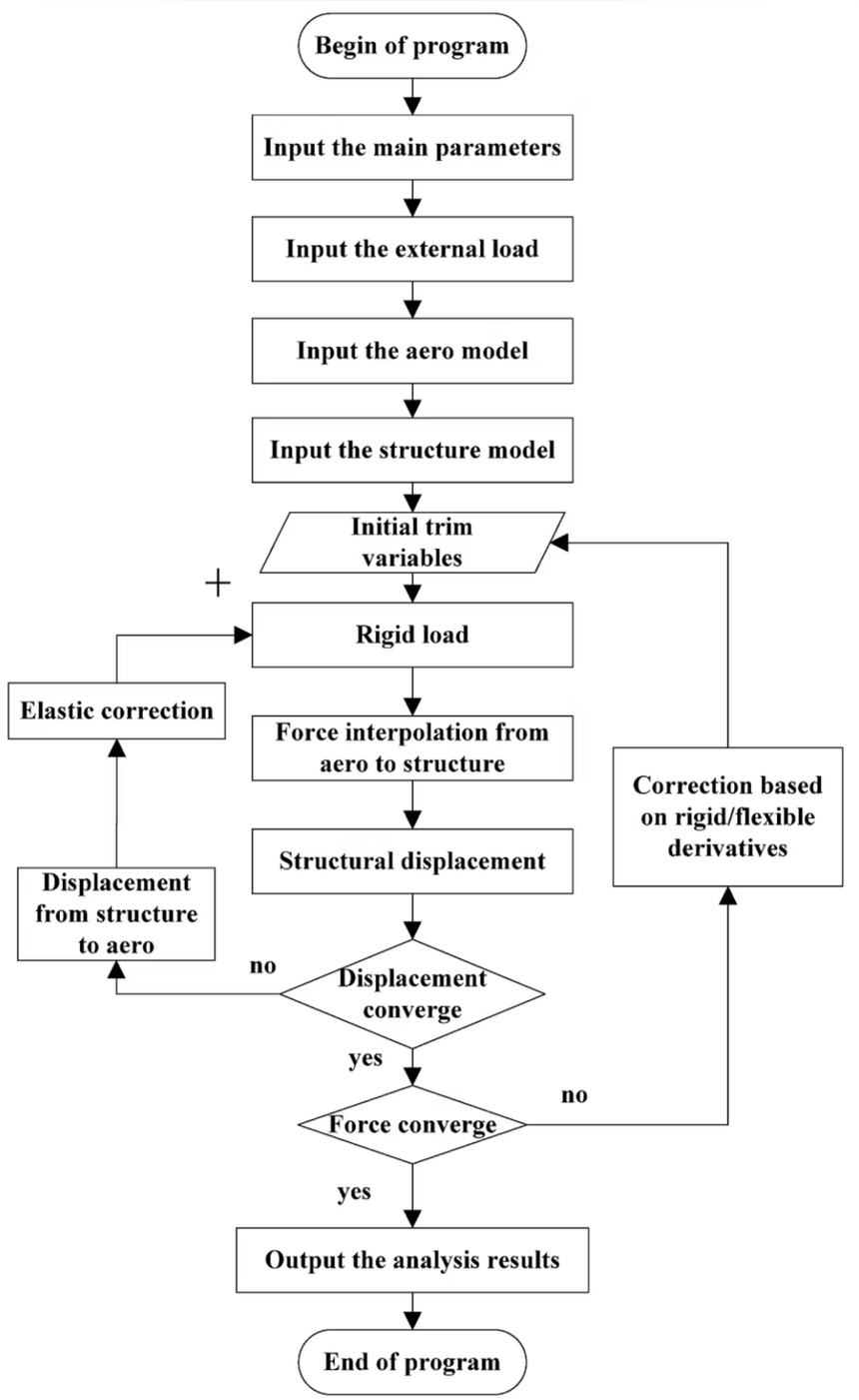
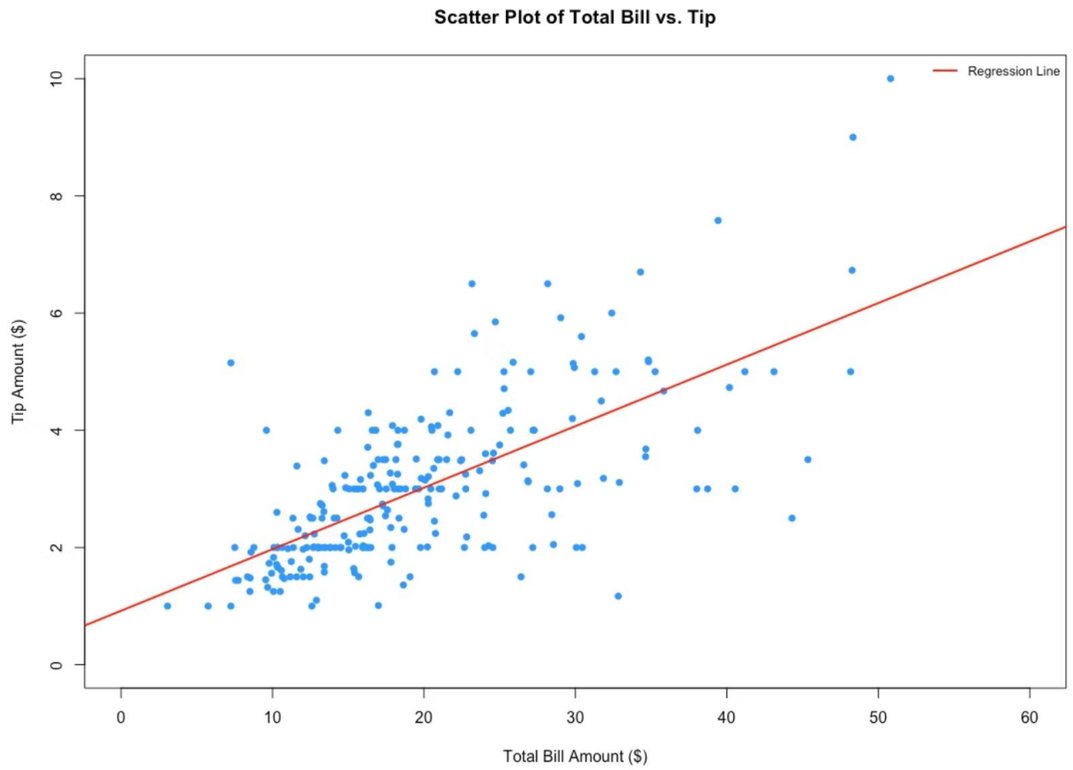
Introduction to Partial Least Squares (PLS) Regression

A powerful statistical technique for handling complex, high-dimensional data



What is Classical Linear Regression?



