Name: Natalya Lavrenchuk (ID: 2141882)
Course: SP25 Machine Learning for Process Engineering
Instructor: Professor Pierantonio Facco
Assignment: Homework #1 – PLS Model
Date: May 26, 2025

UNIVERSITÀ DEGLI STUDI DI PADOVA

The purpose of this assignment is to build a Partial Least Squares (PLS) regression model to predict the final biomass concentration in a batch yeast cultivation process based on historical online process data. MATLAB PLS_Toolbox was used to construct the model and to answer the following questions.

Question 1: Data visualization for both calibration datasets X and Y and discussion

Process data was imported into MATLAB and calibration data was plotted to more easily observe trends across the 7 process variables over all calibration batches (Figure 1, plots A–G). The final yeast biomass concentrations were also summarized and compared across these batches (Figure 1, plots H–I).

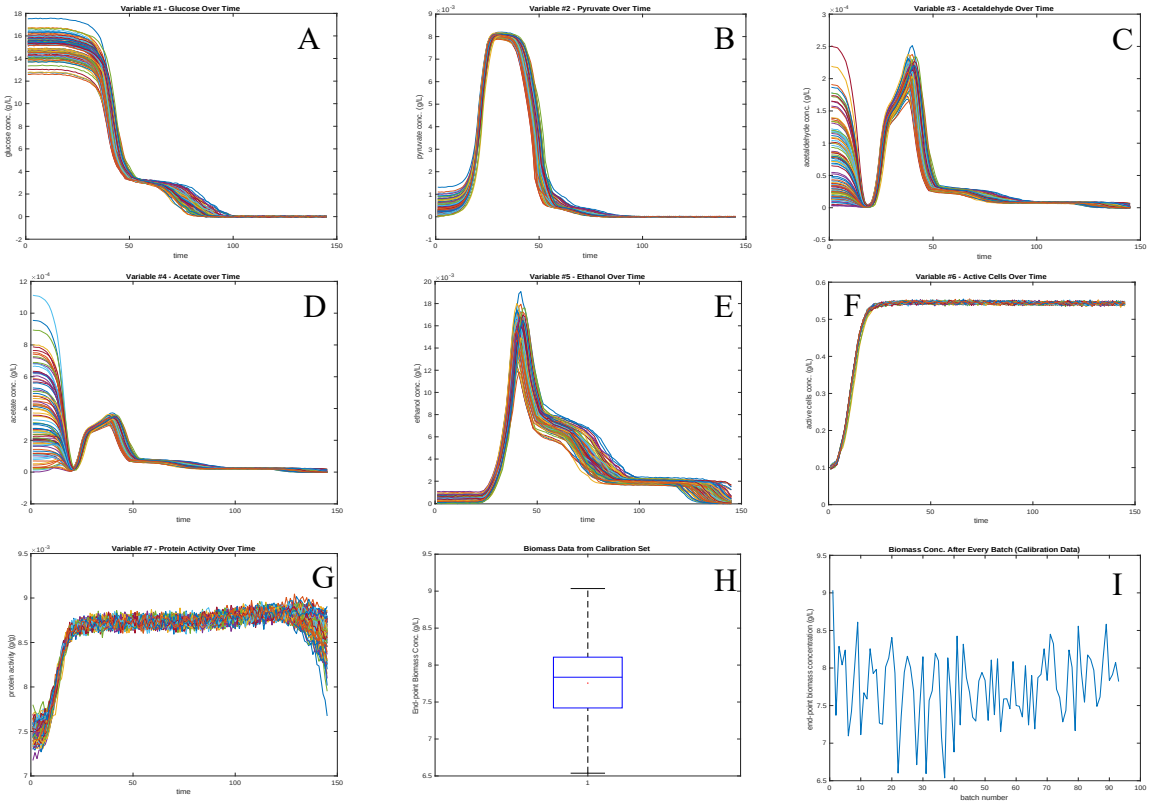

Figure 1. Visualization of the calibration dataset. Plots A–G show the time evolution of the 7 monitored process variables across 93 yeast cultivation batches, each recorded over 145 time points. Plot H presents a boxplot summarizing the final biomass concentrations (output variable), while plot I shows the individual end-point biomass values for each calibration batch.

