Classification with Multivariate Analysis: Focus on PLS-DA

Exploring advanced statistical methods for authenticating food products and ensuring quality control through sophisticated data analysis techniques.

Why Multivariate Analysis?

Complex Food Data

Multiple variables measured simultaneously including chemical composition, sensory attributes, and spectral fingerprints create rich, multidimensional datasets.

Capturing Interactions

Traditional univariate methods fail to reveal interactions between variables and overlook crucial patterns hidden within the data structure.

Enhanced Accuracy

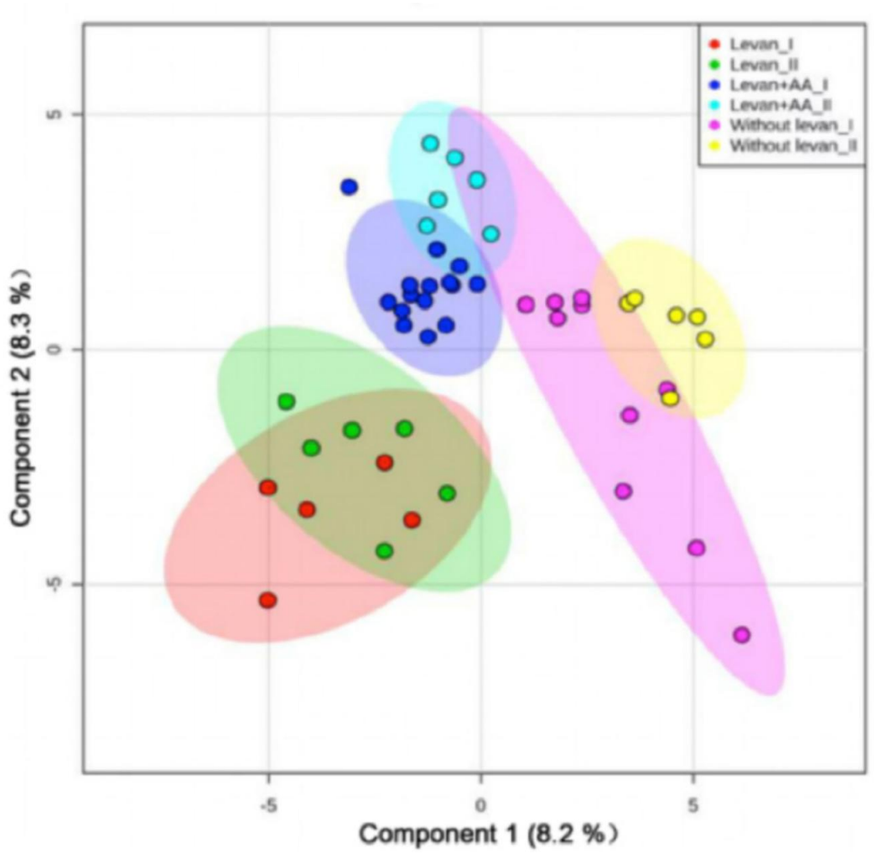
Multivariate methods reveal hidden structures, improving classification accuracy and providing deeper insights into food quality and authenticity.

Key Concepts: Classification vs Clustering

Classification

Supervised learning with predefined, known classes.