Core Purpose

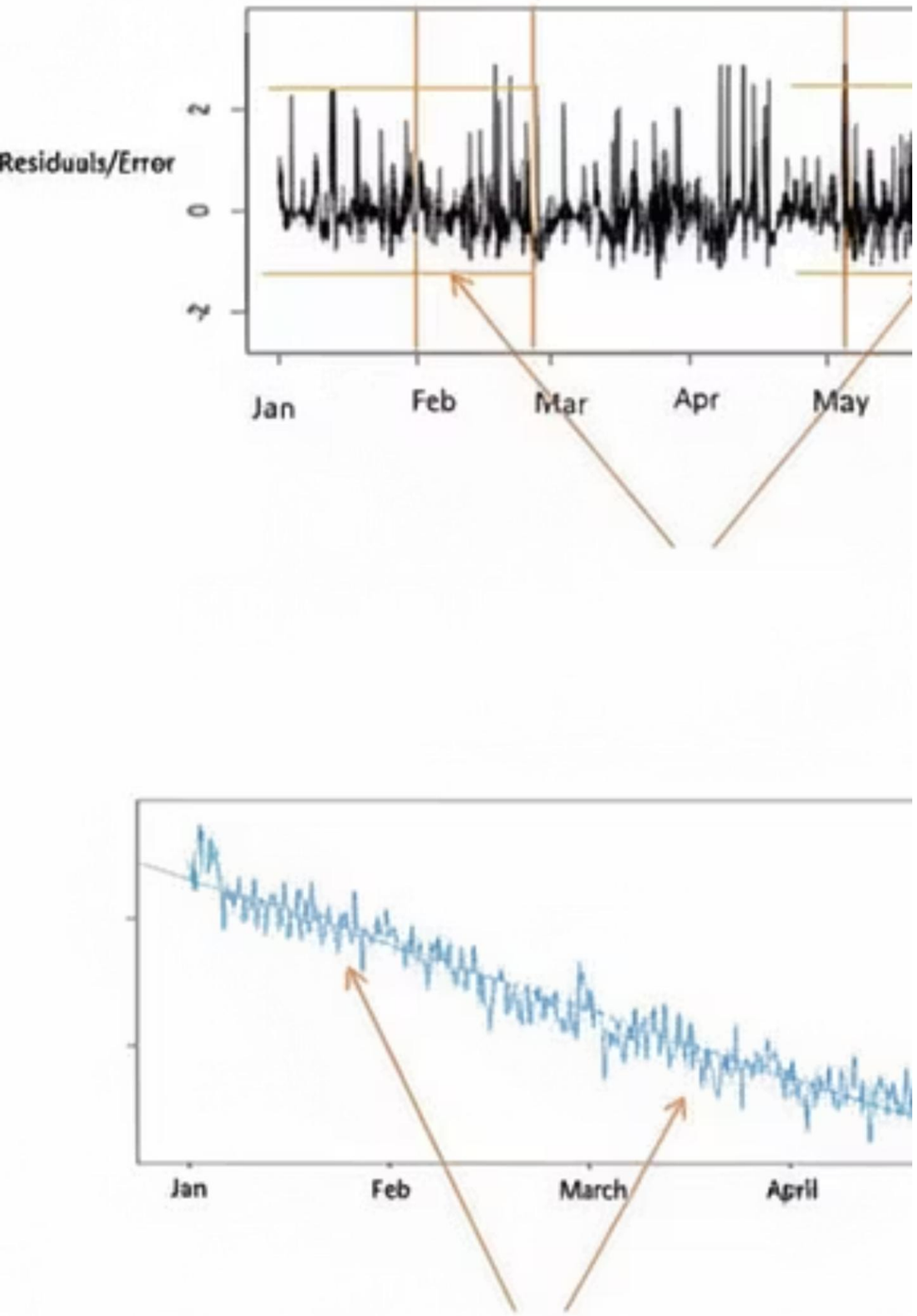
Models the relationship between predictor variables (X) and response variable (Y)

Key Assumptions

Assumes predictors are independent and fewer in number than observations

Estimation Method

Uses Ordinary Least Squares (OLS) to estimate regression coefficients



Limitations of Classical Regression

Multicollinearity

When predictors are highly correlated, coefficient estimates become unstable and unreliable

Dimensionality Problem

More predictors than observations makes matrix inversion mathematically impossible

Overfitting Risk

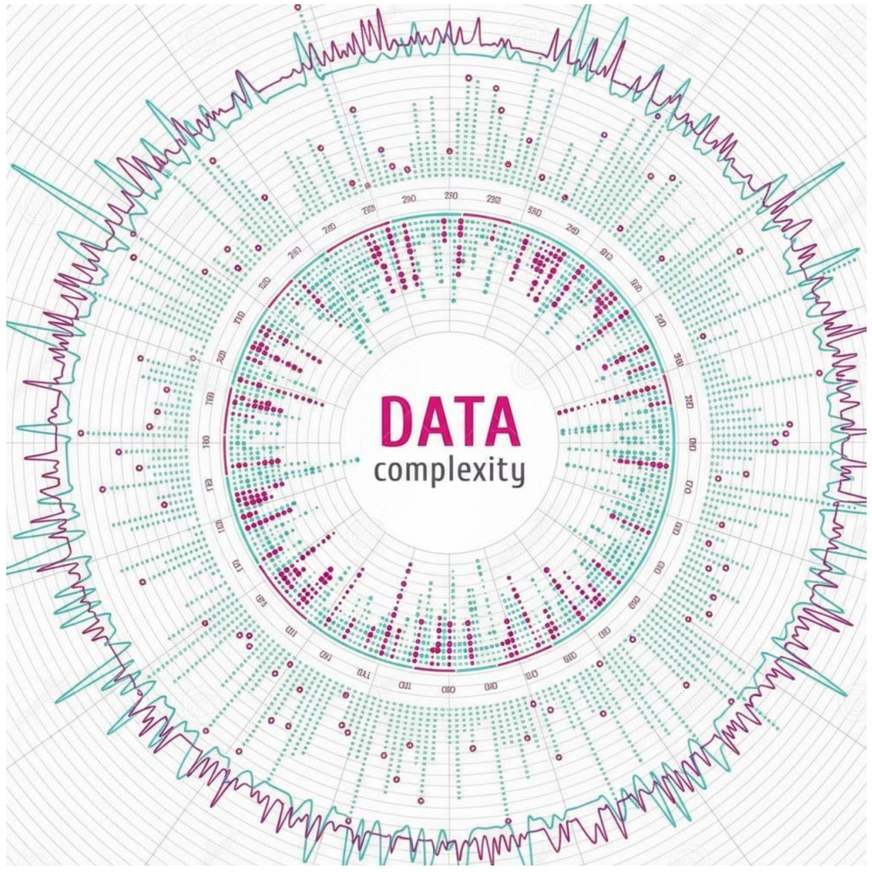
Too many variables lead to excellent fit on training data but poor prediction on new observations

Why Do We Need Something More?

Real-world datasets frequently contain numerous, highly correlated predictors—think of spectroscopic data with hundreds of wavelengths, genomic studies with thousands of genes, or economic models with interconnected indicators.

Classical regression simply cannot handle these scenarios effectively. We need methods that embrace complexity rather than break down under it.

> **Goal:** Achieve robust prediction and meaningful interpretation despite multicollinearity and high dimensionality



Principal Component Regression (PCR): How It Works

Principal Component Regression (PCR) offers an alternative to classical regression by addressing issues like multicollinearity and high dimensionality. It achieves this through a structured, two-step process primarily focused on transforming the predictor variables before modeling the response.

Step 1: Principal Component Analysis (PCA) on X

PCA is performed exclusively on the predictor variables (X matrix) to identify a new set of orthogonal (uncorrelated) variables called principal components. These components are selected to capture the maximum variance within the predictor space, without any consideration for their relationship with the response variable (Y).

Step 2: Regression of Y on Principal Components

After extracting a desired number of principal components, a standard linear regression model is built. The response variable (Y) is then regressed onto these selected principal components, effectively using a reduced and decorrelated set of predictors for the final prediction model.

Crucially, PCR's component extraction in Step 1 is "unsupervised"—it is driven solely by the variance within X, ignoring Y until the second, explicit regression step.

PLS vs Principal Component Regression (PCR)

PCR Approach

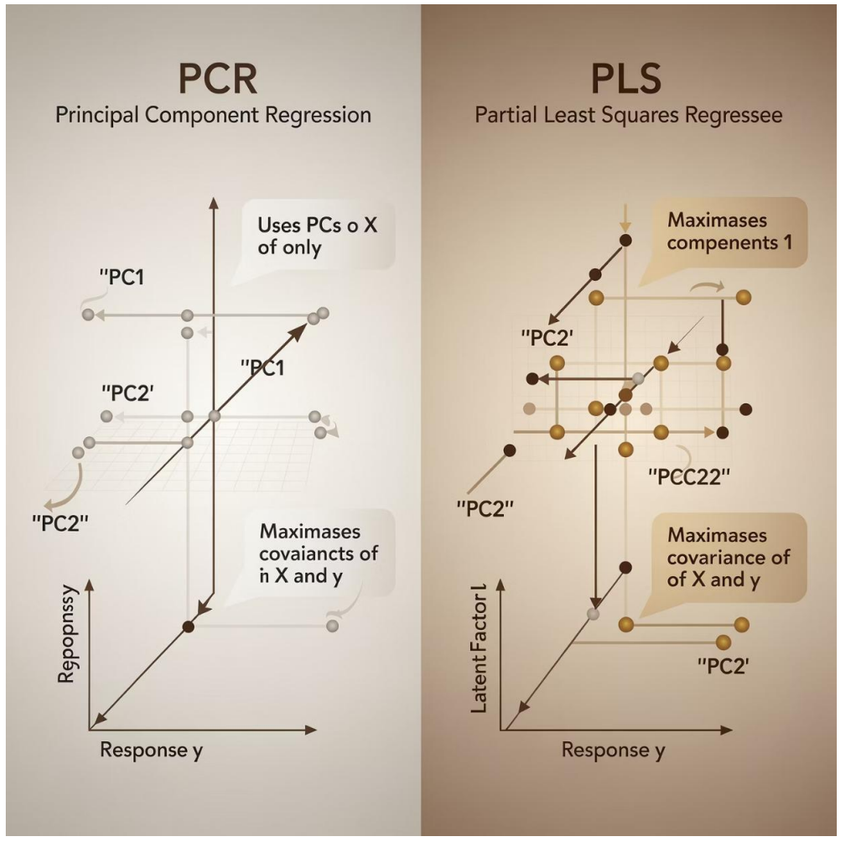
Finds components that explain variance in X only, without considering Y

