PROJECT REPORT

CHE208

MULTIVARIATE ANALYSIS OF MEMBRANE PROPERTIES: CHARACTERISATION OF INTERDEPENDENCIES

Team	Group 2
ABHINAV SINGH	230008002
GARVIT KUMBHAT	230008014
MOHAK DADHICH	230008023
VIGHNEHS MANDWARIA	230008039

Abstract

This research aimed to explore interdependencies between mechanical, electrical, and transport properties, etc. and their correlation to membrane quality. The data consisted of 120 samples, each described by eight predictor variables (X) and one quality response variable (y). Projection to Latent Structures (PLS) modelling was used to represent the connection between the X-space and quality outcome, reducing dimensionality and predictive modelling. A two-component PLS model was shown to provide an effective balance of complexity and performance, explaining approximately 48.4% of variance in X (R^2X(cum) = 0.484) and 45.5% of variance in y (R^2Y(cum) = 0.455). The model had poor predictive capacity with a cross-validated Q^2(cum) of 0.348. Even though the model explained a moderate level of the quality variance (R^2Y(cum) = 0.455), its poor cross-validated predictive capacity (Q^2(cum) = 0.348) indicates difficulties in accurately predicting membrane quality using the predictor variables.

Table of Contents

Introduction

What is PLS?

Projection to Latent Structures (PLS) is a statistical modelling method that is intended to explore complex relationships among groups of independent variables (predictors, X) and dependent variables (responses, Y), particularly when predictors are highly collinear or when there are more predictors than observations. PLS operates by extracting latent variables (factors) that maximise the covariance between X and Y, reducing dimensionality with a priority for prediction.

PLS finds application in almost all fields, including chemometrics, bioinformatics, finance, marketing, and social sciences, owing to its versatility and robustness in coping with high-dimensional, multicollinear, and noisy data.

How Does PLS Work?

PLS functions in three significant steps:

- Dimensionality Reduction: Retains the components (latent variables) from the predictors that have the most relevant information in predicting the responses.
- Regression: Regresses the responses on the latent variables.
- Cross-Validation: Helps find the optimal number of components to prevent overfitting and generalizability.

In contrast to Principal Component Analysis (PCA), which merely seeks to explain variance in X, PLS components are selected to maximise their predictive usefulness for Y.

- PLS at the same time fulfils three tasks:
    - The optimal explanation of the X-space.
    - The optimal explanation of the Y-space.
    - Maximises the connection between the X and Y spaces.
- One model is more effective:
    - Frequently takes fewer factors than Principal Component Regression (PCR).
    - Simpler to interpret.
- PLS accommodates multiple Y-variables.
- PLS presumes that there is an error in both X and Y.

Summary of the Dataset

There are 120 samples, with eight mechanical, electrical, and transport property measurements of a membrane (X1 through X8) and a single quality variable (Y). No missing values in any of the variables. All the variables are continuous and have been summarised using mean and standard deviation, implying the data is adequately prepared for multivariate analysis. The X variables exhibit moderate spread (the standard deviations range from approximately 0.27 to 0.52). In contrast, the quality variable Y exhibits a lower mean and less spread, indicating it is more densely clustered around its mean than are the X variables. This configuration is appropriate for investigating interdependencies between the X variables (e.g., by correlation or PCA) and for describing the relation between X and Y (e.g., by regression or PLS).

Methodology

Data preprocessing consists of cleaning, transforming, and structuring raw data to prepare it for analysis or machine learning models. It is crucial because raw data tends to be messy, inconsistent, or have errors; preprocessing enhances data quality, compatibility with algorithms, and more accurate, consistent results and strong model performance. This initial step avoids biased results and overall data analysis effectiveness. These are the steps involved in it:

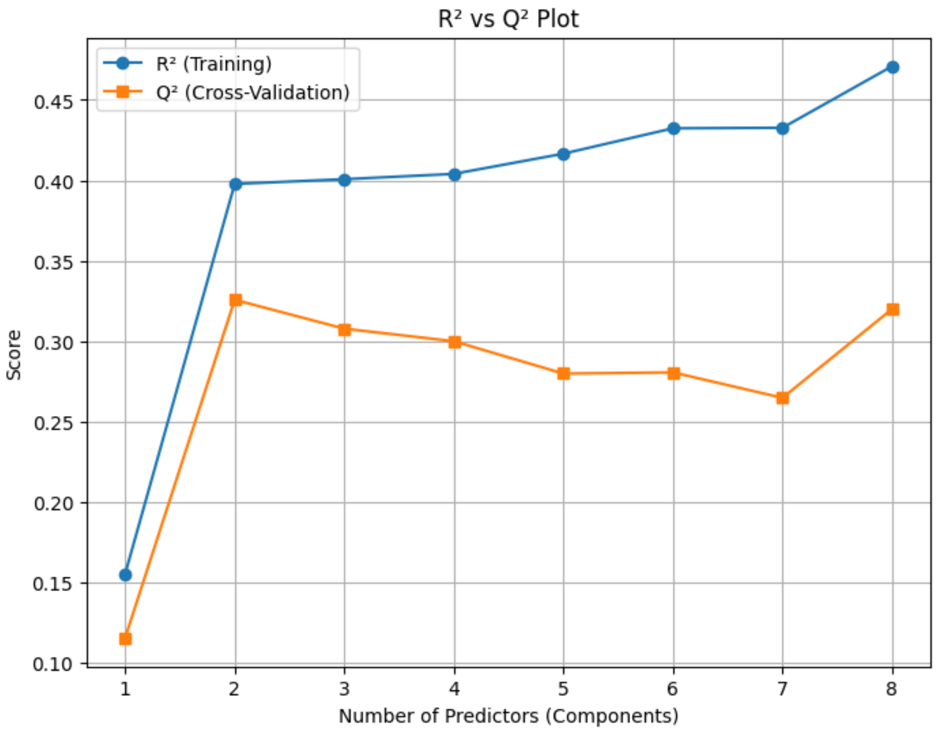
- Data Organisation and Formatting: Arrange the dataset properly, usually with the samples in rows and the variables in columns, and import it into the analysis software in an appropriate format.
- Missing Value Handling: Find and replace any missing data points in the dataset.
- Scaling and Centring

Multivariate data analysis, such as model development and validation, was carried out utilising SIMCA software. Cross-validation was utilized effectively to assist the selection of the best possible two-component structure with respect to cumulative Q^2 (Q^2(cum)) values, while additionally enabling an estimation of predictive capability and resilience of the ensuing PLS model.