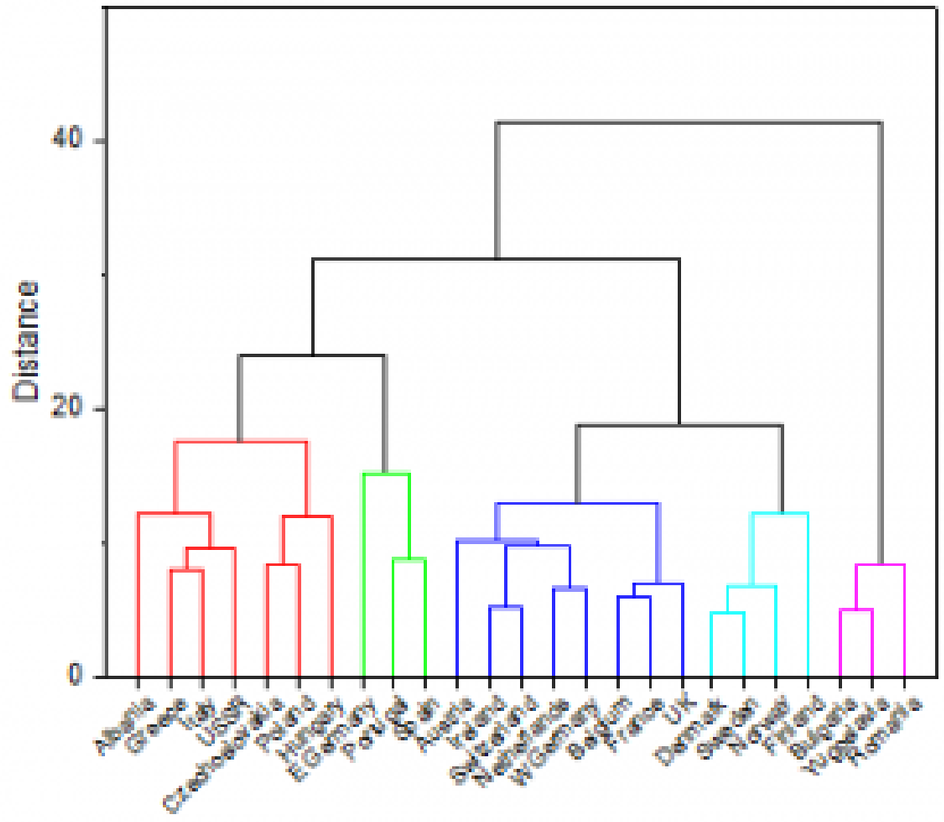
- Requires labelled training data
- Predicts class membership for new samples
- Examples: food origin, cultivar identification, quality grading



Clustering

Unsupervised grouping without prior labels.

- Discovers natural groupings in data
- Exploratory pattern recognition
- No predefined categories needed



PLS-DA is a supervised classification method specifically designed for complex, high-dimensional food datasets where variable interactions matter.

Supervised Analysis: The Key to Food Classification

Supervised analysis is fundamental in PLS-DA to ensure the authenticity and quality of food products, distinguishing itself clearly from unsupervised methods.

Guided Learning

Algorithms learn from a "labeled" dataset, where desired outcomes (e.g., origin, variety) are already known, providing a solid foundation for the model.

Essential Training Data

Requires samples with predefined and known categories to build a robust predictive model, such as different types of cheese or olive oils.

Prediction of New Samples

Once trained, the model can accurately classify new unknown samples, assigning them to previously learned categories.

Benefits for Food Safety

Ideal for authentication, fraud identification, or quality assessment, offering clear and interpretable answers for the food industry.

Achieving Visual Separation in Multivariate Data Analysis

Raw analytical data from complex samples often presents as an intricate mesh of overlapping data points, making direct interpretation and classification challenging. Multivariate classification methods like PLS-DA are powerful tools designed to untangle this complexity by transforming the data space.

Untangling Raw Data Complexity

Initially, raw data from various samples (e.g., different food origins or varieties) typically appears as overlapping clusters in high-dimensional space. Without transformation, distinguishing distinct groups is nearly impossible due to inherent variability and noise.

PLS-DA Transformation Process

Partial Least Squares-Discriminant Analysis (PLS-DA) employs a supervised statistical technique. It constructs new latent variables (components) that maximize the covariance between the predictors (e.g., spectral data) and the response variable (e.g., sample class labels). This process systematically identifies patterns that best differentiate between defined classes.

Interpreting Latent Variables

In the transformed score plot, the new axes represent these latent variables or principal components. These components capture the most significant variance related to class separation, allowing the data to be visualized in a lower-dimensional space where class distinctions become evident.

Measuring Class Separation

Class separation is achieved when samples belonging to different categories form distinct, non-overlapping clusters in the score plot. The quality of separation can be quantified using metrics like R2Y (explained variance of Y) and Q2 (predictive ability), along with visual inspection of confidence ellipses or separation boundaries.

Importance of Visual Inspection

Visual inspection of the score plot is crucial for model validation. It provides an intuitive understanding of how well classes are separated and can reveal outliers or misclassified samples that might not be obvious from statistical metrics alone. This visual check helps confirm the model's reliability and interpretability.