Components may not be relevant for prediction

PLS Advantage

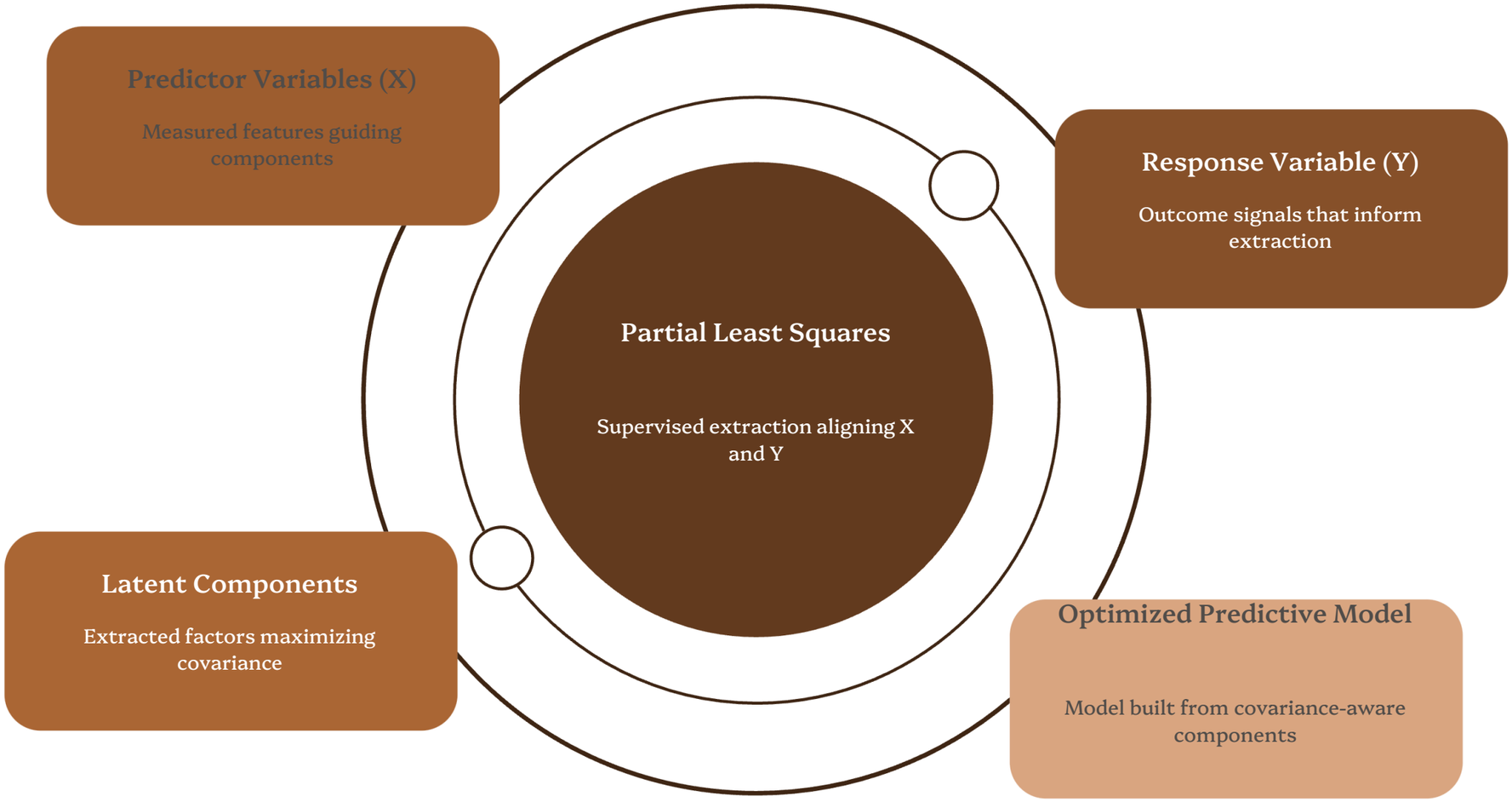
Finds components that explain covariance between X and Y simultaneously

Components are specifically optimised for predicting Y



Partial Least Squares (PLS): How It Works

Unlike Principal Component Regression (PCR), PLS takes a fundamentally different approach to dimension reduction. It's a **supervised** method where the response variable (Y) actively guides the creation of latent components.



PLS works by constructing a set of new latent variables (components) from the predictors (X) that not only explain the variance in X but, more importantly, also **maximize the covariance between X and Y**. This ensures that the extracted components are maximally relevant for predicting the response, directly addressing the limitations of methods that consider X and Y separately.

> This "supervised" component extraction is what gives PLS its predictive power, especially in the presence of multicollinearity and high-dimensional data.

Partial Least Squares Regression: The Big Idea

Hybrid Approach

Combines the dimensionality reduction of PCA with the predictive power of regression

Focused Extraction

Finds latent components that maximally explain the covariance between X and Y

Smart Reduction

Reduces dimensionality whilst maintaining focus on predicting the response variable

How PLS Works Mathematically

01

Simultaneous Decomposition

PLS decomposes both X and Y matrices: $X = TP^T$ and $Y = TQ^T + E$

02

Key Components

T represents scores (latent variables), P and Q are loadings, E captures residuals