Results and Discussion

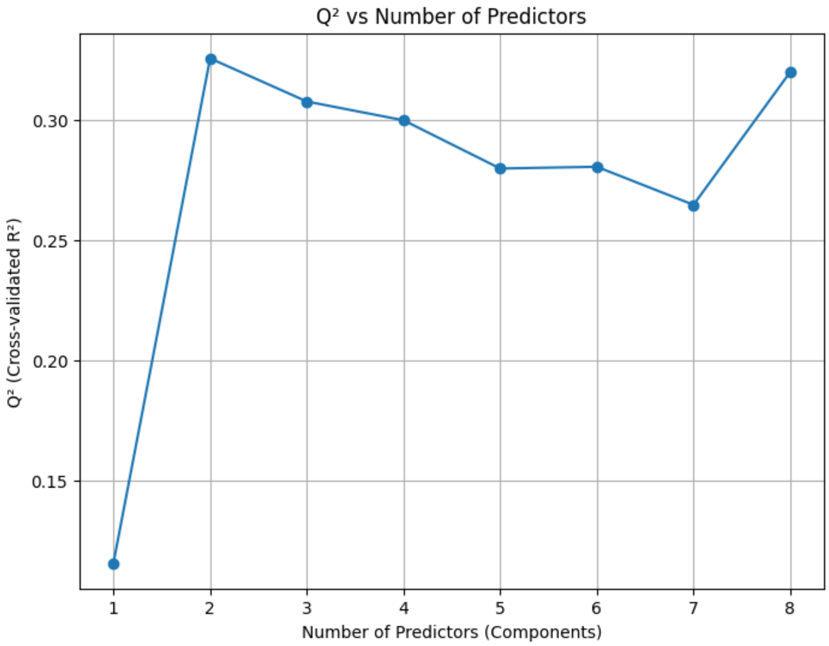
MLR

R^2 vs Q^2 Plot (Model Performance and Overfitting)



This plot compares the training R^2 (goodness of fit) and cross-validated Q^2 (predictive power) as the number of predictors increases. The widening gap between R^2 and Q^2 as more predictors are added suggests overfitting. The optimal model likely uses two predictors, balancing simplicity and predictive performance.

Q^2 vs Number of Predictors Plot

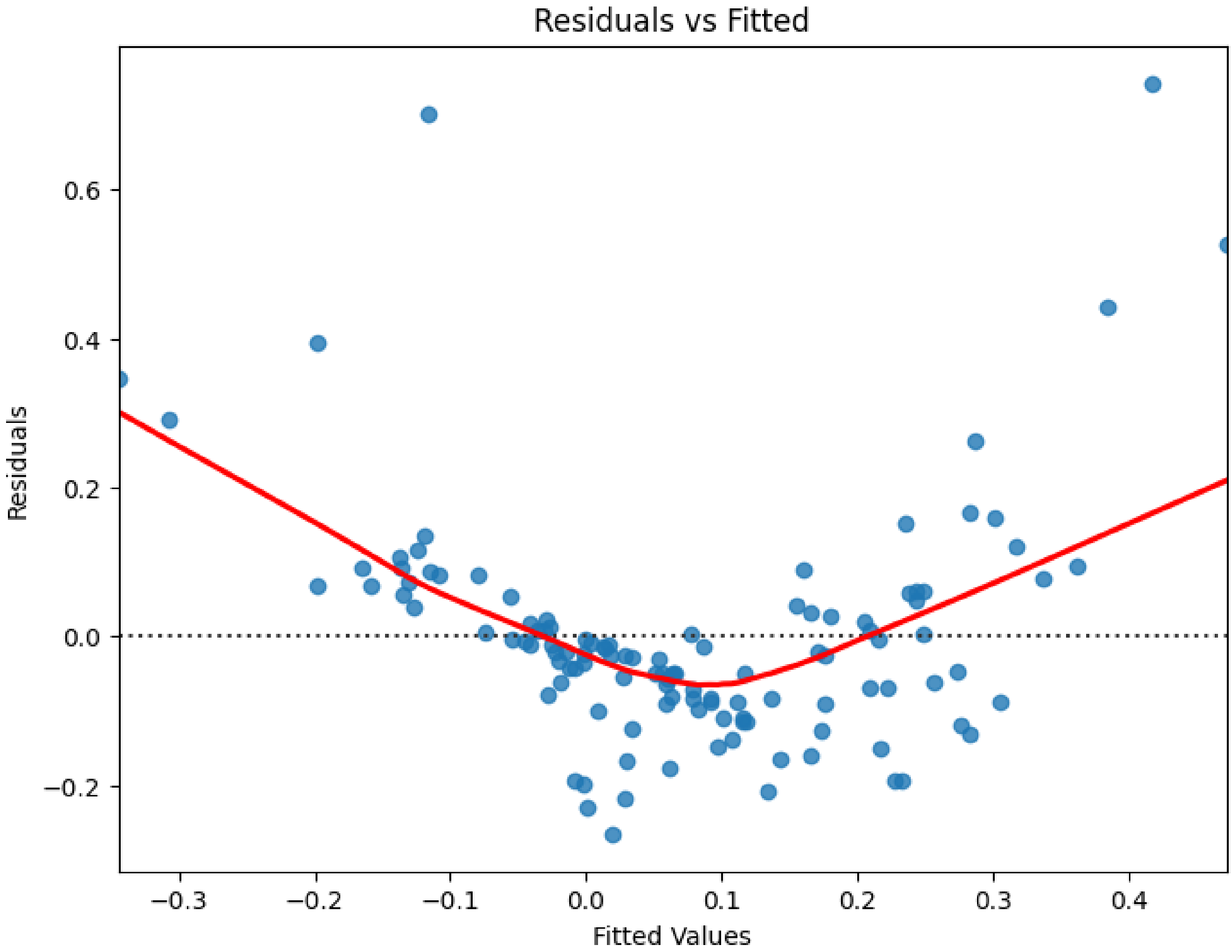


This plot shows the cross-validated Q^2 (predictive R^2) as a function of the number of predictors/components included in the model.

There is a sharp increase in Q^2 from 1 to 2 predictors, suggesting that the first two predictors capture most of the predictive information. After 2 predictors, Q^2 plateaus and fluctuates slightly, indicating that adding more predictors does not substantially improve predictive performance.

The best balance between model simplicity and predictive power is achieved with 2 predictors. Adding more may lead to overfitting or redundancy.

