Jurnal Pendidikan dan Pembelajaran Matematika Indonesia
Vol. 15 No. 1, Tahun 2026
e-ISSN : 2615-7454

PARTIAL LEAST SQUARES FOR STATISTICAL MODELING: CONCEPTS AND APPLICATIONS

N.A.B. Ibrahim

Department of Applied Statistics, Universiti Teknologi MARA, Shah, Alam, Selangor, Malaysia
e-mail: nuraifiahibrahim90@gmail.com

Abstrak

Penelitian ini bertujuan untuk meninjau dan menjelaskan peranan metode Partial Least Squares (PLS) dalam pemodelan statistik, khususnya dalam mengatasi masalah multikolinearitas, data berdimensi tinggi, dan analisis variabel laten. Metode penelitian yang digunakan adalah kajian konseptual dan tinjauan literatur mengenai PLS serta pendekatan terkait, termasuk reduksi dimensi, analisis kovarians, pemodelan jalur, dan estimasi variabel laten, yang didukung oleh pembahasan aplikasi pada bidang kesehatan, ekonomi, teknologi, dan sistem organisasi. Hasil penelitian menunjukkan bahwa PLS menyediakan kerangka analisis yang fleksibel dan robust untuk meningkatkan akurasi prediksi, mengurangi kompleksitas parameter, serta meningkatkan interpretabilitas model melalui estimasi iteratif terhadap konstruk laten dan loading variabel dengan tetap menjaga stabilitas statistik pada data yang kompleks. Kontribusi penelitian ini adalah memberikan sintesis teoritis dan praktis mengenai metode PLS sebagai alternatif terhadap teknik regresi konvensional, sekaligus menegaskan relevansinya dalam penelitian modern yang melibatkan data berukuran besar dan berdimensi tinggi.

Kata Kunci: Analisis Variabel Laten; Data Berdimensi Tinggi; Multikolinearitas; Pemodelan Jalur

Abstract

This study aims to review and clarify the role of Partial Least Squares (PLS) methods in statistical modeling, particularly in addressing multicollinearity, high-dimensional data, and latent variable analysis. The methodology is based on a conceptual and literature-based review of PLS and related approaches, including dimensionality reduction, covariance analysis, path modeling, and latent variable estimation, supported by discussions of practical applications across health, economics, technology, and organizational systems. The findings indicate that PLS provides a flexible and robust framework for improving predictive accuracy, reducing parameter complexity, and enhancing interpretability through iterative estimation of latent constructs and loadings while maintaining statistical stability in complex datasets. The study contributes to the existing literature by synthesizing theoretical and practical perspectives on PLS modeling, emphasizing its usefulness as an alternative to conventional regression techniques and highlighting its applicability for modern data-intensive research environments.

Keywords: *Latent Variable Analysis; Dimensionality Reduction; Multicollinearity; Path Modeling*

1. Introduction

Multicollinearity between predictor variables is a common challenge in statistical modeling. When strong correlations exist, it becomes necessary to identify latent or underlying variables that account for shared variance. Standardizing equations and examining covariance structures do not always adequately resolve collinearity, which often requires reformulating the model to ensure stability and interpretability over time. In such cases, the constant term and parameter weights must be carefully estimated to avoid distorted or unstable solutions.

Modeling equations involve combining observed values and stochastic components to explain relationships among variables. When there is an imbalance between the number of predictor variables and the sample size, the estimated weights may not fully represent the underlying structure of the data. This limitation motivates the use of dimensionality-reduction techniques, such as principal component analysis (PCA), which aim to achieve proportional representation by measuring distances that depend on the magnitude of variation. Model development is therefore an iterative learning process that seeks equilibrium between explanatory power and statistical reliability.

Although principal component analysis (PCA) is widely used as a dimensionality-reduction technique to address multicollinearity, it is often less effective than partial least squares (PLS) when the primary objective is prediction and interpretation of relationships between predictor and response variables. PCA focuses solely on maximizing the variance among predictor variables without considering the dependent variable, which may result in components that explain large amounts of variation but have limited predictive relevance. Consequently, important information associated with the response variable can be overlooked, particularly in situations involving small sample sizes, highly correlated predictors, or complex latent structures. In contrast, PLS simultaneously considers both predictor and response variables by extracting latent components that maximize covariance between them, thereby producing more stable and interpretable parameter estimates. This integrated approach allows PLS to better capture the underlying relationships within high-dimensional datasets while reducing estimation bias and improving predictive accuracy. Therefore, compared with PCA, PLS provides a more efficient framework for handling multicollinearity and constructing robust statistical models in applied research contexts.