The trends across all batches seem fairly consistant with no noticable irregularities, which demonstrates a relatively controlled and reprudicble proces between batches. With the increased growth of active yeast cells (Fig 1F) we see respective increase in protein activity in the cell culture (Fig. 1G), increase in glucose consumption (Fig 1A), and other expected metabolic patterns in the data (spikes in pyruvate, ethanol, etc.). Plots H and I show that the final biomass concentrations are also regular with minimal variation between batches. These trends suggest the calibration data represents stable process behavior under normal operating conditions (NOCs).

<u>Question 2:</u> Build a PLS model for the prediction of the end-point biomass from the time trajectories of the process variables

Prior to constructing the PLS model, the data was scaled and unfolded to match the required format for model construction. The original dataset, structured as a 3D array with dimensions batch × variable × time (e.g., 93 × 7 × 145), was reshaped using a batch-wise unfolding strategy into a 2D matrix (e.g., 93 × 1015), where each row represents one batch, and each column corresponds to a variable at a specific time point. Batch-wise unfolding was selected over variable-wise unfolding since it retains the full time profile of each process variable and is better suited for modeling the development of batches and quality outputs. In addition, autoscaling was applied to the data, by subtracting the mean from each data point and dividing by its standard deviation. This allows for comparability across variables of different original magnitudes.

In constructing the model, the number of latent variables (LVs) was selected based on the Root Mean Squared Error of Cross Validation (RMSECV) curve shown in Figure 2. The RMSECV decreases up to 4 LVs and then stays constant above that value, indicating not much added

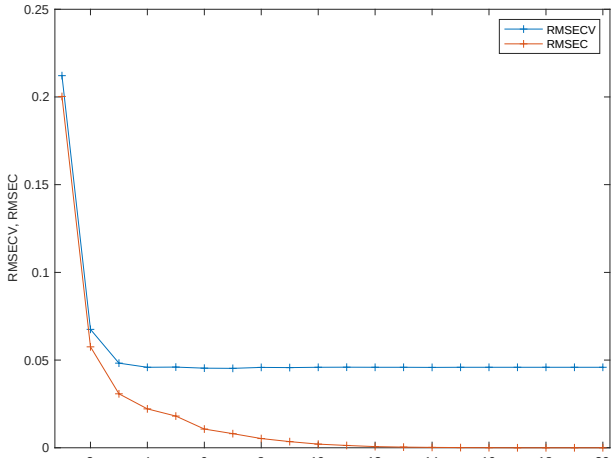

Figure 2. Cross-validation and calibration error (RMSECV and RMSEC) as a function of the number of latent variables (LVs). The optimal number of LVs was selected to be 4 LVs.

benefit with higher LVs. For this reason, 4 LVs were chosen as the optimal point between balancing model complexity and predictive performance. This minimizes chances of overfitting the model while still capturing the relevant variance in the input data. With 4 LVs, the model explains approximately 47.5% of the variance in X and 99.8% of the variance in Y, as shown in the PLS table (Table 1). This suggests the model captures the underlying structure of the process effectively and is suitable for predicting the end-point concentration of yeast in new batches.

Table 1: Explained variance in X and Y by latent variables (LVs) for the PLS model.

LV	% Variance in X	Cumulative X	% Variance in Y	Cumulative Y
1	32.26	32.26	82.4	82.4
2	9.41	41.68	16.15	98.55
3	2.54	44.22	1.04	99.58
4	3.26	47.48	0.2	99.78
5	6.11	53.59	0.07	99.86
6	1.92	55.51	0.09	99.95
7	2.41	57.93	0.02	99.97
8	1.95	59.88	0.02	99.99

<u>Question 3:</u> Plot and discuss critically the X score plot of LV1 vs. LV2

The score plot (Figure 3) shows how the calibration batches are distributed in the reduced latent variable space defined by LV1 and LV2. As can be seen from the figure, LV1 alone captures 32.3% of the variance in X with LV2 accounting for an additional 9.4% of the variance in X.