Residuals vs Fitted Values Plot

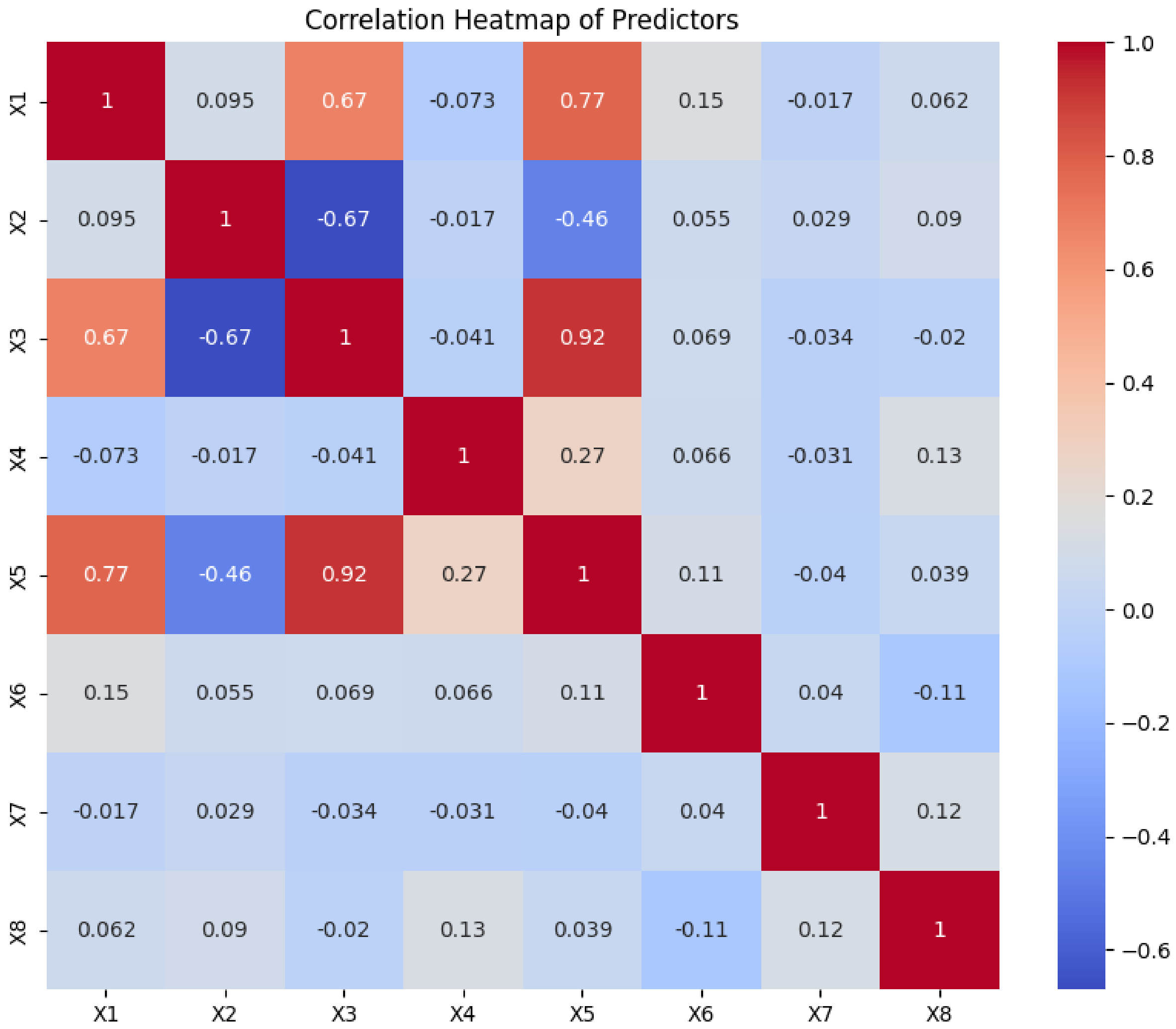


This plot shows the residuals (errors) versus the fitted values from the model, with a smoothing line (red).

The residuals display a clear curved (U-shaped) pattern, rather than being randomly scattered around zero. This suggests **non-linearity**: the linear model does not fully capture the relationship between predictors and the response.

Model improvement may be needed, such as adding polynomial terms, interaction terms, or using a non-linear model.

Correlation Heatmap of Predictors (Multicollinearity)

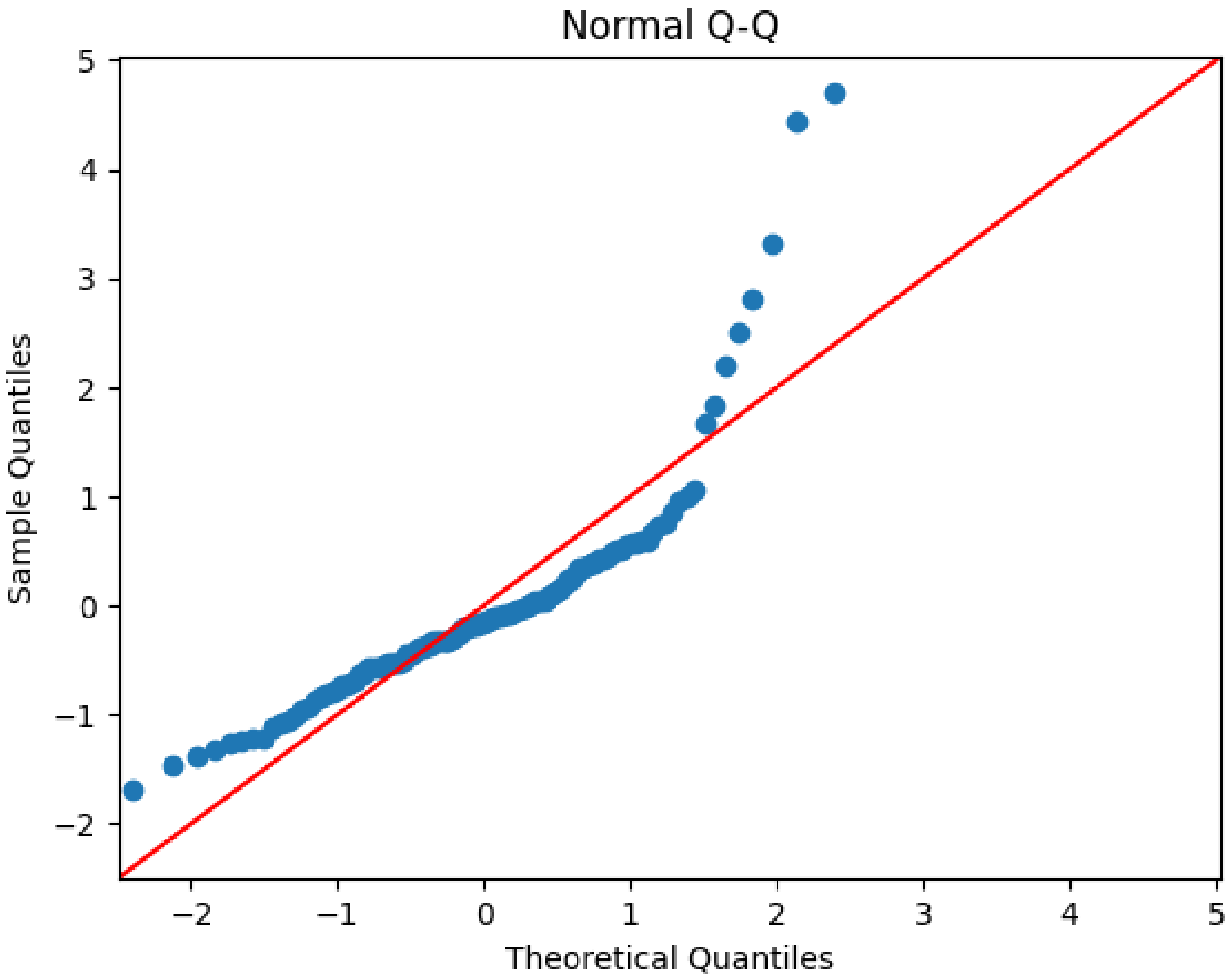


This heatmap shows pairwise correlations among the eight predictors.