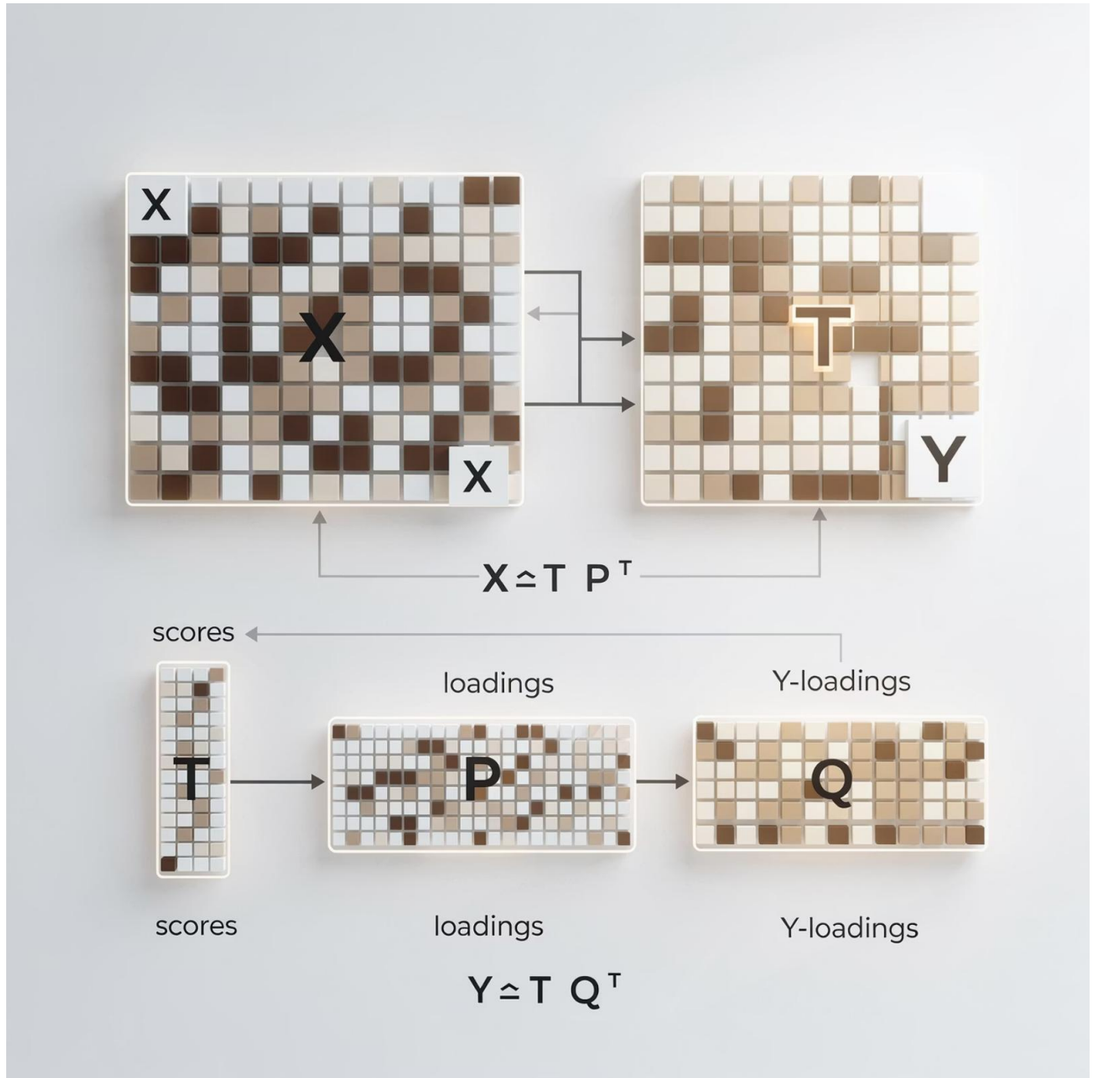
03

Optimisation Criterion

Components (columns of T) are extracted to maximise covariance between X and Y

Key Insight: Unlike classical regression, PLS creates new variables (latent components) that are optimal linear combinations of the original predictors

PLS Matrix Decomposition: Visual Overview



The matrices **T**, **P**, and **Q** form the core of PLS decomposition.

T (scores matrix) contains the new latent variables or components. These components capture the essential information from both the predictor (X) and response (Y) variables. Each column of T represents a component, which is a linear combination of the original predictors, designed to maximize the covariance between X and Y.

P (X loadings matrix) shows the weights or contributions of each original X variable to each latent component in T. It describes how the original predictors relate to these newly formed latent components.

Q (Y loadings matrix) illustrates how the response variable(s) Y relate to the latent components. It describes the relationship between Y and the extracted components.

This decomposition allows PLS to effectively reduce dimensionality while maintaining strong predictive power, as the components in T are specifically chosen to explain both the structure within X and predict Y simultaneously.

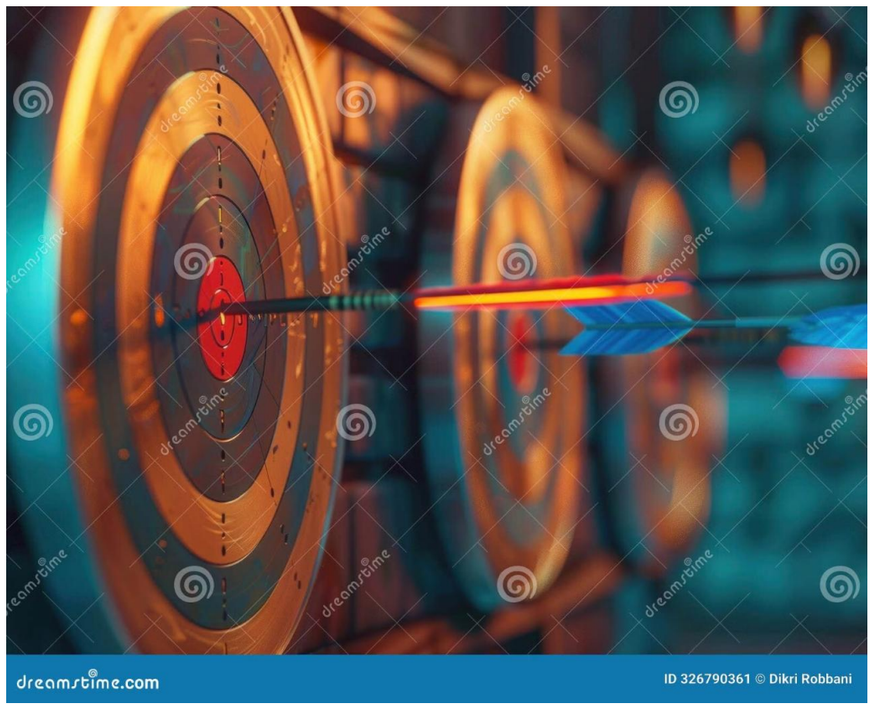
Key Differences: PCR vs PLS

While both PCR and PLS address multicollinearity and high dimensionality, their fundamental approaches to component extraction lead to distinct advantages in different scenarios.

Component Extraction	Unsupervised; solely based on variance within predictor variables (X).	Supervised; based on maximizing covariance between X and response variable (Y).
Objective	To find components that best explain the variance in X.	To find components that best explain both the variance in X and the covariance with Y.
Y's Role	Ignored during the component extraction phase; considered only in the subsequent regression step.	Actively guides the selection and weighting of components, ensuring relevance to prediction.
Best For	Situations where understanding the internal structure of X is paramount, or when Y is not available for component selection.	When the primary goal is robust prediction of Y, especially in presence of strong multicollinearity.
Predictive Power	Potentially weaker for predicting Y, as components are not optimized for this purpose.	Typically stronger for predicting Y, due to direct optimization for the response variable.

This table highlights why PLS is often preferred in predictive modeling tasks, as its component construction inherently prioritizes the relationship with the outcome.

PLS1: Single Response Variable



The Focused Approach

PLS1 is designed specifically for modelling one response variable (Y) at a time. The algorithm extracts latent components that are optimally tuned to predict that single response.

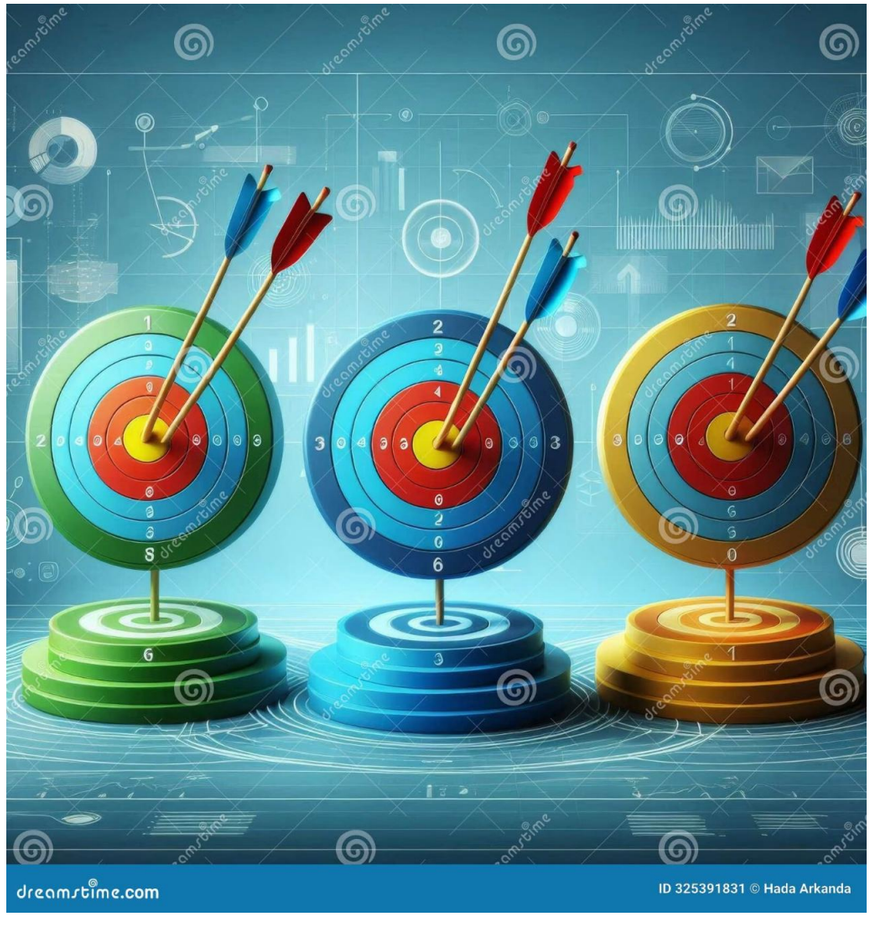
- Simpler interpretation of results
- Component extraction focuses on a single prediction task
- Most common variant in practical applications

PLS2: Multiple Response Variables

The Multivariate Solution

PLS2 handles multiple response variables (multiple columns in Y matrix) simultaneously. It extracts components that explain the covariance structure across all responses at once.