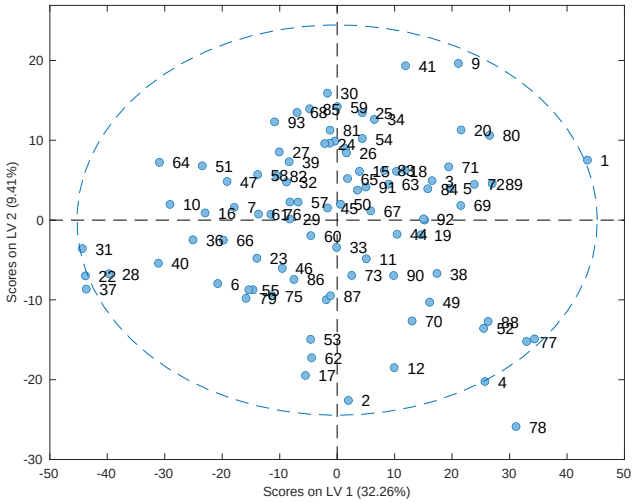

Figure 3. X score plot showing the projection of calibration batches onto latent variables LV1 and LV2.

Most of the points stay near the origin, indicating quite consistent behavior between batches. Some batches lie outside the edges of the Hotelling's T^2 ellipse (blue dotted ellipse), indicating some batch variability or minor deviations from the NOCs, such as in batch number 78. No apparent trends in the data are seen and data points are generally randomly scattered suggesting minimal signs of process drifts between batches. In addition, there doesn't appear to be any extreme outliers, further suggesting all calibration batches perform relatively consistently without major abnormalities.

Question 4: Plot and discuss critically the weights for the first LV

A graph depicting the weight contributions of each of the seven process variables to the first latent variable (LV1) across all time points was constructed (Figure 4). The weights plot shows how each variable contributes at each time instance within the batches to the prediction of final biomass concentration. Higher absolute values indicate greater importance in modeling the end-point yeast concentration.

Many of the variables, including acetaldehyde, acetate, and ethanol, show a strong influence only during specific time intervals within the batch, potentially indicating certain important metabolic activity for yeast growth. In contrast, glucose and pyruvate exhibit decreasing weight contributions as the batch progresses. On the other hand, protein activity and active cells

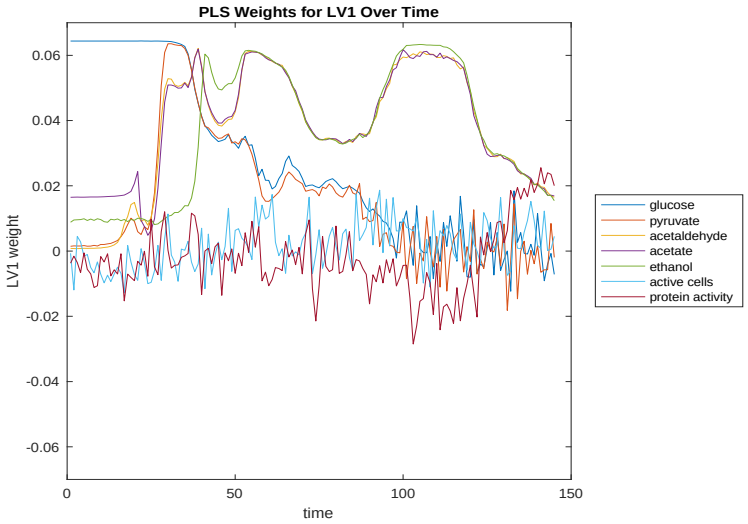

Figure 4. PLS weights for the first latent variable (LV1), showing the relative contribution of each process variable across time.

demonstrate increasing weight importance over time, suggesting they also play a key roll in predicting final biomass concentration.

