Review of: "On Optimal Linear Prediction"

Chris Rohlfs[1]

1 Electrical Engineering, Columbia University, United States

Potential competing interests: No potential competing interests to declare.

This paper explores the theory behind partial least squares (PLS), a dimensionality reduction approach with significant applications in chemistry and other fields involving high-dimensional data. A key distinction of PLS, compared to methods like principal component analysis (PCA), is its incorporation of outcome information to select predictors that maximize explanatory power in the context of a specific response variable. This focus on maximizing covariance between predictors and outcomes sets PLS apart in its ability to handle collinear and high-dimensional data.

The author highlights a gap in the existing theoretical literature on PLS, specifically regarding the conditions under which it is optimal or desirable compared to other methods. This paper makes an important contribution by seeking to provide a formal foundation for the use of PLS, clarifying the assumptions under which it can be considered an optimal method for linear prediction.

General Impressions:

While the paper contributes to the theoretical understanding of PLS, I feel that it could benefit from being more directly aligned with the practical concerns of researchers. As it stands, the paper seems to explore interesting theoretical connections across different perspectives on PLS, but it may not fully address key questions practitioners or applied researchers have about when and why PLS should be chosen over other methods. With some adjustments, particularly in terms of application and examples, this paper could become a valuable resource for both theoreticians and applied scientists.

Specific Points:

1. The Focus on PLS as 'Optimal': The paper's exploration of PLS as an optimal method is an interesting theoretical contribution. However, PLS is inherently the result of an optimization procedure (i.e., it maximizes covariance between X and Y), which raises the question of what specific broader cases are being considered when evaluating its optimality.

One of the key ways to assess the contribution of this paper is to examine whether it provides a useful framework for choosing PLS in a broader context. While the paper lays out conditions for PLS to be optimal, I have some concerns about the selection of models considered for comparison:

a) **Comparison to Other Methods**: PLS is only compared to estimators that use the same number mmm of latent variables. While comparisons to methods like ridge regression, lasso, elastic net, and PCA are valid, the reasons for choosing among these methods are often based on trade-offs such as interpretability, error handling, or overfitting, which could be discussed more clearly.

b) **Rationale for Dimensionality Reduction**: The assumed data generating process doesn't clearly justify why the dimensionality is reduced to m components. In practice, dimensionality is often reduced for reasons such as enhancing interpretability, accounting for measurement error, or mitigating redundancy in the predictors. It would be helpful for the paper and theoretical framework to capture these trade-offs more explicitly, as they are critical to real-world decision-making.

c) **Omission of Nonlinear Models**: The paper does not consider some of the most compelling modern alternatives, such as random forests, XGBoost, or support vector machines (SVM), which are nonlinear and increasingly used in practice. Although the paper's focus is on linear predictors, incorporating or at least discussing how PLS compares to popular nonlinear methods would broaden its relevance.

2. Accessibility and Visual Presentation: The paper is densely written, with heavy use of technical mathematics that spans multiple fields. This complexity is valuable for a theoretical audience, but it may hinder accessibility for a broader audience. Introducing **visual aids**—such as diagrams illustrating the optimization process or flowcharts comparing PLS to alternative methods—could greatly enhance clarity and engagement.

Additionally, while the analogy involving statisticians A and B is meant to provide intuition, it may not resonate with all readers. Simplifying or providing more practical examples could strengthen the paper's appeal to applied researchers.

3. Balancing Theory with Application: The paper is heavily theoretical, and while this is valuable, it might be more impactful if it demonstrated how PLS performs in **real-world applications**. If the goal is to identify the best method for predicting chemical relationships in high-dimensional settings, the paper would benefit from including examples that compare the empirical performance of PLS with other techniques. These examples could showcase how different methods trade off between **explainability**, **noise handling**, and **prediction accuracy** in practical scenarios.

Real-world applications would help readers assess whether the theoretical improvements discussed are substantial enough to influence model selection in practice. Even a single relevant application would strengthen the paper and provide concrete validation of the theoretical work.

4. Discussion of Machine Learning Alternatives: As computational power continues to grow, it is difficult to imagine that the research community will remain focused solely on linear methods like PLS. Machine learning methods, particularly neural networks, are gaining ground in high-dimensional prediction problems, including in chemistry. Including a discussion of these nonlinear alternatives, even if only in the context of future research, would make the paper more forward-looking and relevant to contemporary practices.

Conclusion:

This paper offers an important theoretical contribution to understanding the optimality of PLS under specific conditions.

However, its impact would be greatly enhanced by connecting the theory to practical applications, expanding the range of methods for comparison, and making the content more accessible. With these adjustments, the paper could serve as a valuable reference for both theoretical and applied researchers looking to understand the role of PLS in modern data analysis.