Distinguishing Good from Poor Separation

Good separation is characterized by tight, well-defined clusters for each class with clear boundaries and minimal overlap, indicating a robust classification model. Poor separation, conversely, shows overlapping or dispersed clusters, suggesting insufficient distinction between classes or a need for model refinement.

What is PLS-DA?

01

Supervised Method

Partial Least Squares Discriminant Analysis integrates dimensionality reduction with classification capabilities.

02

Class-Focused Separation

Incorporates predefined class labels to maximise separation between groups, ensuring discrimination is the primary objective.

03

Latent Variable Extraction

Identifies and extracts latent variables that best capture class differences in high-dimensional food analysis data.

04

Optimal for Complexity

Particularly effective when the number of variables exceeds sample size, a common scenario in modern food analysis.

Mathematical Foundations of PLS-DA

Partial Least Squares-Discriminant Analysis (PLS-DA) provides a robust framework for classifying complex food science data by identifying latent variables that maximize the covariance between predictor and response matrices. It transforms high-dimensional data into a lower-dimensional space, where class separation is optimized for clearer distinction.

1. Matrix Decomposition

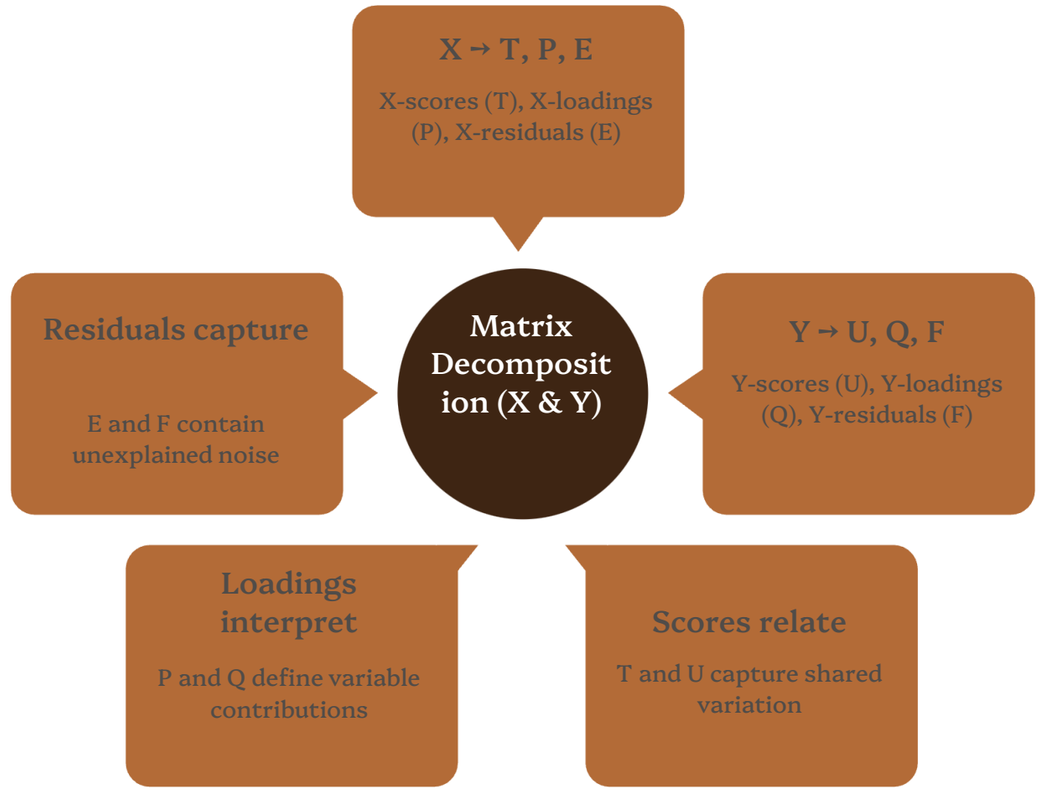
PLS-DA decomposes both the predictor matrix (X, e.g., spectral data) and the response matrix (Y, class labels) into scores and loadings, along with residual matrices.