- Useful when responses are conceptually related
- More efficient than running separate PLS1 models
- Captures shared structure across multiple outcomes



Why Preprocess Data?

Scale Matters

Variables measured on different scales (e.g., kilograms vs milligrams) can dominate component extraction unfairly

Centring Benefits

Mean-centering (subtracting the mean) ensures components represent variation rather than absolute values

Standardisation

Scaling to unit variance (autoscaling) puts all variables on equal footing for fair comparison

Typical Preprocessing Steps

1

Step 1: Mean-Centring

Subtract the mean from each variable in both X and Y matrices

2

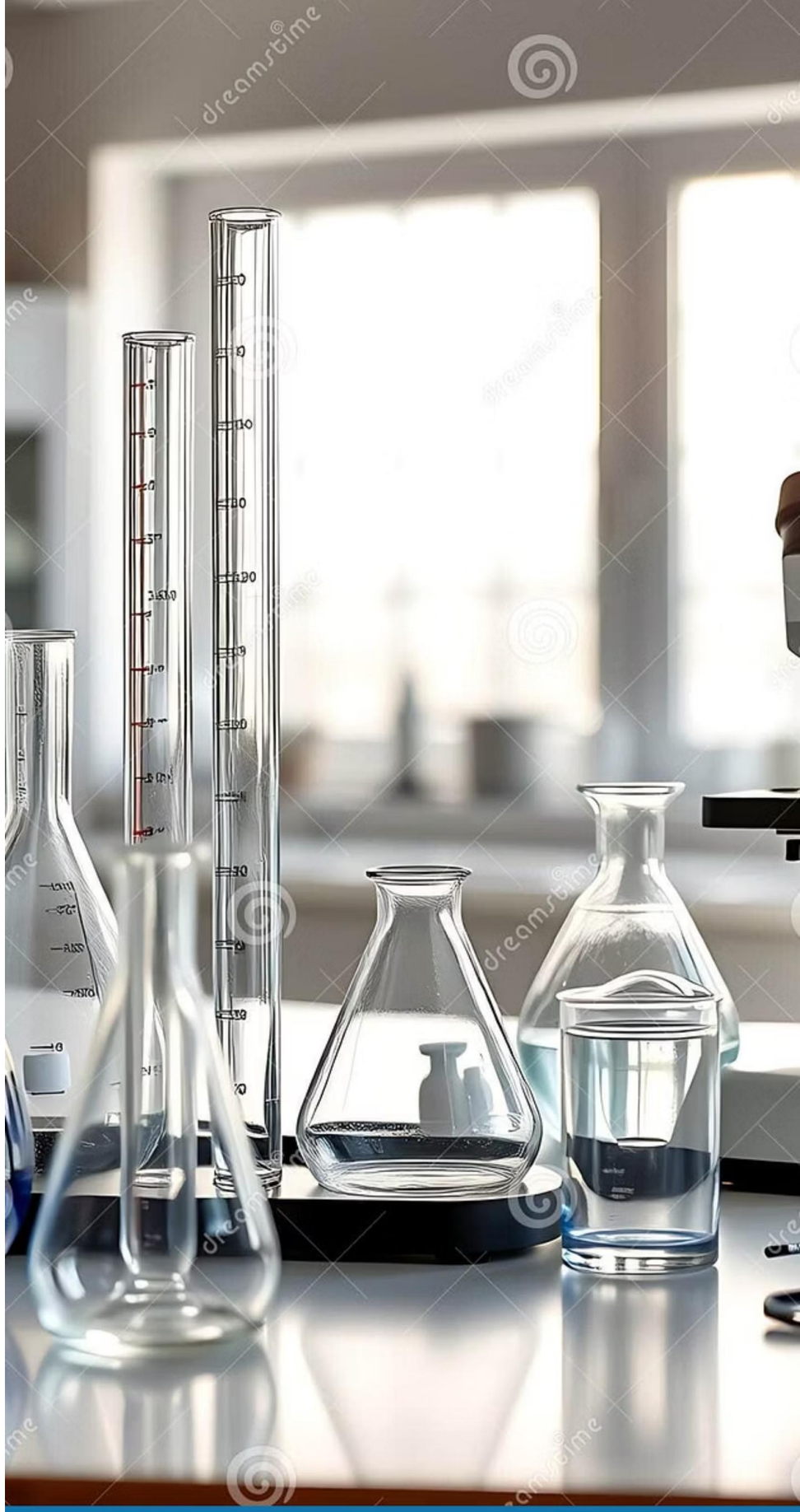
Step 2: Scaling (Optional)

Divide by standard deviation to achieve unit variance if variables have different measurement scales

3

Step 3: Data Quality

Handle missing data through imputation or removal, and identify/address outliers before modelling



Key Outputs and Their Meaning

1

Number of Components

Balance between model fit and overfitting, typically selected through cross-validation

2

Scores Plots

Reveal clusters, outliers, and underlying patterns in the data structure

3

Loadings/Weights

Show variable importance and relationships—which predictors drive the response

4

Model Quality Metrics

Explained variance (R^2) and predicted variance (Q^2) indicate fit and predictive ability

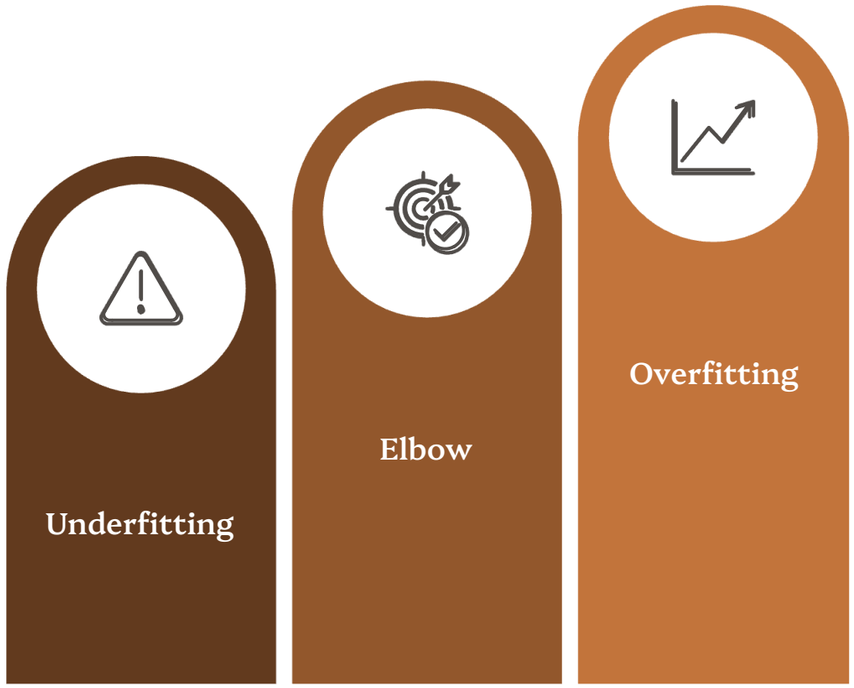
5

Regression Coefficients

Used for making predictions on new data, translating back to original variables

Selecting the Number of Components

The number of components is a critical choice in PLS modeling, directly impacting model performance and interpretability.



1

Underfitting

Using too few components results in an overly simplistic model that fails to capture the underlying relationships, leading to high bias and poor predictive accuracy.

2

Overfitting

Conversely, too many components can cause the model to capture noise and specific characteristics of the training data, degrading its ability to generalize to new, unseen data.

3

Cross-Validation

The standard method for selection involves splitting data into training and test sets, then evaluating predictive performance (e.g., RMSECV or Q^2) for a varying number of components.

4

The "Elbow" Rule

The optimal number of components is often found at the "elbow" point in the cross-validation plot, where adding more components provides diminishing returns or starts to increase the prediction error.

Understanding Scores Plots

Scores plots are a powerful visualization tool in PLS, showing how individual observations are positioned within the new, reduced dimensional space created by the latent components (the T matrix). By typically plotting the first two or three components against each other (e.g., Component 1 vs. Component 2), these plots reveal critical patterns and relationships within your data.

Clusters

Groups of similar observations, indicating underlying categories or shared characteristics within the dataset.

Outliers

Unusual samples that deviate significantly from the main data patterns, potentially indicating errors, unique events, or anomalous behavior.

Trends & Gradients

Systematic patterns or gradients in the data, suggesting underlying processes, time-dependent changes, or concentration shifts.