<u>Question 5:</u> Plot and discuss critically the plot of the regression coefficients

The same results found in the weight plot (Question 4) on impact of process variables for predictability of biomass concentration, can be found by plotting the regression coefficients of the model over time (Figure 5). In the same way, higher absolute values indicate stronger impact on the model's output. We find similar trends hold, many variables have strong impact on the model's predictability at only certain time instances within the batch. Starting glucose concentration has particularly high impact on the final biomass concentration as well. Together, the weights and regression coefficients help identify which variables, and when during the batch, they are most influential in predicting the final outcome.

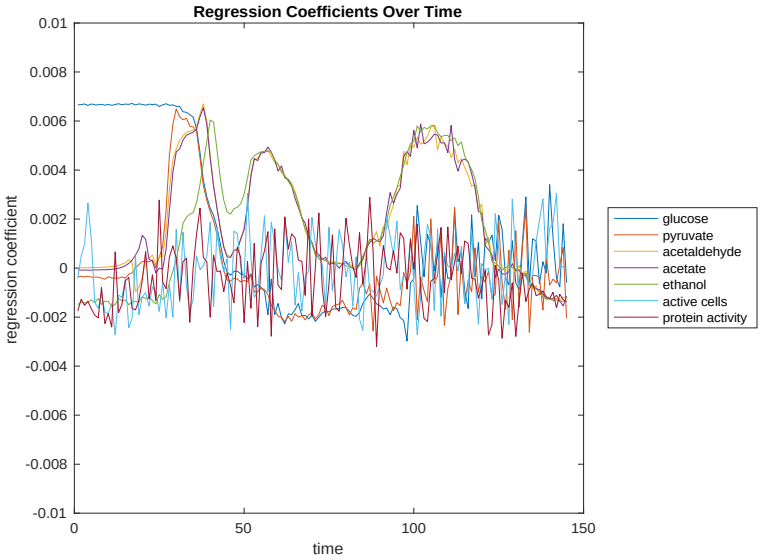

Figure 5. Regression coefficients over time for each of the seven process variables. The plot illustrates the influence of each variable at different time points on the prediction of the final biomass concentration.

<u>Question 6:</u> Verify if the linear structure of the PLS model is appropriate

The scores of X an Y, stored in T and U respectively, in latent variable 1 were plotted to observe the appropriateness of the linear structure of the PLS model for this application. The resulting graph is shown in Figure 6. A red linear regression line has been added to visually assess the degree of linearity between the components. As can be seen, the strong linear relationship between T and U in between T and U in LV1 indicates that the linear PLS model structure is

appropriate, as it effectively captures the underlying linear correlation between the predictors and the response.

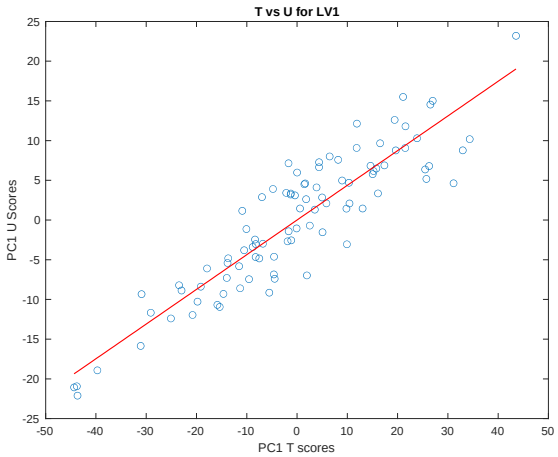

Figure 6. Score plot comparing T (X scores) and U (Y scores), with linear regression line indicating strong linearity between predictor and response variables.