Not all variables contribute meaningfully to a given model. Feature selection methods, such as filtering and wrapper approaches, are used to exclude irrelevant variables based on logical and statistical relationships. A representative model is typically evaluated using least squares error criteria, with the goal of minimizing overall error magnitude. As model complexity increases, traditional geometric interpretations may lose explanatory power, necessitating adjustments to prevent overfitting. Simplifying the model by reducing parameters and correcting inverted or unstable directional relationships improves generalizability (Cheng, 2024).

Parameter weights, including the intercept, are often selected in accordance with Occam's Razor, favoring simpler models that adequately explain the data (Cheng, 2024). Small changes in parameter estimates can have substantial effects on linear relationships among variables. During preprocessing, parameter reduction and lag adjustments may be applied to construct a more parsimonious model in advance. Model complexity is thus tailored to optimize productivity and interpretability, particularly in applied contexts. Relationships among variables are assumed to be proportional in magnitude and direction, with latent variables playing a central role in capturing synchronized cause-and-effect patterns (Hair et al., 2020). Variations in significance can indicate past issues with linearity or collinearity that warrant further refinement (Hair et al., 2020).

Reevaluation of key indicators is essential for identifying influential variables and integrating multiple sources of information within a unified modeling framework (Hair et al., 2020). Advanced techniques such as tetrad analysis further improve understanding of latent structures and parameter relationships, particularly in the presence of sampling and specification challenges (Chinnaraju, 2025). In this context, Partial Least Squares (PLS) provides a practical and robust approach for estimating coefficients and latent constructs while maintaining model interpretability and stability (Schuberth et al., 2023). Therefore, the aim of this study is to review and clarify the role of PLS methods in identifying statistically significant parameters, reducing model complexity, and improving predictive capability in high-dimensional and multicollinear data environments.

The underlying assumptions guiding the analysis emphasize objectivity in model specification and parameter estimation. Structural models and their associated estimation parameters are designed to ensure consistency between numerical data and algorithmic procedures (Schuberth et al., 2023). As a result, this methodological approach has gained increasing popularity across disciplines for its ability to stabilize data distributions and reconcile differences among analytical techniques (Sarstedt & Liu, 2024). The expansion of data repositories has further encouraged the development of complementary and more sophisticated models, enhancing analytical perspectives (Sarstedt & Liu, 2024). Sample size remains a critical determinant of analytical capacity, influencing statistical power and the reliability of results.

To ensure valid inference, selection procedures must remain randomized, and parameter weights should fall within plausible ranges, as estimation errors tend to increase in the presence of uncontrolled mediating variables. Least squares–based path estimation is commonly employed to minimize variance and prioritize influential relationships within the model. Data reduction techniques are particularly important when working with smaller samples or high-dimensional vectors, allowing the number of latent variables to be aligned with the available information without distorting data flow. Such reductions enable broader sample inclusion while preserving meaningful differences among observations. Partial least squares methods, in particular, aim to identify the strength and proportion of relationships among variables, thereby reducing estimation gaps and improving model stability (Kwong & Wong, 2015).

Path modeling combined with data visualization supports meaningful ranking and interval estimation across constructs. Data reduction also enhances predictive capability by facilitating classification and grouping based on predefined relational structures. Multivariate statistical approaches extend beyond single-variable analyses by enabling the simultaneous extraction and integration of complex information, which is increasingly relevant in data-intensive environments (Perdana et al., 2023). Advanced analytical techniques, including deep learning, require robust sampling strategies and systematic optimization to manage large parameter spaces and loading structures within linear modeling frameworks (Ravand & Baghaei, 2016). Importantly, partial least squares methods remain suitable as nonparametric techniques, particularly for analyzing smaller sample sizes and establishing baseline models (Ravand & Baghaei, 2016).

Finally, effective use of time and computational resources allows additional data to be collected and presented through visualization tools that enhance interpretability and learning. Historically, such approaches preceded comprehensive latent variable models and laid the groundwork for subsequent regression-based analyses. A key advantage of these models is their ability to constrain estimation space and directional magnitude, ensuring balanced parameter estimation across different equation systems. As data volumes continue to grow exponentially, least squares estimation in conjunction with principal component analysis has become increasingly important for managing complexity. Ongoing research emphasizes adaptability and practical implementation, particularly through software applications that support advanced modeling once conceptual understanding has been established (Nascimento & Macedo, 2025).