$$X = TP^T + E$$

$$Y = UQ^T + F$$

Where:

- **X:** Predictor variables matrix (n samples x p variables)
- **Y:** Response variables matrix (n samples x m class labels)
- **T:** X-scores matrix (latent variables from X)
- **U:** Y-scores matrix (latent variables from Y)
- **P:** X-loadings matrix (weights for X variables)
- **Q:** Y-loadings matrix (weights for Y variables)
- **E, F:** Residual matrices (unexplained variance)



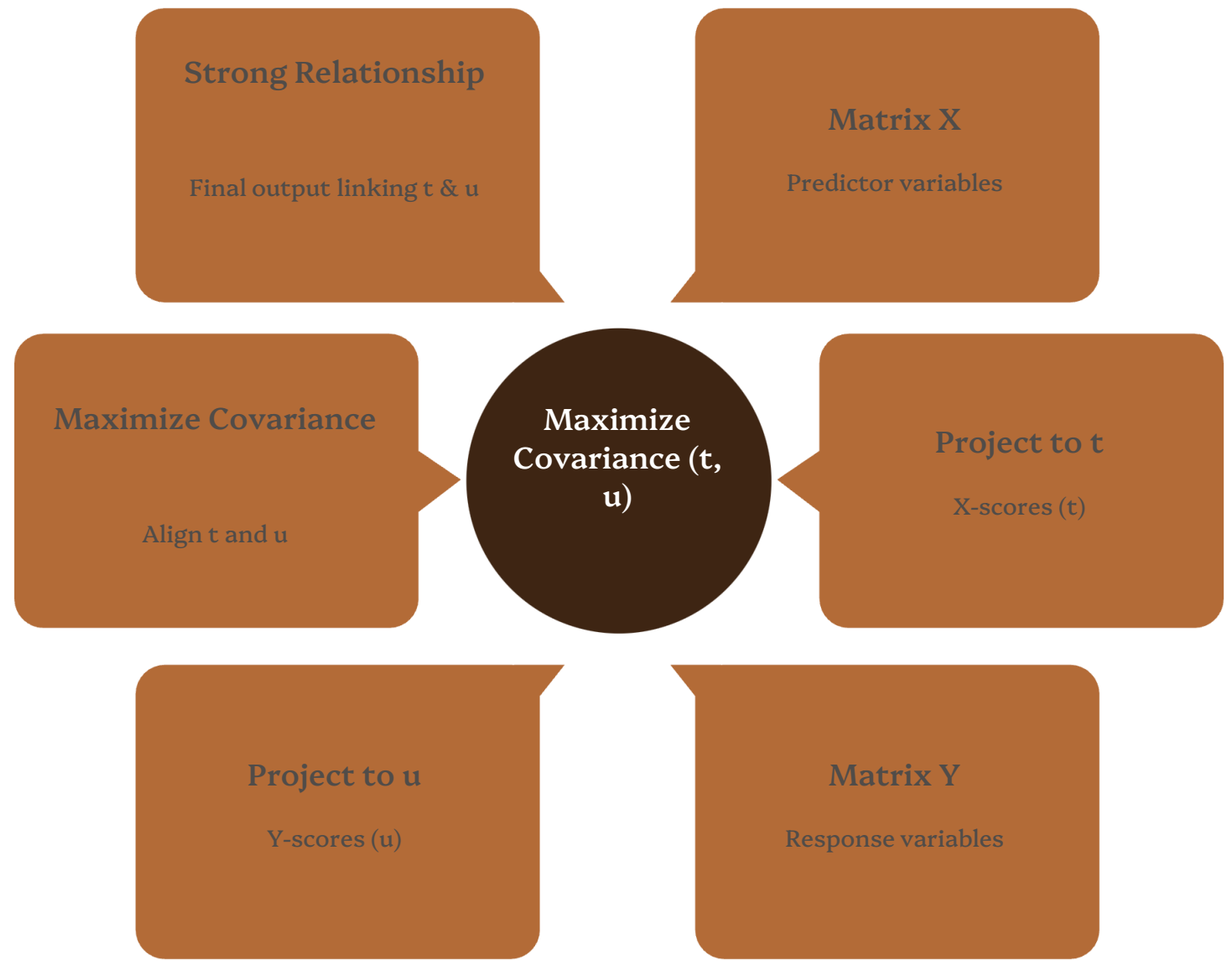
2. Maximizing Covariance: Unlocking Relationships

At the heart of PLS-DA is the principle of maximizing the covariance between the projected scores of the predictor variables (X) and the response variables (Y). This ensures that the latent variables extracted best explain the relationship between the two datasets.