Group Separation

How well different predefined groups (e.g., control vs. treatment) are distinguished from each other based on their component scores.

Scores plots are invaluable for quality control, identifying problematic samples, and gaining a deeper understanding of the overall data structure.

Interpreting Loadings and Weights

Loadings and weights are fundamental outputs of a PLS model, providing deep insights into how original variables contribute to the latent components and influence the prediction of the response variable(s).

Loadings (P Matrix)

Quantify the contribution of each original predictor variable (X) to the latent components. They reveal the direction and strength of the relationship between each X variable and a specific component.

Weights (W Matrix)

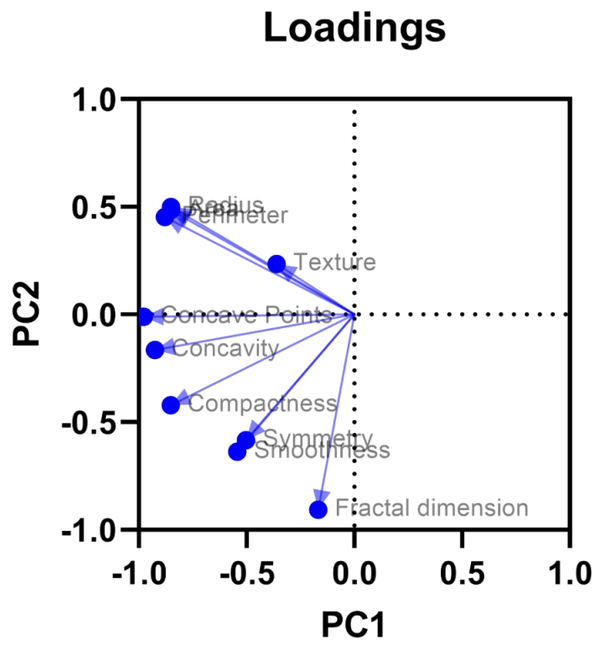
Represent the importance of each predictor variable in constructing the components. These components are specifically designed to maximize the covariance with the response variable(s) (Y).

High Absolute Values

A high absolute value for a variable in either loadings or weights indicates a strong influence on that particular latent component, making it a significant driver for the model.

Loadings Plots & Variable Importance

- **Variable Influence:** Identify which variables are most influential in shaping the latent components and thus the model's predictions.
- **Variable Correlation:** Variables that plot close together are positively correlated, while those on opposite sides of the origin are negatively correlated.
- **Group Separation:** Understand which variables are responsible for the distinct clustering or separation observed in the scores plots.
- **Key Predictors:** Pinpoint the most critical predictors for the response variable by examining their position relative to the response in a loadings plot.
- **VIP Scores:** (Variable Importance in Projection) provide a single metric summarizing the overall importance of each predictor variable across all components in the model. Variables with a VIP score greater than 1 are generally considered significant.



Evaluating Model Quality: R^2 and Q^2

In Partial Least Squares (PLS) modeling, two primary metrics, R^2 and Q^2, are critical for assessing how well your model performs, both in fitting the existing data and predicting new, unseen data.

R^2 (R-squared): Explained Variance

Measures how well the model fits the training data. It indicates the proportion of variance in the dependent variable (Y) that is predictable from the independent variables (X).

- Range: 0 to 1 (higher is better).
- **R^2X:** variance explained in predictor matrix X.
- **R^2Y:** variance explained in response variable Y.
- Caution: A high R^2 alone doesn't guarantee good predictions on new data, as it can be inflated by overfitting.

Q^2 (Q-squared): Predicted Variance

Measures the predictive ability of the model, typically through cross-validation. It indicates how well the model predicts new, unseen data points, reflecting its generalization power.

- Range: Typically 0 to 1 (can be negative for very poor models).
- **Relationship to R^2:** Q^2 should be reasonably close to R^2 for a robust model.
- **Overfitting Indicator:** A large gap between R^2 and Q^2 suggests the model is overfitting the training data.
- **Rule of Thumb:** A Q^2 value greater than 0.5 generally indicates good predictive ability.

Together, R^2 and Q^2 provide a comprehensive view of model quality, balancing model fit with its ability to generalize to future observations, crucial for reliable statistical analysis.

Using Regression Coefficients for Prediction

After building a Partial Least Squares (PLS) model, we obtain regression coefficients that translate the model back into the original variable space, showing how the original X variables relate to the predicted Y variable.

$$Y_{predicted} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

→ Applying Coefficients

Apply these coefficients to new observations (sets of X variables) to generate predictions for the response variable (Y).

→ Interpreting Effects

Each coefficient (\$beta_i$) indicates the unique effect of its corresponding variable (X_i) on the response. A positive coefficient implies a positive relationship, while a negative one suggests an inverse relationship.

→ Magnitude of Effect

The absolute magnitude of a coefficient reflects the strength of that variable's influence. Larger absolute values mean a stronger impact on the predicted Y.

→ Standardization

Coefficients can be standardized or unstandardized. Standardized coefficients allow for direct comparison of variable importance, especially when original variables have different scales.

Regression coefficients are crucial for making predictions, understanding variable effects, and deploying the PLS model for real-world applications.

Understanding Overfitting in PLS Models

Overfitting occurs when a Partial Least Squares (PLS) model learns the training data and its random fluctuations too precisely. This leads to exceptional performance on the data it was trained on, but a dramatic decrease in accuracy when encountering new, unseen data.

1

High R^2, Low Q^2

A large disparity between R^2 (model fit) and Q^2 (predictive ability), such as R^2=0.95 and Q^2=0.40, strongly indicates overfitting.

2

Poor Test Data Performance

The model's predictive accuracy drops significantly when applied to independent test datasets not used during training.

3

Excessive Components

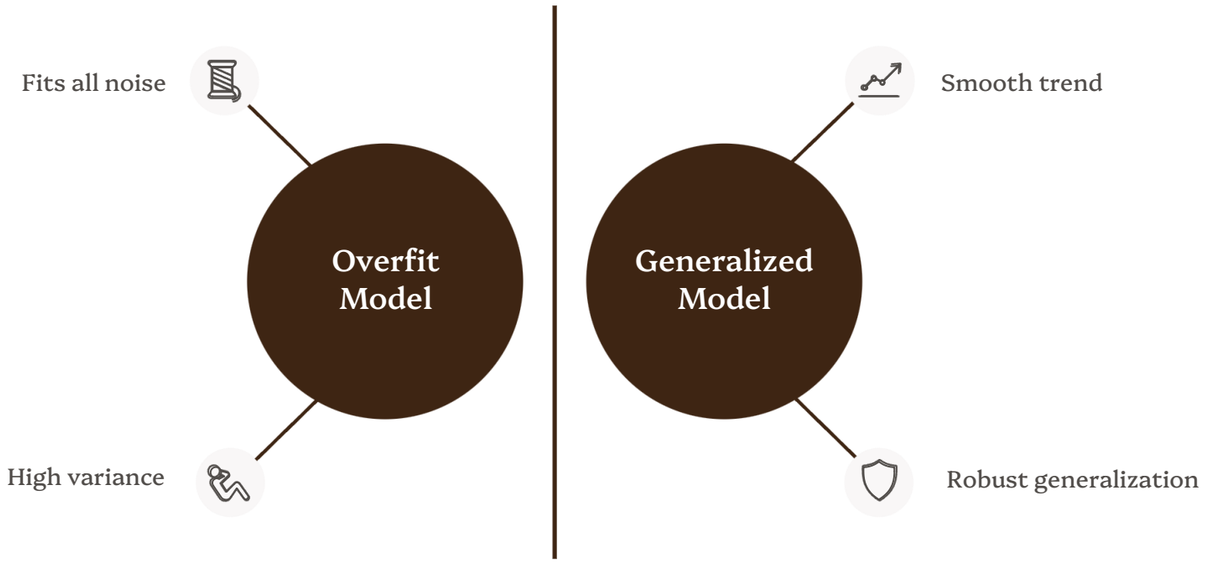
Using too many latent components relative to the sample size, capturing noise rather than underlying patterns.

Causes of Overfitting

- **Too Many Components:** Including more latent variables than necessary to explain the variance.
- **Small Sample Size:** Insufficient data relative to the number of predictor variables, making the model sensitive to noise.
- **Lack of Cross-Validation:** Not properly validating the model's performance on unseen data.