Strong positive correlations are observed between X1, X3, and X5, and between X3 and X5 (e.g., X1–X5: 0.77, X3–X5: 0.92, X1–X3: 0.67). A strong negative correlation is seen between X2 and X3 (−0.67). Most other variable pairs have low correlations, indicating some predictors are independent.

The presence of multicollinearity (high correlation between some predictors) can affect the stability and interpretability of regression coefficients. Dimensionality reduction (e.g., PCA, PLS) or variable selection may help.

Normal Q-Q Plot (Normality of Residuals)



This plot assesses whether residuals are normally distributed.

The points deviate from the straight line, especially in the upper tail (right side), indicating the presence of **outliers** and some **right-skewness**. This violation of the normality assumption can affect the reliability of statistical inference (e.g., p-values, confidence intervals).

Key Problems with MLR on This Membrane Dataset

Multicollinearity Between Predictors

- Evidence: Extremely high pairwise correlations (e.g. X_1–X_5: r = 0.77; X_3–X_5: r = 0.92). Variance Inflation Factors are greater than 5 for these variables.
- Consequence: Standard errors are inflated—coefficient estimates are unstable: small changes in data will reverse signs or magnitudes (e.g. X_1 versus X_5). This defeats any hope of assigning effects to individual predictors.

Non-Linearity

- Evidence: Residuals vs. fitted-values plot has a strong U-shape instead of random scatter

- Consequence: The linear model does not capture curvature in the true response surface. Estimates of coefficients are biased, and predictive performance deteriorates until quadratic or interaction terms are added.

Non-Normal Residuals

- Evidence: Q–Q plot shows right-skewed residuals that are not on the 45° reference line.
- Consequence: Breaks the assumption of normally distributed errors, which makes p-values, confidence intervals, and other inferential statistics calculated from the MLR framework invalid.

Overfitting

- Evidence: R^2 for Training = 0.47, but cross-validation Q^2 = 0.33. Adding more predictors increases R^2 without enhancing Q^2.
- Consequence: The model is picking up noise instead of the signal. Its generalisation capability to new membranes samples is impaired, resulting in overly optimistic in-sample statistics and low out-of-sample performance.

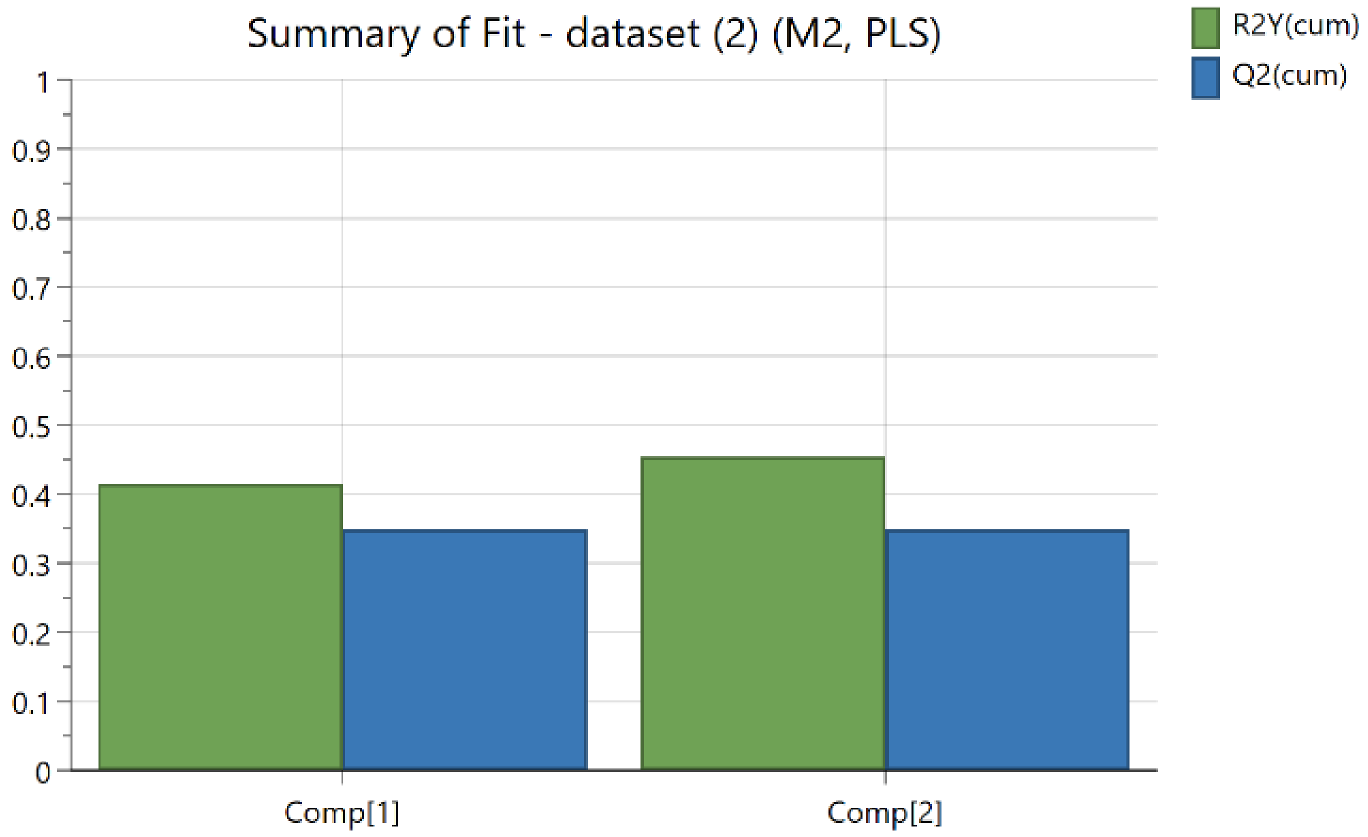
Dominant Variables Masking Subtle Effects

- Evidence: Large values of coefficients for X_0, X_1, and X_2 ($|\beta| \approx 20$), but β for X_3–X_7 close to zero.
- Consequence: Potentially significant contributions from X_3–X_7 are masked. Without handling collinearity and overfitting, we risk eliminating variables affecting membrane quality.