The algorithm identifies optimal projection vectors, **w** for X and **c** for Y, to create new scores **t** and **u** respectively, such that their covariance is maximized.

$$maxCov(t, u) = \quad max(t^T u)$$

Here, **t** and **u** are column vectors representing the X-scores and Y-scores (latent variables) derived from the original matrices T and U. This maximization is achieved through an iterative process, sequentially extracting components that capture the maximum remaining covariance between X and Y.



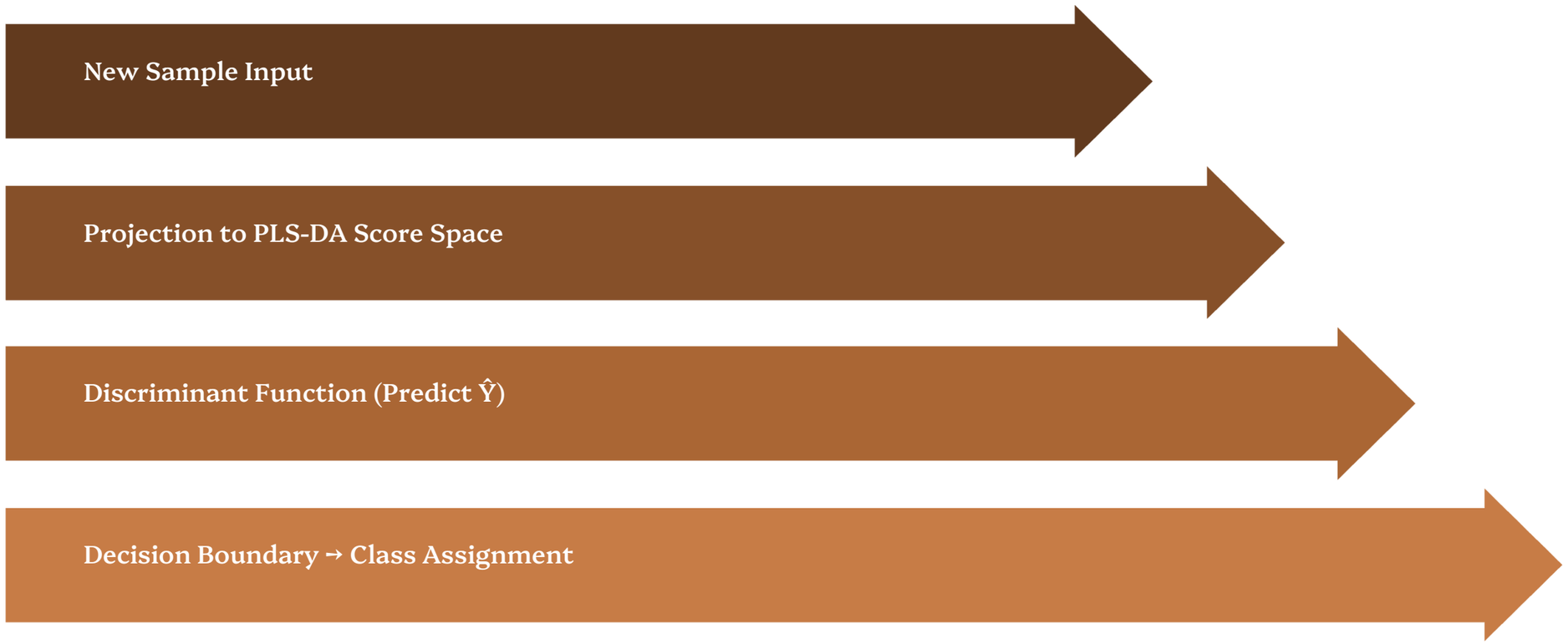
Mathematical Foundations of PLS-DA: Discriminant Function and Classification

After extracting the latent variables that maximize covariance, PLS-DA builds a linear regression model between the X-scores (T) and Y-scores (U). This model defines a discriminant function used to classify samples.

