V=1) to 36.072 (V=25). This is consistent with theoretical expectations that higher error variance leads to less precise parameter estimates.
3. Sample Size Influence: Across all scenarios, TMSE decreases as the sample size increases, highlighting the stabilizing effect of larger datasets. For example, with ρ=0.95 and V=25, TMSE under W1 and AIC decreases from 117.392 (n=20) to 18.387 (n=100) and further to 7.866 (n=250). This aligns with classical regression theory: larger samples reduce estimation variability.
4. Comparison of Model Selection Criteria (AIC vs. BIC): Generally, PLS models selected via BIC yield lower TMSE than those selected via AIC, demonstrating that BIC tends to choose more parsimonious models that improve predictive accuracy. Exceptions occur when the variability is low (V=1), where AIC occasionally outperforms BIC (e.g., ρ=0.8, n=30, V=1). This suggests that in very low-noise scenarios, AIC may capture subtle predictive information that BIC misses.
5. Effect of Weight Attachment (W1 vs. W2): When BIC determines the best latent variables, TMSE values are identical regardless of whether equal weight (W1) or variance-based weight (W2) is used, indicating robustness of BIC-based selection to weighting schemes. In contrast, under AIC selection, TMSE values vary with the weight attachment: equal weight (W1) consistently yields lower TMSE than variance-based weight (W2), implying that W1 is generally preferable when using AIC.
6. Overall PLS vs. OLS Performance: Across all scenarios, PLS estimation substantially outperforms ordinary least squares (OLS) in the presence of multicollinearity. OLS TMSE is consistently higher, especially at high levels of multicollinearity and variability (e.g., 43082.610 at n=20, ρ=0.999, V=100 for OLS compared to 27639.680 for PLS with W2 and BIC). This reinforces the advantage of PLS for handling correlated predictors.

Collectively, the results suggest that, when handling multicollinearity in linear regression:

i. BIC is the more reliable criterion for selecting the best latent variables.
ii. Equal weighting (W1) is advantageous when using AIC, though BIC selection remains robust to weight choice.
iii. TMSE decreases with larger sample size and lower variability, and increases with higher multicollinearity, confirming theoretical expectations.

5. Conclusion

The study confirms that Partial Least Squares (PLS) estimation is superior to OLS when multicollinearity is present among explanatory variables. Key conclusions are: PLS consistently achieves lower TMSE than OLS, particularly under high multicollinearity and high error variance, demonstrating its effectiveness in addressing multicollinearity; the Bayesian Information Criterion (BIC) is the preferred method for selecting the optimal set of latent variables. BIC-based selection consistently leads to lower TMSE, regardless of the weight attachment, while AIC is sometimes suboptimal, especially with variance-based weighting (W2); Equal weights (W1) perform better than variance-based weights (W2) when AIC is used for latent variable selection, while the choice of weight has no impact on TMSE with BIC, indicating the robustness of BIC in guiding latent variable selection. In summary, the best practice for handling multicollinearity in linear regression with PLS is to select latent variables using BIC, ensuring robust and efficient estimation while minimizing predictive error.

REFERENCES

Ayinde, K. (2007b). Equations to generate normal variates with desired inter-correlation matrix. *International Journal of Statistics and System*, 2(2), 99–111.

Bamidele, T. T., & Alabi, O. O. (2024). A robust estimator for causal inference: Integrating two-stage least squares with principal component. *International Journal of Recent Research in Mathematics, Computer Science, and Information Technology*, 11(1), 27–32. https://doi.org/10.5281/zenodo.12671069

Bastien, P., Vinzi, V. E., & Tenenhaus, M. (2005). PLS generalized linear regression. Computational Statistics & Data Analysis, 48(1), 17–46. https://doi.org/10.1016/j.csda.2004.02.005

Höskuldsson, A. (2015). PLS regression methods. *Journal of Chemometrics*, 29(10), 569-582.

Kondylis, A. (2006). PLS methods in regression model assessment and inference (Unpublished thesis). Université de Neuchâtel.

Naes, T., & Martens, H. (1989). *Multivariate calibration.* John Wiley & Sons.

Tenenhaus, M. (1998). La régression PLS: Théorie et pratique. Technip, Paris.

Westerhuis, J. A., van Velzen, E. J. J., Hoefsloot, H. C. J., & Smilde, A. K. (2016). Multivariate data analysis of complex datasets: Applications in metabolic fingerprinting. *Journal of Chemometrics*, 30(7), 421-430.

Wold, H. (1975). Soft modelling by latent variables: The nonlinear iterative partial least squares (NIPALS) approach. In J. Gani (Ed.), Perspectives in Probability and Statistics: Papers in Honour of M. S. Bartlett (pp. 520–540). Academic Press.

Wold, S., Ruhe, A., Wold, H., & Dunn, W. J. (2016). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *Journal of Econometrics*, 67(1), 121–139. https://doi.org/10.1137/0905052