PLS

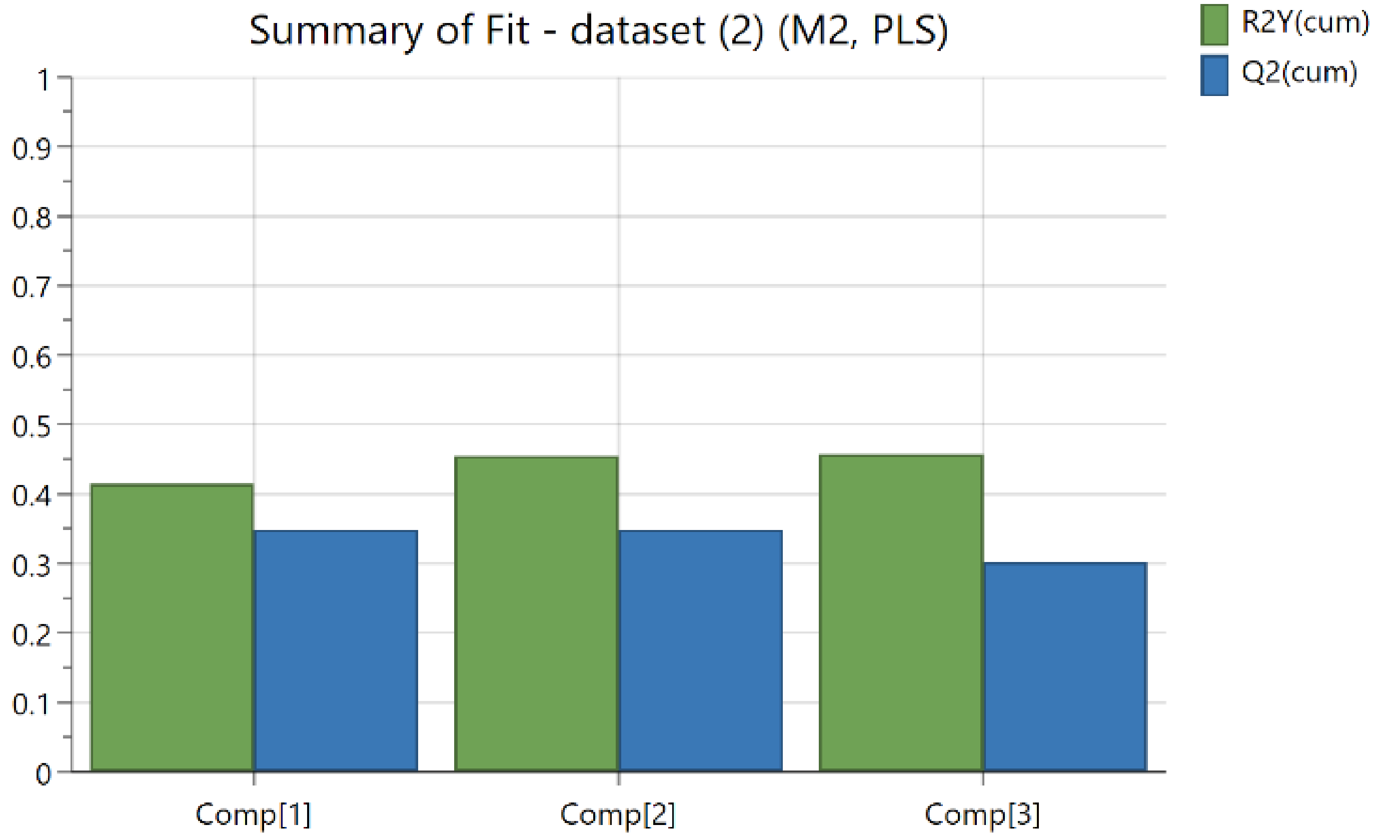
Summary of Fit

Following the given values and screenshot, below is the explanation of the performance metrics for the PLS model and how the two components were chosen:



The performance of the PLS model was measured in terms of cumulative explained response variable variance (R^2Y(cum)) and cumulative cross-validated predicted variance (Q^2(cum)). The R^2Y(cum) indicates the model fit, which reflects the extent to which it explains the response (Y) variation by the predictor (X) data within the training set. The Q^2(cum) from cross-validation estimates the predictive or goodness of prediction performance of the model for new data.

For the selected two-component model, R^2Y(cum) = 0.455, meaning that the model explains 45.5% of the variance in the response variable(s). The corresponding Q^2(cum) = 0.348, meaning that the model can predict approximately 34.8% of the response variance, suggesting positive predictive relevance ($Q^2 > 0$). While the model explains a moderate level of variance ($R^2Y = 0.455$), its predictive capability ($Q^2 = 0.348$) is comparatively limited but still exhibits some predictive capability.



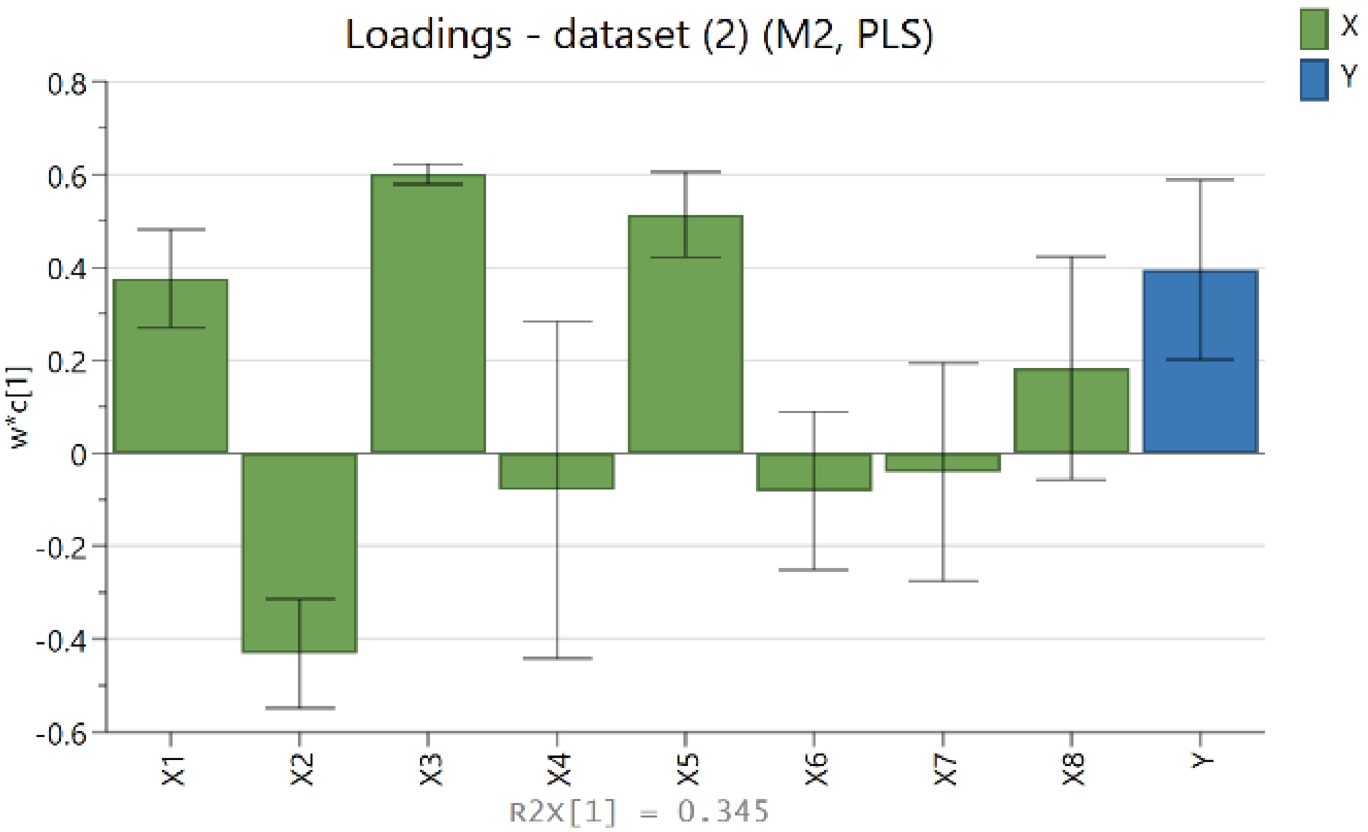
The decision to retain two components arose from reflecting on the development of these measures over the components added. The first component accounted for $R^2Y = 0.414$ and $Q^2 = 0.3481$. Including the second component increased variance explained a bit to R^2Y(cum)=0.455, while predictive power was not significantly changed at Q^2(cum)=0.3481. Adding the third component had little increase in variance explained (R^2Y(cum) increased only to 0.457) but at the expense of predictability (Q^2(cum) fell to 0.302). Since predictability is often the first consideration, and Q^2 fell when the third component was added, the two-component model was the best balance between variance explanation and predictability stability.