The effective use of computational resources and visualization tools enables larger datasets to be analyzed more efficiently while improving interpretability and analytical learning. Historically, these approaches established the foundation for latent variable and regression-based modeling techniques. One important advantage of such methods lies in their ability to constrain estimation space and stabilize parameter magnitudes across complex equation systems. As data dimensionality and volume continue to increase, least squares estimation combined with dimensionality-reduction techniques, such as principal component analysis and Partial Least Squares (PLS), has become increasingly important for managing analytical complexity. However, despite the growing application of PLS-based methods across various disciplines, existing studies remain fragmented in explaining how PLS simultaneously addresses multicollinearity, latent variable estimation, and predictive stability within high-dimensional data environments. In addition, limited review studies provide a comprehensive synthesis of the methodological advantages of PLS compared with conventional approaches. Therefore, a research gap exists in clarifying the conceptual integration, practical implementation, and statistical relevance of PLS modeling in modern data-intensive applications. To address this gap, the present study first reviews the theoretical foundations of multicollinearity and latent variable modeling, followed by a discussion of dimensionality-reduction approaches and PLS estimation procedures. The study then examines the roles of loadings, covariance structures, and residual estimation in improving predictive performance

and model interpretability, before concluding with practical implications and future directions for statistical modeling research.

3. Methodology

The methodology section of this study is based on a qualitative literature review approach focusing on the theoretical and methodological development of Partial Least Squares (PLS) modeling. Relevant academic sources, including journal articles, books, and methodological studies related to PLS, PLS-SEM, latent variable analysis, multicollinearity, and dimensionality reduction, were systematically reviewed and analyzed. The review process involved identifying key concepts, estimation procedures, covariance structures, loading interpretations, and predictive capabilities associated with PLS methods. Comparative evaluation was also conducted between PLS and conventional approaches such as principal component analysis (PCA) and ordinary least squares (OLS) to examine their respective strengths and limitations in handling high-dimensional and complex datasets. The synthesized findings were then organized to provide a comprehensive understanding of the applicability, robustness, and interpretability of PLS methods across different research contexts.

4. Result and Discussion

The initial assumption of no relationship among variables may be challenged when empirical evidence reveals statistically meaningful associations. Such findings encourage reexamination of the state space to better explain latent matrix transitions that evolve over time. Simplified modeling frameworks are often adopted to reduce complexity while maintaining explanatory power. Contemporary data collection and reduction strategies therefore focus not only on data volume but also on the efficiency with which predictive and sustainable models can be developed. As sufficient data become available, the number of latent variables within the state space determines the structure of vector representations. Variability among observed variables can often be explained using ordinary least squares (OLS), while partial least squares (PLS) offers a complementary data-reduction approach that represents multiple variables within a modular latent-variable framework.

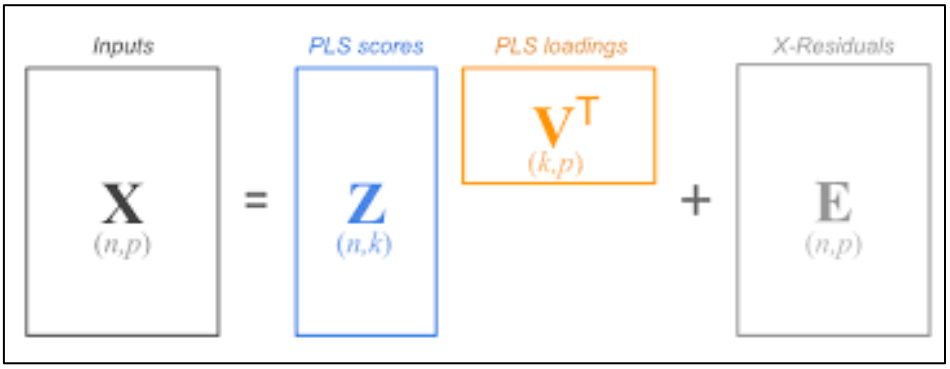

Figure 1: The Loadings and Error Vector in A Partial Least Squares Methods.

Figure 1 illustrates the relationship between loadings and error vectors in a PLS model. The outcome variable cannot be fully understood from individual predictors unless relationships are examined across varying levels of dispersion. Modeling therefore begins with a set of observed variables that are transformed into latent constructs, ensuring that relational magnitudes remain within a stable and interpretable range. From a methodological perspective, data are optimized through selective procedures that adjust numerical values according to the direction and strength of relationships among variables.

PLS modeling is inherently iterative, allowing predictor variables to be estimated under randomized selection processes that ensure fair representation. Covariance matrices are used

to estimate correlations among variables, supporting robust prediction while minimizing unnecessary interference. Component loadings are derived from latent constructs, and estimation errors are captured as residuals within the model. These residuals reflect unexplained variance rather than serving as determinants of model quality.