For classification, the response variable Y is typically encoded using dummy variables (e.g., 0 for one class, 1 for another, or one-hot encoding for multiple classes). The prediction equation for $\hat{Y}$ (predicted Y-scores) is derived from this relationship:

$$\hat{Y} = T(T^T T)^{-1} T^T Y$$

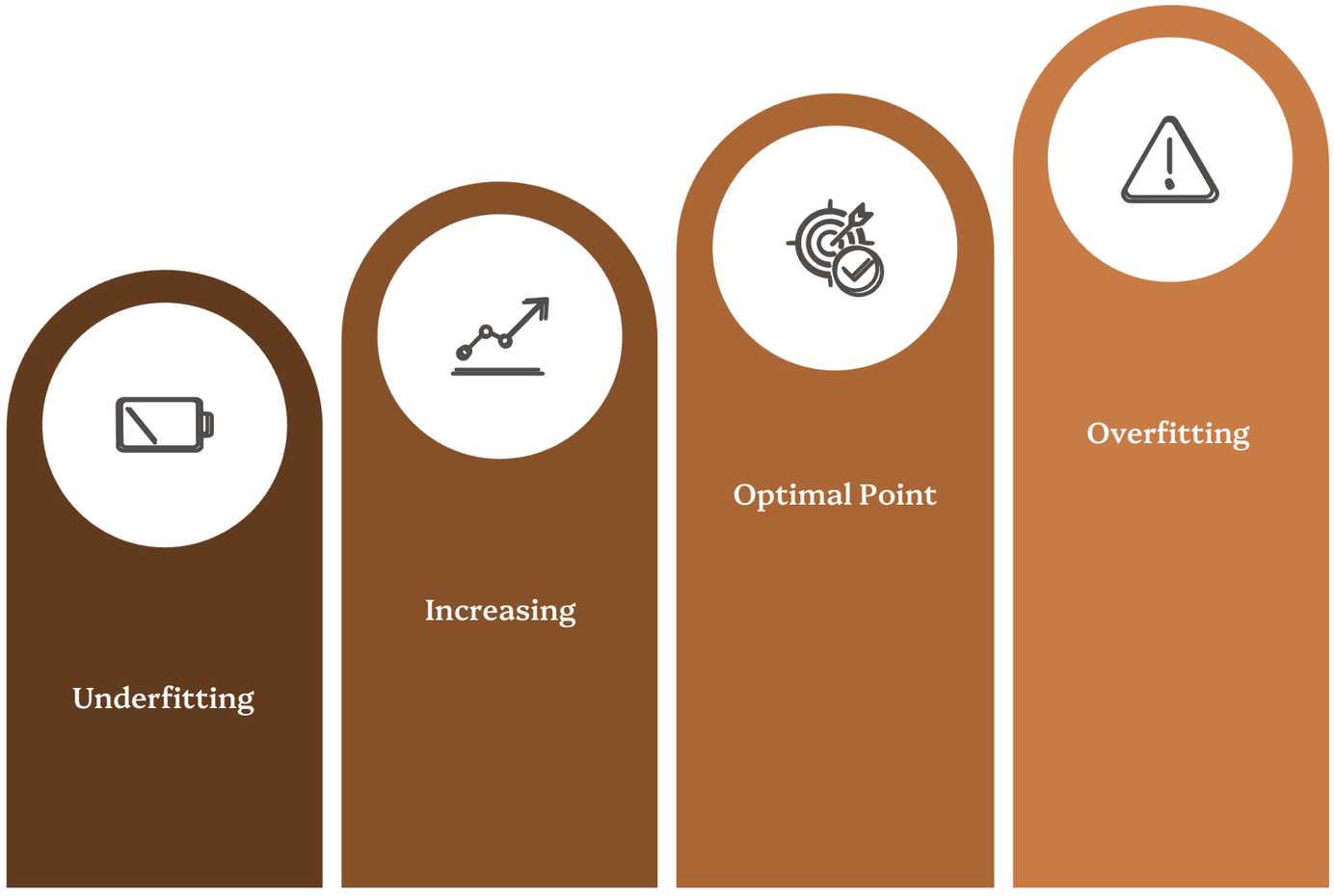
New samples are then classified by projecting their data onto the PLS-DA model to obtain their X-scores. These scores are fed into the discriminant function to predict their Y-values. Finally, each sample is assigned to the class corresponding to its predicted Y-value (e.g., closest to 0 or 1, or highest probability in multi-class scenarios).