<u>Question 7:</u> Build for the calibration dataset a Q vs. T2 monitoring chart with the respective 95% confidence limits

A Q vs T^2 monitoring chart was built for the calibration dataset of the model and depicted in Figure 7. A 95% confidence interval was calculated and added to the graph to demonstrate the regions of normal operation. Some batches, such as Batch 78, exhibit high T^2 values, indicating that the batch lies far from the model center in the score space, possibly suggesting some unusual behavior. In contrast, Batches 87 and 89 have high Q values, meaning they deviate from the model in the residual space, indicating poorly explained variance. Batch 43 is high in both T^2 and Q, suggesting it deviates both in modeled and unmodeled cases, meaning it is likely an outlier that may have deviated from NOCs and should be further examined.

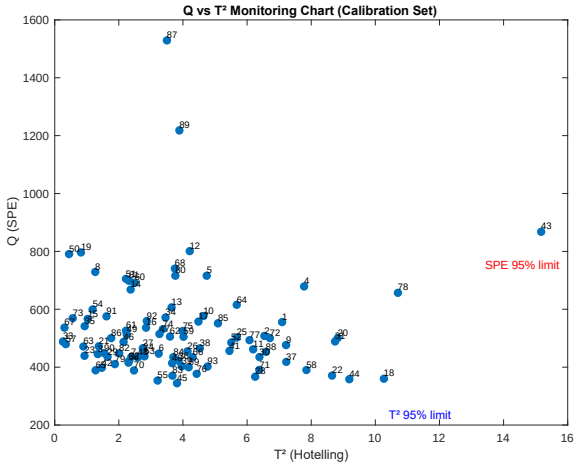

Figure 7. Q (SPE) vs. T² (Hotelling) monitoring chart for the calibration set. The 95% confidence limits are shown as reference lines. Batches above the horizontal (Q) or vertical (T²) limits indicate deviations from the model.

Question 8: Build the matrices of the residuals E and F

The X residual matrix, stored in E, shows mostly low residuals across the dataset, indicating the PLS model captures the majority of variance in the input data. Some deviations appear in specific variable-time instances and suggests some model limitations. The Y residual matrix, stored in F, shows overall low error in predicted biomass, with a few outlier batches showing poor predictability. These correspond to batches with high Q or T² values, confirming their deviation from normal behavior.

Question 9: Compute the mean relative error MRE for the calibration Y matrix

The calculated mean relative error (stored in variable MREc) for the calibration set is 0.0023, or 0.23%. This low value indicates that the predicted yeast concentrations are very close to the true values. A small relative error demonstrates that the PLS model is highly accurate on the training data and has effectively captured the relationship between the process variables and the output. The very low MRE, however, may indicate that the model is overfitting the data and should be tested with validation data.

<u>Question 10:</u> Plot and discuss the parity plot in calibration

A parity plot depicting the predicted Y vs the measured Y in the calibration data is plotted in Figure 8. As can be seen, the predicted values of yeast biomass concentration from the PLS model align very well to the measured, true value of the actual yeast concentration from the calibration batches. The R^2 value was calculated to be 0.998, or 99.8%, which means the model is highly predictive of the training set data, and potentially overfitting the data.

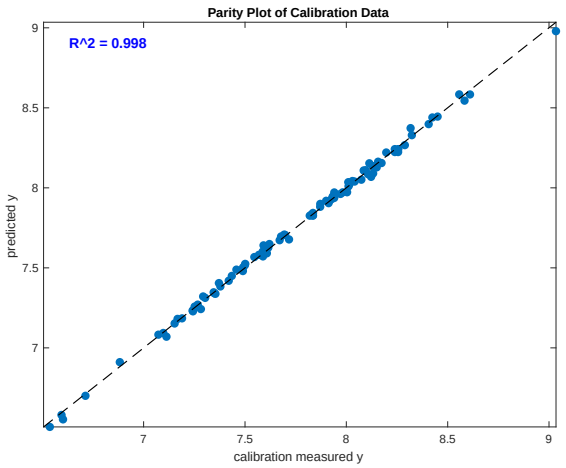

Figure 8. Parity plot comparing predicted versus measured end-point biomass concentrations for the calibration dataset.