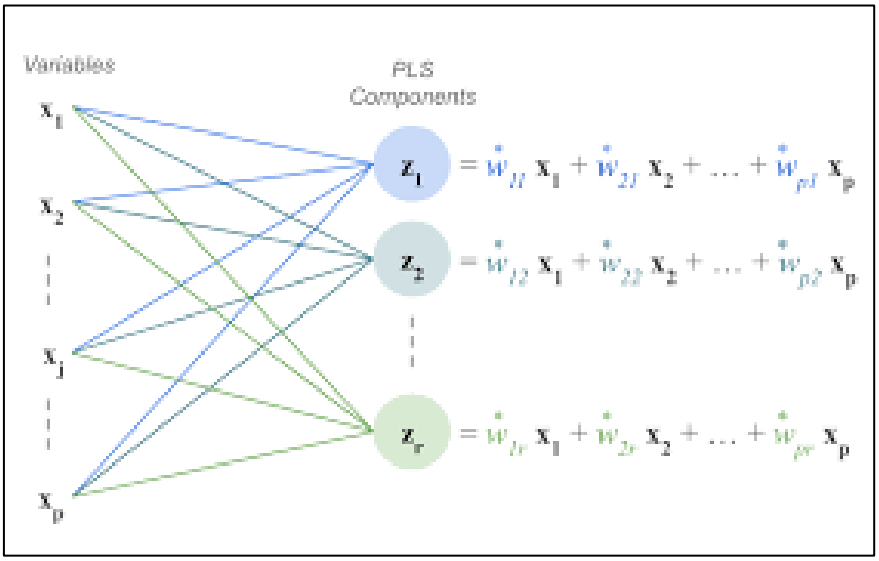

Figure 2: The Workable Equation for Each Lag of Variables.

Figure 2 presents the functional equations associated with different variable lags. Statistical power is essential for achieving precision and accuracy in applied problem-solving contexts. For example, health-sector risk models rely on factor loadings to classify biochemical indicators and other measurable attributes. In economic contexts, small variations in pricing or distribution can significantly influence consumer decision-making, such as limiting product substitution to guide purchasing behavior. Similarly, sustainable food practices—including the redistribution of staple foods and reduction of waste—can yield social and health benefits for communities.

Technological advancements in contactless payment systems during the COVID-19 pandemic further illustrate the need for robust statistical modeling approaches (Singh et al., 2020) . Data generated through barcode scanning, digital queuing systems, and secure access controls raise challenges related to privacy, redundancy, and data alignment (Singh et al., 2020). In some cases, transaction data may fail to accurately capture true ordering patterns or consumer classifications.

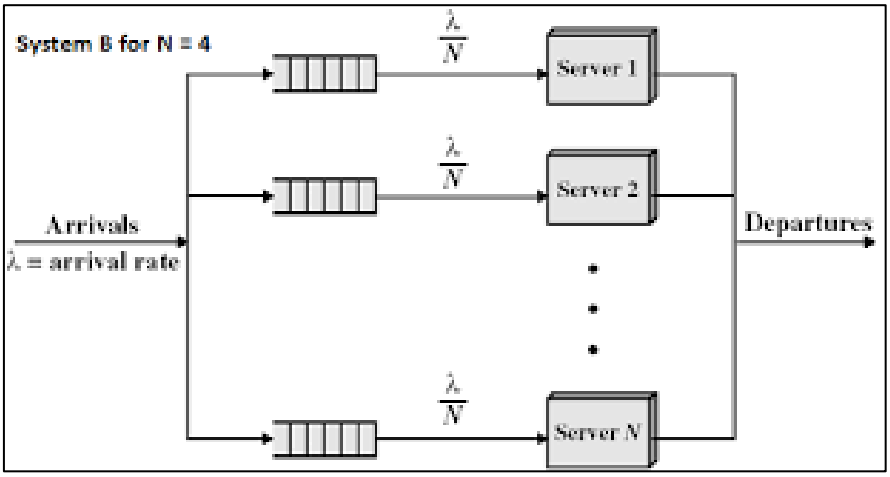

Figure 3: The Alignment Problem for the Assembly Line of Servers.

Figure 3 highlights alignment challenges within server-based systems. While multiple purchasing platforms coexist, queuing and sampling remain imperfect but necessary mechanisms for managing information flow. Profiling systems often categorize users into similar groups, enabling more efficient future interactions. High-demand service environments benefit from data-reduction strategies, such as PLS, which reduce parameter complexity in linear models while maintaining interpretability. Organizational perspectives vary widely across industries, reflecting differing philosophical and operational priorities.