Preventing Overfitting

- **Cross-Validation:** Systematically use cross-validation to select the optimal number of components.
- **Balance R^2 and Q^2:** Aim for Q^2 values that are close to R^2 to ensure generalization.
- **Simpler Models:** Prioritize parsimonious models with fewer components.
- **More Samples:** Increase the sample size if feasible to provide more robust data.
- **Preprocessing:** Apply appropriate data cleaning and scaling to reduce noise.



Cross-Validation: Ensuring Model Reliability

Cross-validation is a critical statistical technique used to evaluate how well a model generalizes to an independent dataset. It provides an objective assessment of the model's performance on new, unseen data, which is vital for building robust predictive models, especially in complex multivariate analysis like PLS.

Why it's Essential in PLS

- **Prevents Overfitting:** Ensures the model doesn't simply memorize training data noise, leading to poor performance on new data.
- **Optimal Component Selection:** Helps identify the ideal number of latent components, balancing model complexity and predictive power.
- **Realistic Performance Estimate:** Provides an unbiased measure of how well the model is expected to perform on future, unseen observations.

Common Cross-Validation Methods

- **K-fold Cross-Validation:** Data is divided into K subsets (folds). The model is trained on K-1 folds and tested on the remaining fold, repeating this process K times.
- **Leave-One-Out (LOO):** A specific case of K-fold where K equals the number of samples, making each sample a test set once.
- **Monte Carlo Cross-Validation:** Involves repeatedly and randomly splitting the data into training and validation sets to average out performance estimates.

Cross-Validation Process in PLS

How Cross-Validation Works in PLS:

01

Data Splitting

The dataset is initially split into distinct training and validation (or test) sets for each fold or iteration.

02

Model Building & Prediction

A PLS model is built on the training data using a varying number of components. Predictions are then made on the corresponding validation set.

03

Error Calculation

Prediction errors, such as RMSECV (Root Mean Square Error of Cross-Validation), are calculated for each model run on the validation set.

04

Iteration & Averaging

Steps 1-3 are repeated across all folds. The prediction errors are then averaged across all iterations.

05

Optimal Component Selection

The number of latent components that yields the lowest average prediction error is selected as optimal for the final model.

Key Metrics from Cross-Validation:

RMSECV (Root Mean Square Error of Cross-Validation)

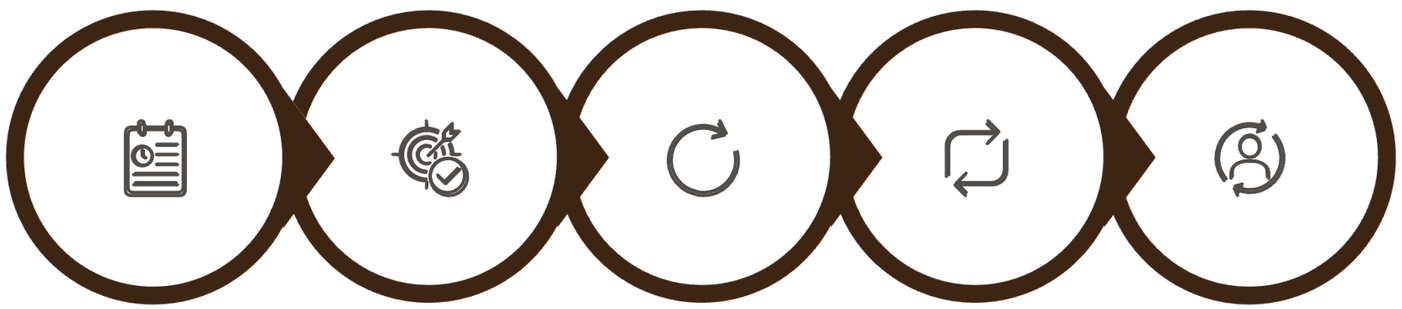
- A measure of the average magnitude of the errors. Lower values indicate better predictive accuracy.

Q^2 (Cumulative Predicted Variance)

- Indicates the predictive power of the model on new data. A higher Q^2 (e.g., >0.5) suggests good generalization.

Prediction Error Plots

- Visualizations that help identify the optimal number of components by showing the trend of prediction error as more components are added.



Practical Tips for Interpretation

Cross-Validation is Essential

Always use cross-validation to select the optimal number of components—don't rely on fit statistics alone

Visualise Your Scores

Examine score plots carefully for clusters, trends, and outliers that reveal data structure

Understand Variable Importance

Interpret loadings and weights to identify which original predictors are driving your responses

Watch for Overfitting

If predicted R^2 (Q^2) is substantially lower than R^2, your model may be overfitting—consider fewer components

Summary & Takeaways

Power of PLS

PLS regression excels with high-dimensional, collinear data where classical methods fail

Balanced Approach

It elegantly balances dimensionality reduction with prediction accuracy through latent components

Flexible Variants

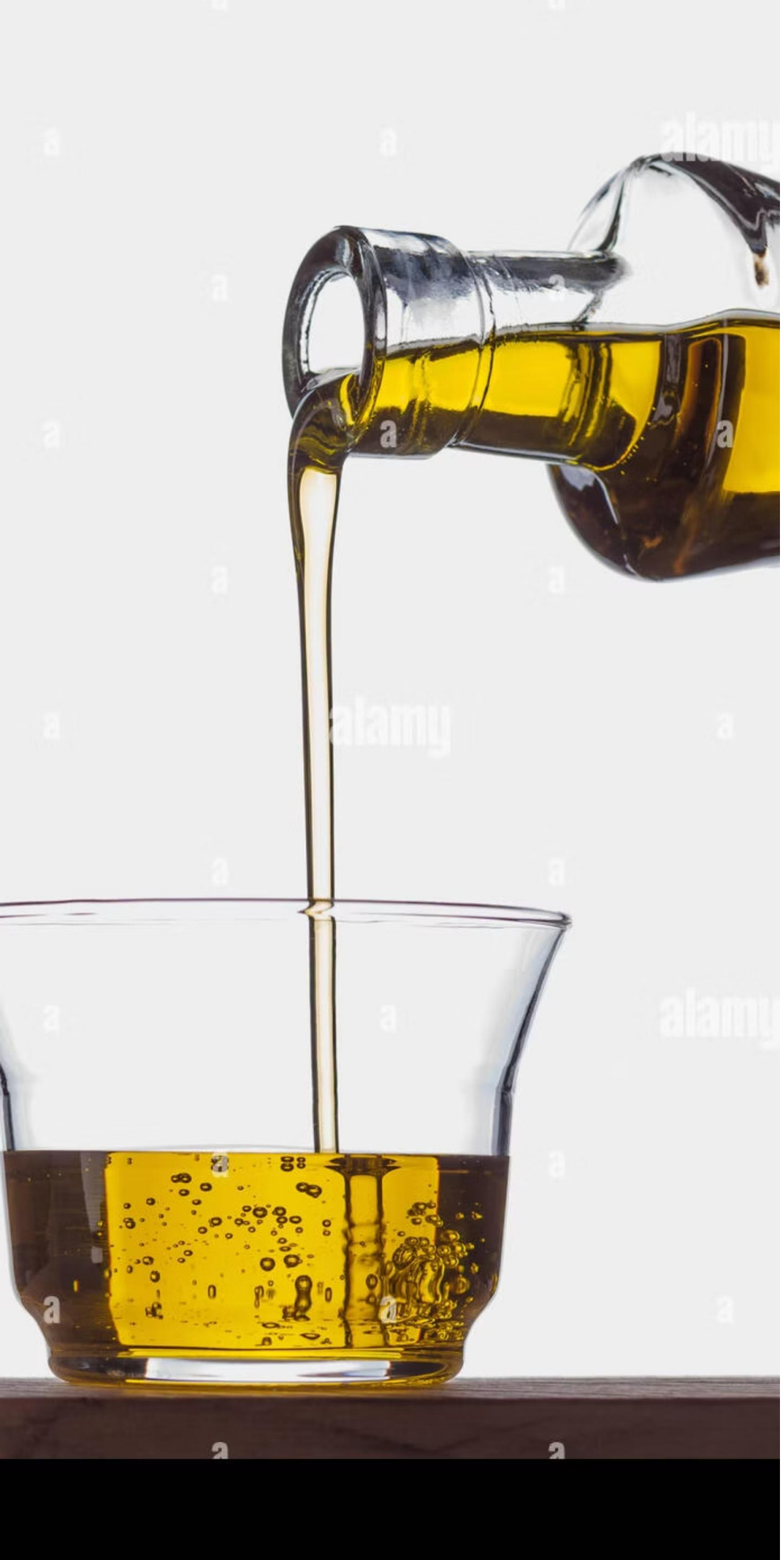
PLS1 and PLS2 adapt seamlessly to single or multiple response variable scenarios

Critical Practices

Proper preprocessing and rigorous validation are essential for building reliable, trustworthy models

Actionable Insights

Rich interpretation tools help translate PLS mathematical results into practical, actionable insights



FTIR Spectroscopy for Virgin Olive Oil Quality Analysis

A rapid analytical method combining Fourier Transform Infrared (FTIR) spectroscopy with Partial Least Squares (PLS) regression for monitoring fatty acid composition and peroxide value in virgin olive oil. This approach offers a faster, more cost-effective alternative to traditional chromatographic methods.