<u>Question 11:</u> Project the validation batches into the PLS model, estimate the quality variables, calculate the errors of estimation, compute the MRE in validation and discuss them with respect to the variability of the real measurements

The two validation batches were projected into the PLS model to estimate the final biomass concentration. The predicted values were compared against the true measured values for these batches, yielding an absolute error between predicted and measured of 0.0435 and 0.0441. The mean relative error (MRE) for validation was calculated to be 0.003, which was slightly higher than the MRE for the calibration data (0.0023), as should be expected. This low MRE means the model generalizes well and has good predictive performance on unseen test data. It should be noted that a validation size of only two batches may be causing the model to perform better than expected. Increasing the amount of validation batches would allow for a better understanding of model performance on test data.

<u>Question 12:</u> Plot and discuss the parity plot in validation

The parity plot shows the predicted biomass concentrations from the PLS model against the actual measured values for the two validation batches (Figure 9). The points lie close to the diagonal line, indicating good predictive performance. An R^2 value of 0.853 is slightly poorer than the calibration data, which is to be expected, but still appropriate for the application. It also confirms a strong linear relationship between predicted and true values, indicating that the model performs well on the validation set.

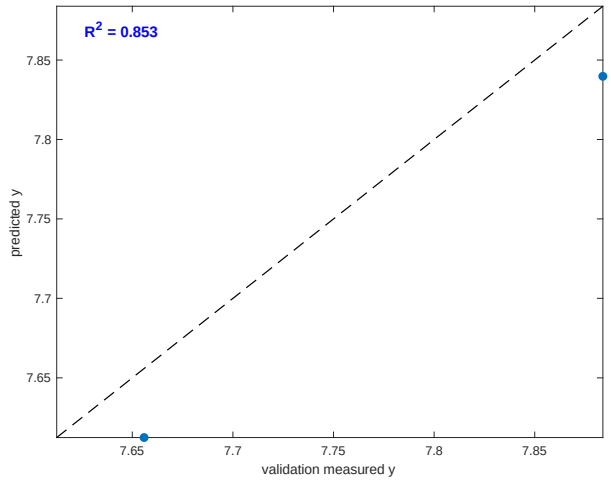

Figure 9. Parity plot of predicted vs. measured biomass concentration for the validation set.

<u>Question 13:</u> discuss the projection of the validation batches in the Q vs. T2 monitoring chart

Two validation batches (red) are projected onto the Q vs T^2 plot (Figure 10). One validation batch lies within the 95% confidence region for both T^2 and Q, meaning it aligns well with the modeled normal operating conditions defined by the calibration batches. The other exceeds the Q threshold, suggesting deviation in the residual space, while still being within the modeled T^2

limits. Further investigation using variable contribution plots on Q should be performed to understand the deviations.

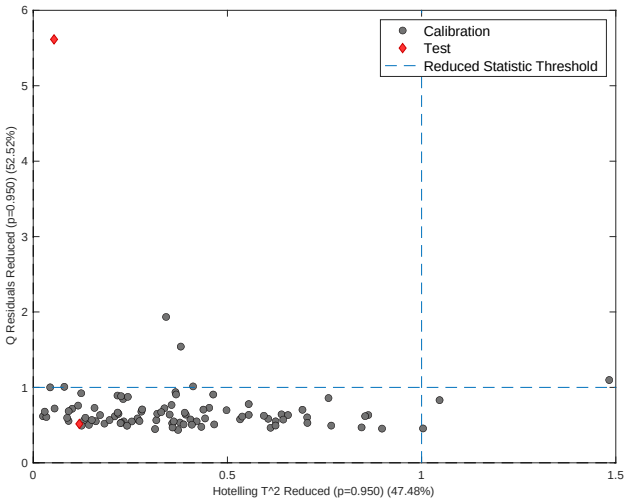

Figure 10. Projection of validation batches onto the Hotelling's T^2 vs. Q residuals space.