Figure 4: The High Dimensionality in the Gap of A Philosophical Belief.

Figure 4 illustrates the challenge of high dimensionality in theoretical and conceptual modeling, emphasizing the importance of Partial Least Squares (PLS) methods in simplifying complex relationships among variables. In hypothesis-driven research, theories and axioms form the basis for understanding organizational and empirical phenomena; however, increasing data volume and multidimensional structures often complicate the interpretation and estimation of statistical relationships. Consistent with the objective of this study, dimensionality-reduction approaches such as PLS provide a practical framework for identifying latent constructs, reducing multicollinearity, and improving predictive accuracy while maintaining model interpretability. The ability of PLS to integrate observed and latent variables enables researchers to bridge gaps in empirical understanding and develop more stable analytical models for data-intensive environments. Economic applications further demonstrate the usefulness of these approaches in supporting efficient production and decision-making strategies across industries, including food and beverage manufacturing, where labor costs,

supply chain dynamics, and technological outsourcing significantly influence operational competitiveness and sustainability.

5. Conclusion

In conclusion, this review highlights the importance of partial least squares and related data-reduction techniques in addressing high-dimensional data, multicollinearity, and complex latent structures. By balancing model simplicity with explanatory power, PLS provides a flexible framework for analyzing relationships across diverse domains, from health and economics to technological systems. The iterative estimation of loadings, careful treatment of residuals, and emphasis on statistical power enhance model stability and interpretability. As data continue to grow in volume and complexity, the adoption of parsimonious, robust modeling approaches remains essential for generating reliable insights and supporting informed decision-making across applied research contexts.

References

Cheng, X. (2024). Signals and Systems Ins. In *Comput. Signal Syst,* 1 (1). https://soapubs.com/index.php/ICSS

Chinnaraju, A. (2025). Partial least squares structural equation modeling (PLS-SEM) in the AI Era: Innovative methodological guide and framework for business research. *Magna Scientia Advanced Research and Reviews*, *13*(2), 062–108. https://doi.org/10.30574/msarr.2025.13.2.0048

Nascimento, J. C. H. B. D., & Macedo, M. Á. D. S. (2025). The Potential of Partial Least Squares Structural Equation Modeling (PLS-SEM) with a Formative Approach in Accounting Research. *Revista de Educação e Pesquisa Em Contabilidade (REPeC)*, *19*. https://doi.org/10.17524/repec.v19.e3679

Hair, J. F., Howard, M. C., & Nitzl, C. (2020). Assessing measurement model quality in PLS-SEM using confirmatory composite analysis. *Journal of Business Research*, *109*, 101–110. https://doi.org/10.1016/j.jbusres.2019.11.069

Kwong, K., & Wong, K. (2015). *Partial least squares structural equation modeling (PLS-SEM) techniques using SmartPLS*. http://www.researchgate.net/publication/268449353

Perdana, P. N., Armeliza, D., Khairunnisa, H., & Nasution, H. (2023). Research Data Processing Through Structural Equation Model-Partial Least Square (SEM-PLS) Method. *Jurnal Pemberdayaan Masyarakat Madani (JPMM)*, *7*(1), 44–50. https://doi.org/10.21009/jpmm.007.1.05

Ravand, H., & Baghaei, P. (2016). Partial Least Squares Structural Equation Modeling with R. *Practical Assessment, Research & Evaluation,*. *21*(11). https://www.researchgate.net/publication/308169920_Partial_Least_Squares_Structural_Equation_Modeling_with_R

Sarstedt, M., & Liu, Y. (2024). Advanced marketing analytics using partial least squares structural equation modeling (PLS-SEM). *Journal of Marketing Analytics,* 12(1), 1–5. https://doi.org/10.1057/s41270-023-00279-7

Schuberth, F., Rosseel, Y., Rönkkö, M., Trinchera, L., Kline, R. B., & Henseler, J. (2023). Structural Parameters under Partial Least Squares and Covariance-Based Structural Equation Modeling: A Comment on Yuan and Deng (2021). *Structural Equation Modeling*, *30*(3), 339–345. https://doi.org/10.1080/10705511.2022.2134140

Singh, S., Srivastava, R. K., & Singh, A. (2020). Analysis of Queueing System and Impact of Digital Payments in Supermarket. *International Journal of Mathematics Trends and Technology*, 66(5), 106-116. https://doi.org/10.14445/22315373/IJMTT-V66I5P515