Comp No.	R^2Y(cum)	Q^2(cum)
Comp 1	0.414231	0.348194
Comp 2	0.455198	0.348486
Comp 3	0.457052	0.302089

Loadings Plot Interpretation

The given loadings plot for the first component (Comp1) of the PLS model:

This component explains the main covariance structure between the predictor variables (X) and the response variable (Y).



The loadings (w*c[1]) show the weight and direction of influence each variable has on this component.

- Response Variable (Y): The Y variable has a positive loading (approximately +0.4). It shows that this first component represents a pattern positively associated with the response variable; high scores for this component have higher values for Y.

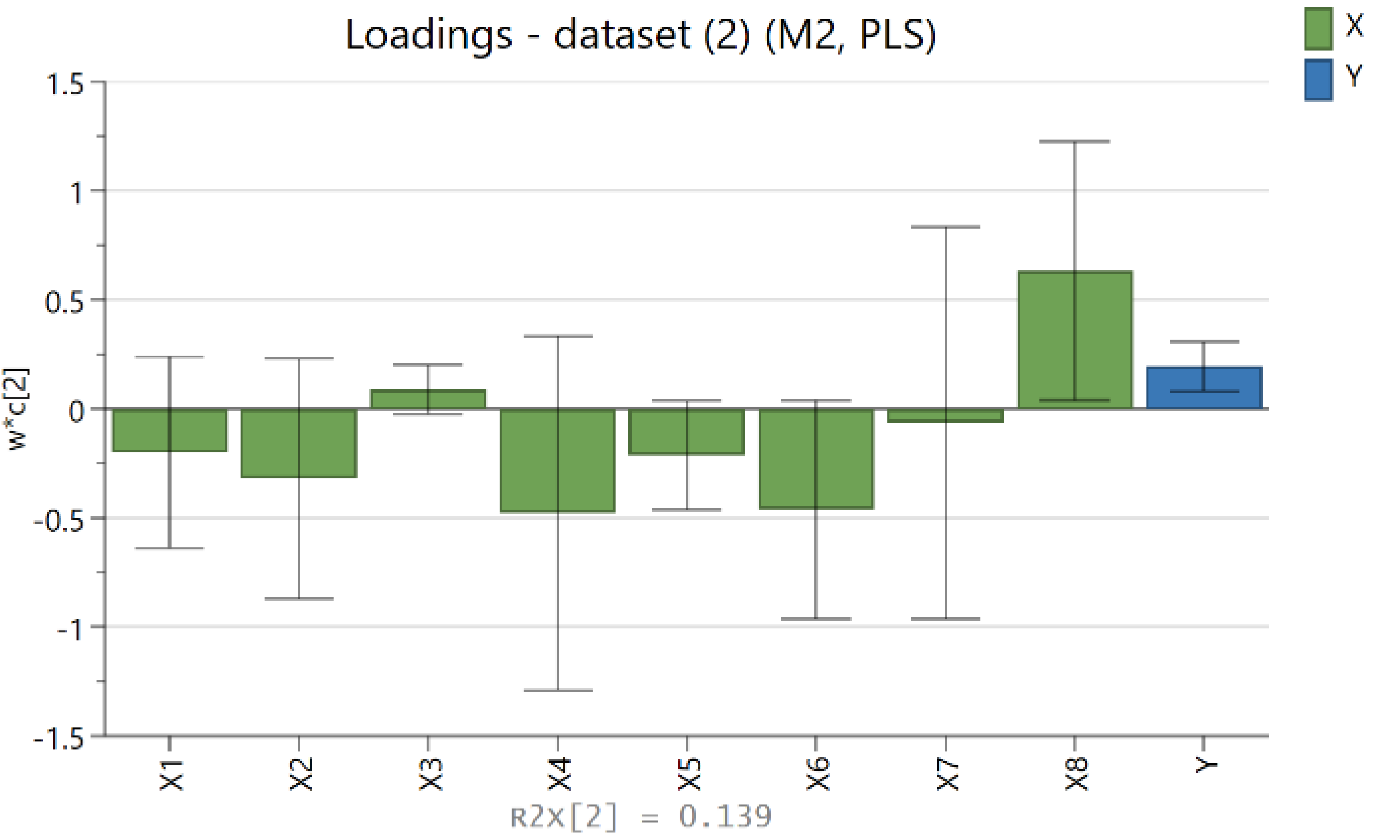
- Predictor Variables (X): The most significant positive loadings are for variables X3 (loading ≈ +0.6) and X5 (loading ≈ +0.5). This indicates that increases in these predictor variables are strongly associated with increases in response Y, as accounted for by this component.

Variable X1 (loading ≈ +0.38) and X8 (loading ≈ +0.2) are positively loaded, showing a positive association with Y, though less so than X3 and X5.

Variable X2 has a significant negative loading (-approximately -0.4) and indicates that high levels of X2 have a corresponding negative relationship with Y in the described relationship by this component.

X4, X6, and X7 are variables with very near-zero loadings, demonstrating that they tend to have minimum influence on the primary X-Y relationship captured in this first component.

In summary, the first factor predominantly observes a relationship where greater values of the response (Y) have bigger values of X1, X3, X5, and X8, but smaller values of X2. The X4, X6, and X7 variables contribute hardly any to this specific direction of variability2. This one factor explains approximately 34.5% of the variability in the X-variables (R2X1 = 0.345).



This second factor picks up on another covariance relationship between the predictors (X) and response (Y) perpendicular to the first factor. Loadings (w*c) indicate how each variable affects this specific pattern.