METHODOLOGY

Analytical Workflow

Sample Collection

86 virgin olive oil samples from Italian regions (Abruzzo, Marche, Puglia) across 2006-2007 harvest seasons

Reference Analysis

GC-FID for fatty acid methyl esters and titrimetric method for peroxide value determination

FTIR Spectroscopy

ATR-FTIR spectra acquired (4000-700 cm^{-1}, 32 scans/sample, 4 cm^{-1} resolution) using ZnSe crystal

Chemometric Analysis

PLS regression models built with spectral pre-treatment and validated using independent sample sets

STATISTICAL APPROACH

PLS Regression Strategy and Optimization

Why Partial Least Squares?

PLS regression was selected as the multivariate calibration method because it effectively handles complex spectral data with multiple overlapping peaks. Unlike univariate methods, PLS decomposes spectral data into latent variables (LVs) that capture maximum covariance between spectra and analyte concentrations.

The method excels with collinear data and can extract useful information even when spectral features are not easily detectable by visual inspection.

Model Optimization

Spectral Pre-treatment: Mean-centering applied to all models. First derivative used for peroxide value to enhance sensitivity.

Latent Variables: Optimal number determined using Haaland and Thomas criterion (α=0.75), ranging from 13-15 LVs for fatty acids and 10 LVs for peroxide value.

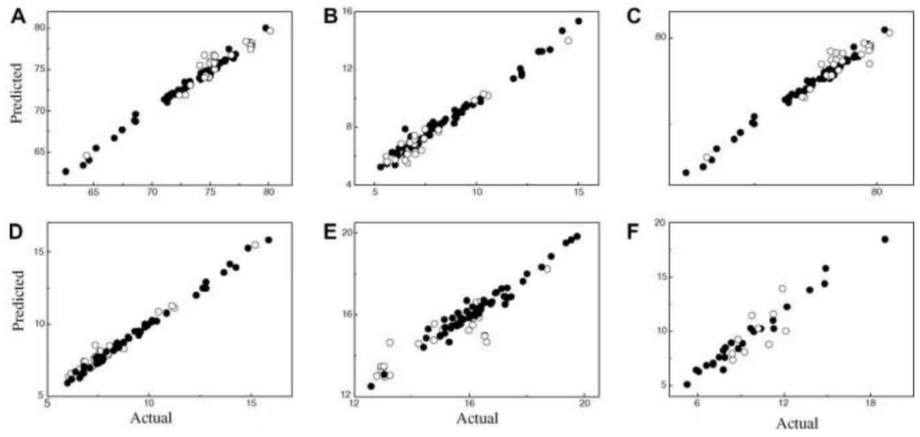
Spectral Ranges: 3033-700 cm^{-1} for fatty acids (excluding 2400-2260 cm^{-1} noise region); full spectrum 4000-700 cm^{-1} for peroxide value.

RESULTS

Fatty Acid Profile: Calibration Performance

PLS models demonstrated excellent predictive capability for major fatty acid components. The wide concentration ranges in the sample set (oleic acid 62.0-80.0%, linoleic acid 5.3-15.0%) enabled robust calibration models suitable for diverse olive oil samples.

0.99	0.98	0.99	0.99
Oleic Acid r^2	**Linoleic Acid r^2**	**MUFA r^2**	**PUFA r^2**
RMSD: 0.42%, REC: 0.51%, 14 latent variables	RMSD: 0.39%, REC: 4.64%, 13 latent variables	RMSD: 0.42%, REC: 0.56%, 14 latent variables	RMSD: 0.20%, REC: 2.23%, 15 latent variables



VALIDATION

External Validation and Method Performance

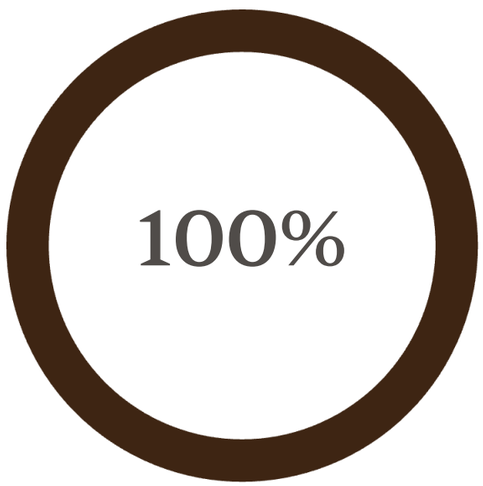
Cross-Validation Strategy

Independent validation sets were used to assess model performance: 25 samples for fatty acid parameters and 10 samples for peroxide value. This external validation approach provides unbiased assessment of predictive capability.

Recovery Rates

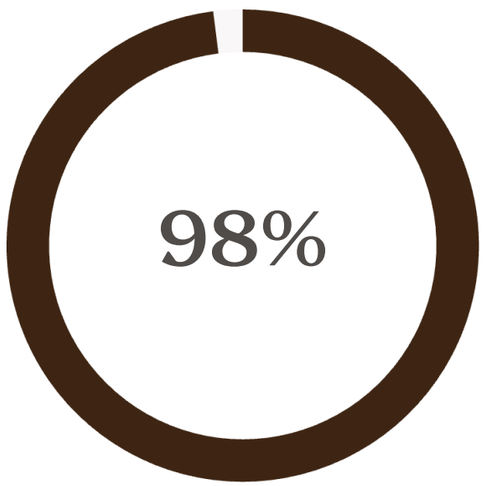
- **Oleic acid:** 100% recovery, REP 1%
- **Linoleic acid:** 98% recovery, REP 7%
- **MUFA:** 100% recovery, REP 1%
- **PUFA:** 103% recovery, REP 4%
- **SFA:** 98% recovery, REP 6%
- **Peroxide value:** 100% recovery, REP 10%

Regression slopes (0.93-0.98) and intercepts near zero indicate low bias and absence of systematic errors.



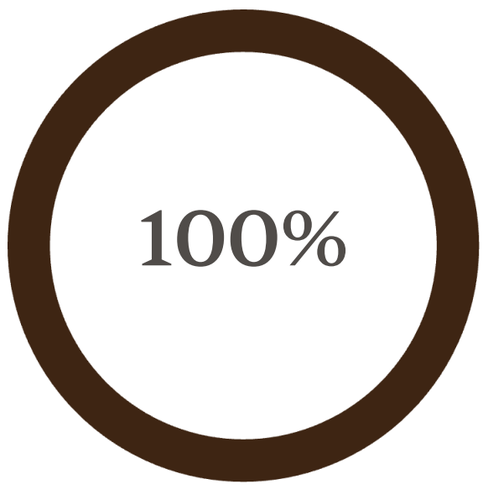
Oleic Acid

Recovery rate



Linoleic Acid

Recovery rate



MUFA

Recovery rate

Key Findings and Method Advantages

Superior Speed

Complete analysis in minutes vs. 30 minutes for titrimetric PV and 1 hour for GC-FID fatty acid analysis

No Sample Preparation

Direct ATR-FTIR measurement eliminates derivatization and extraction steps required by traditional methods

Environmental Benefits

Virtually no solvent waste produced, making it more environmentally friendly than chromatographic techniques

Excellent Detection Limits

LODs: 3.0% (oleic), 0.5% (linoleic), 1.3% (SFA), 3.0% (MUFA), 0.3% (PUFA), 1.0 meq O_2/kg (PV)

The FTIR-PLS method provides results statistically comparable to official procedures while offering significant advantages in throughput, cost, and environmental impact. First derivative spectral treatment proved essential for peroxide value determination, achieving expanded measurable range (3.4-15.7 meq O_2/kg) compared to previous NIR methods (0-10 meq O_2/kg) without requiring reagent addition.