<u>Question 14:</u> Discuss critically the validation batches for both prediction performances and their position in the Q vs. T2 monitoring chart; if either Q or T2 are out of the confidence limits build the contribution plots to understand what variables time trajectories and what instants deviate from the NOCs

To further examine the variability, a contribution plot was generated to investigate the contributions of each variable to the SPE for the second validation batch, which exceeded the Q limit (Figure 11). Variables 6 (active cells), 2 (pyruvate), and 4 (acetate) contribute the most to the high Q value, indicating they are the main sources of deviation from normal operating conditions, potentially due to process shifts or sensor issues.

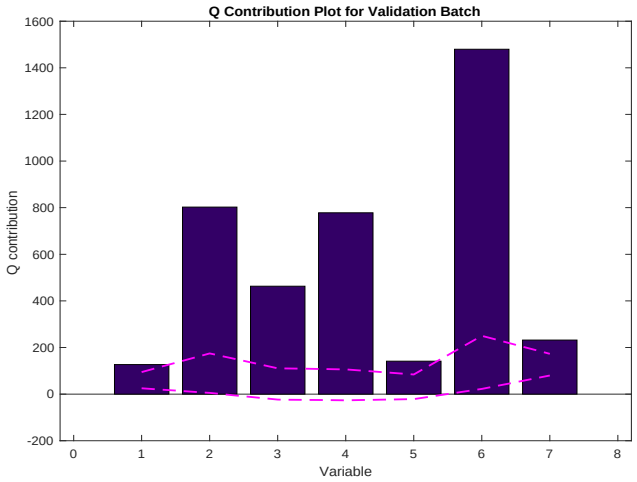

Figure 11: Q contribution plot for the second validation batch. Bars represent variable-wise contributions to the SPE, with magenta dashed lines showing ±2σ bounds from the calibration.

<u>Question 15:</u> If LV is the number of selected latent variables in the PLS model, what happens to the prediction performance if a total number of LV+3 is selected?

A new PLS model with LV+3, in this case with 7 LVs, was built in order to assess the contribution of higher LVs to model performance. Switching from 4 to 7 latent variables improved numerical prediction performance: RMSE and MRE decreased in both calibration and validation, and R^2 increased—especially in validation (from 0.85 to 0.94), indicating better generalization. However, the Q vs T^2 monitoring chart suggests a trade-off in that more calibration points now fall outside the statistical limits, and the validation point appears more extreme. This suggests potential overfitting—while the model captures more variation, it may

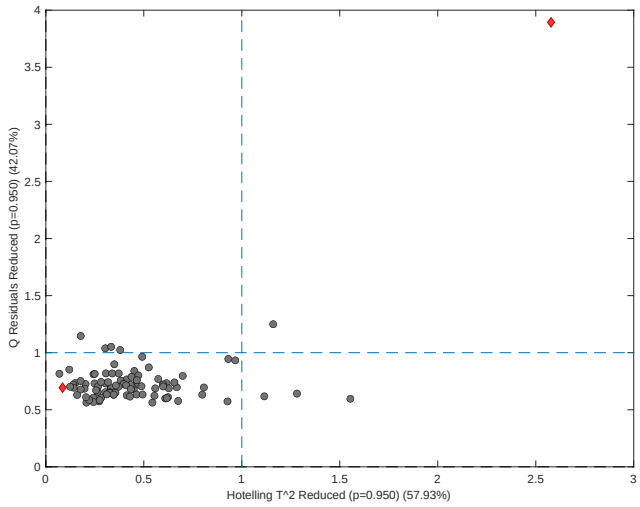

Figure 12. New model with 7 LVs projection of calibration and validation into Q vs T^2 space.

also be modeling noise or minor fluctuations not representative of normal operation. While the 7LV model is more accurate in prediction, it is less robust for monitoring, and the 4LV model may offer a better balance between prediction and process interpretability.

Conclusion

In this project, a PLS regression model was successfully developed to predict final biomass concentration from fermentation process data in yeast cultivation. The model demonstrated strong predictive accuracy, especially with 4 latent variables, balancing performance and interpretability. Monitoring tools like Q vs T^2 charts and contribution plots provided valuable insight into batch deviations and variable influence. Overall, the model proved effective for both prediction and process understanding.