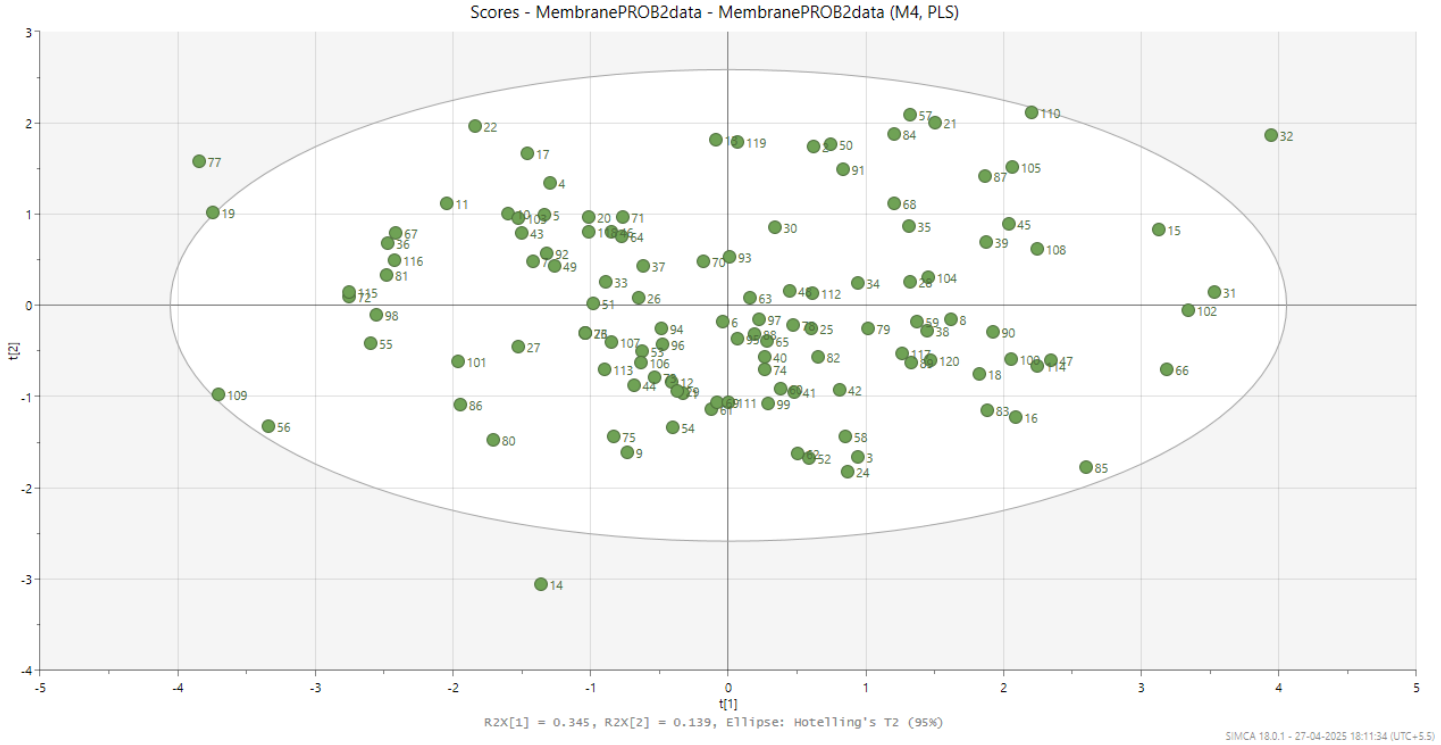
- Response Variable (Y): The Y variable carries a positive loading (about +0.2), so this component also describes a positive relationship with the response, albeit less powerfully than the first component.
- Predictor Variables (X): Variable X8 has the most significant positive loading (approximately +0.6), which means increases in X8 are strongly associated with Y in the relationship established by this second factor.

Variables X3 and X7 have very small positive and negative loadings, respectively, which means an almost zero contribution to this factor.

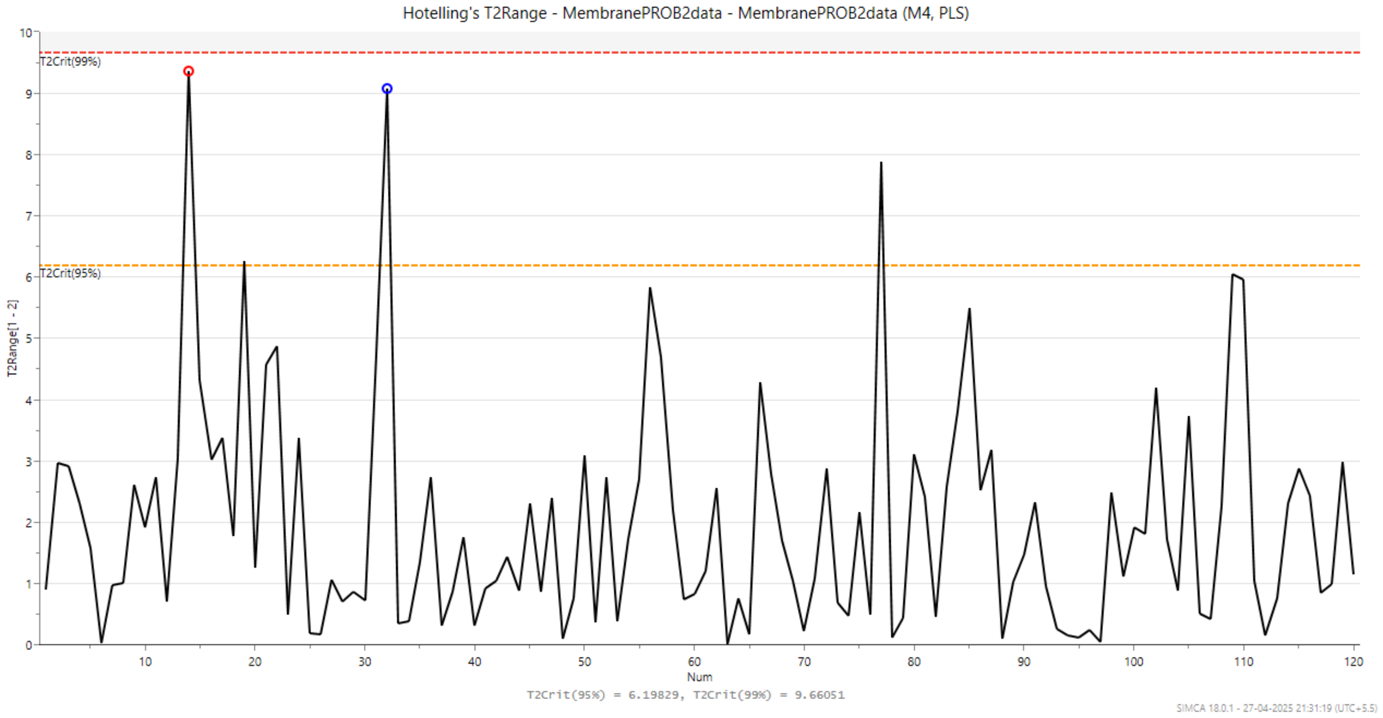
Overall, the second component primarily captures a relationship wherein larger values of Y are strongly associated with larger values of X8 but smaller values of X1, X2, X4, X5, and X6 (particularly X4). This component explains an additional 13.9% of the variance in the X-variables (R2X = 0.139).

Scores Plot Interpretation

The scores plot shows the distribution of the 120 membrane samples in the model space as defined by the first two PLS components, t1 and t2. The first component, on the x-axis, accounts for 34.5% of the X-variance, and the second component, on the y-axis, accounts for a further 13.9%. The ellipse represents the Hotelling's T^2 statistic at a 95% confidence level, which marks the region anticipated for normal samples in the model.



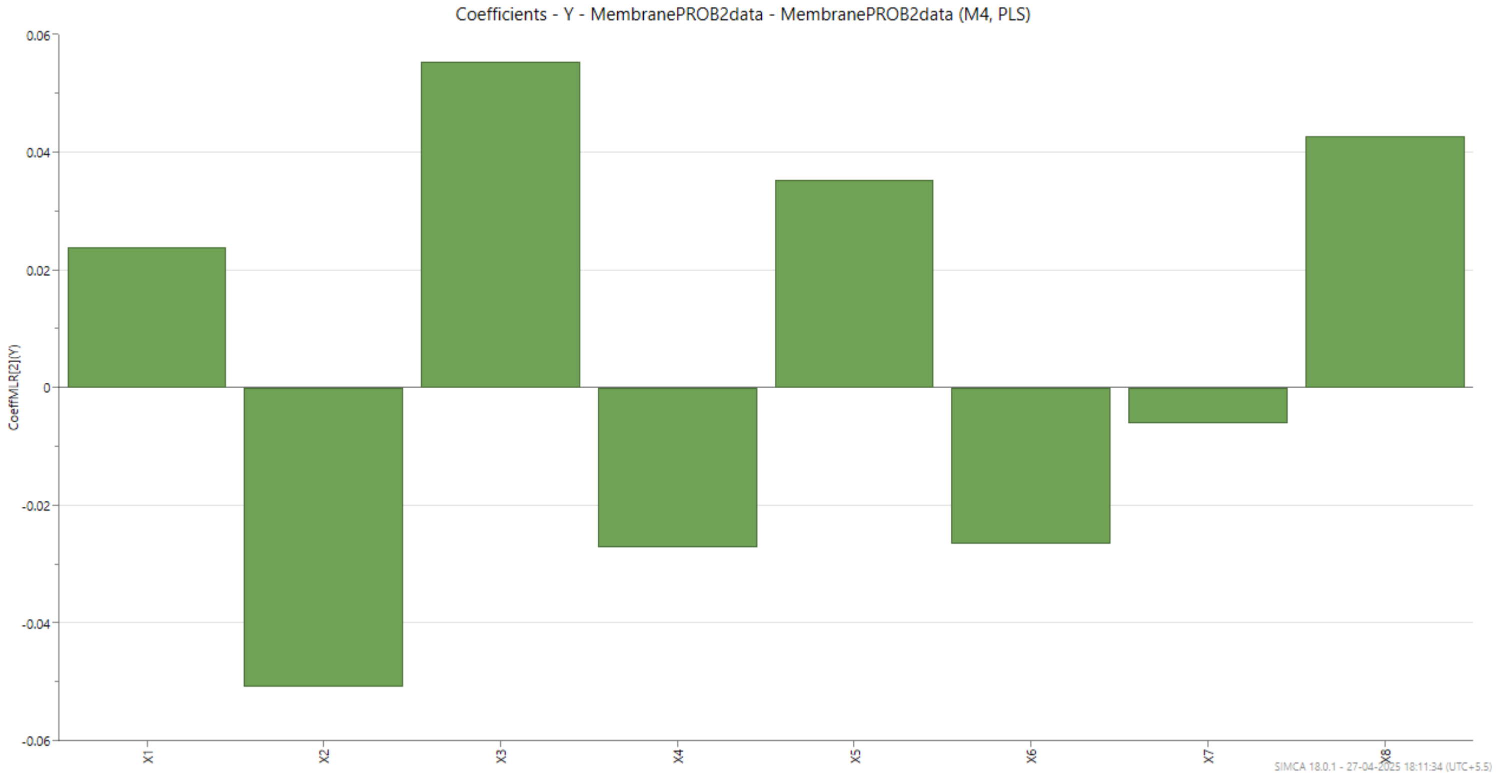
Most of the samples fall inside this ellipse, indicating that they fit well with the model that results from most of the data. However, some samples are outside or near the boundary, which means they might be outliers or possess some special characteristics not entirely explained by the principal relations. Specifically, sample 14 is far from the ellipse in the negative t1 and negative t2 quadrant, indicating that it greatly differs from the typical pattern.



Overall, the plot shows reasonable structure but highlights potential outliers (*especially sample 14*) and the spread of samples according to the two main dimensions of variation captured by the PLS model, which can also be seen by the line plot of Hotelling's T^2.

PLS Regression Coefficients Plot Interpretation

The above plot shows the two-component PLS model's scaled and centred regression coefficients, showing the modelled relationship between each predictor variable (X1-X8) and the response variable (Y)1. These coefficients are the net effect of each predictor on the response in the context of the model.



- Positive Influence: Variables X3, X8, X5, and X1 have positive coefficients. As per the model, this means that all increases in these variables are related to increases in the response variable Y. The strongest positive influence is shown by X3, followed by X8 and then X5.

- Negative Influence: Variables X2, X4, X6, and X7 contain negative coefficients. According to the model, this indicates that increases in these variables relate to decreases in the response Y. X2 shows the most extreme negative influence.
- Magnitude of Influence: The variables with the most significant coefficients (positive or negative) are the most influential in the prediction of Y in this model. These are X3 (strong positive), X2 (strong negative), and X8 (strong positive). Variables X4, X5, and X6 have a moderate influence, X1 has a lesser positive impact, and X7 has a negligible effect on the response based on this model.

In conclusion, the model indicates that the attainment of higher values of the response Y is mainly linked to raising X3, X8, and X5, while lowering X2.

Conclusion

Projection to Latent Structure (PLS) regression was used in this research to explore interdependencies among membrane characteristics and their relation to membrane quality, offering a robust alternative to Multiple Linear Regression (MLR) in cases of multicollinearity, non-linearity, and potential overfitting. A two-component PLS model was the best fit, explaining 45.5% of the variance in the response variable and exhibiting a moderate cross-validated predictive ability ($Q^2 = 0.348$). The analysis identified X3, X5, and X8 as significant variables that had a positive relationship with membrane quality, and X2 with a strong negative impact. While the model can explain high variance, its prediction accuracy is moderate, and a small number of outliers were discovered, suggesting that future research may focus on augmenting the dataset size, bringing in non-linear modelling techniques, and employing strict variable selection techniques to improve model robustness and prediction accuracy.