Key Parameters: Selecting Latent Components

The number of latent components (factors) is a critical parameter in PLS-DA, significantly influencing model accuracy and generalization. Selecting too few components results in an underfit model, unable to capture essential data patterns. Conversely, too many components can lead to overfitting, where the model learns noise in the training data, performing poorly on new, unseen samples.

This optimal number is typically determined through cross-validation. This technique evaluates the model's predictive performance on independent subsets of the data, helping to identify the point where predictive ability is maximized without compromising generalizability. The aim is to strike a balance between model complexity and robust predictive performance, ensuring the model is both informative and reliable.



PLS vs. PLS-DA: Distinguishing Applications

While both Partial Least Squares (PLS) methods leverage latent variables to handle complex data, their application differs fundamentally based on the nature of the outcome variable.

PLS Regression: Continuous Prediction

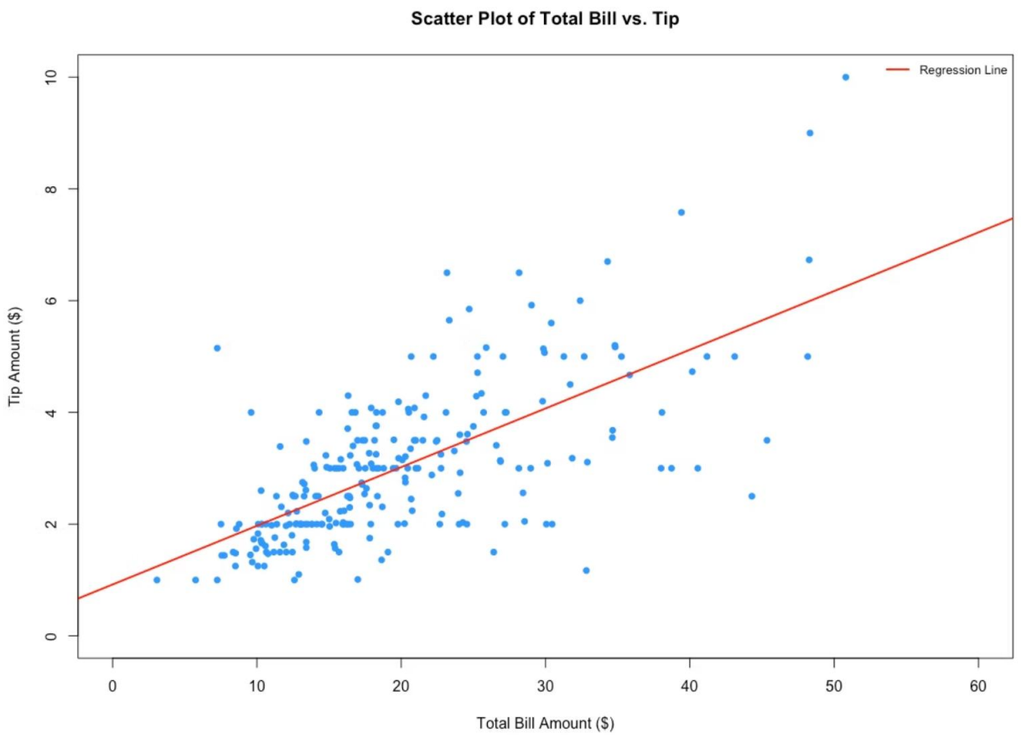
This technique is employed when the outcome variable (Y) is **continuous and numerical**, such as predicting sugar content, acidity levels, or a quantitative quality score. It focuses on modelling the linear relationship between predictor variables and one or more continuous responses.

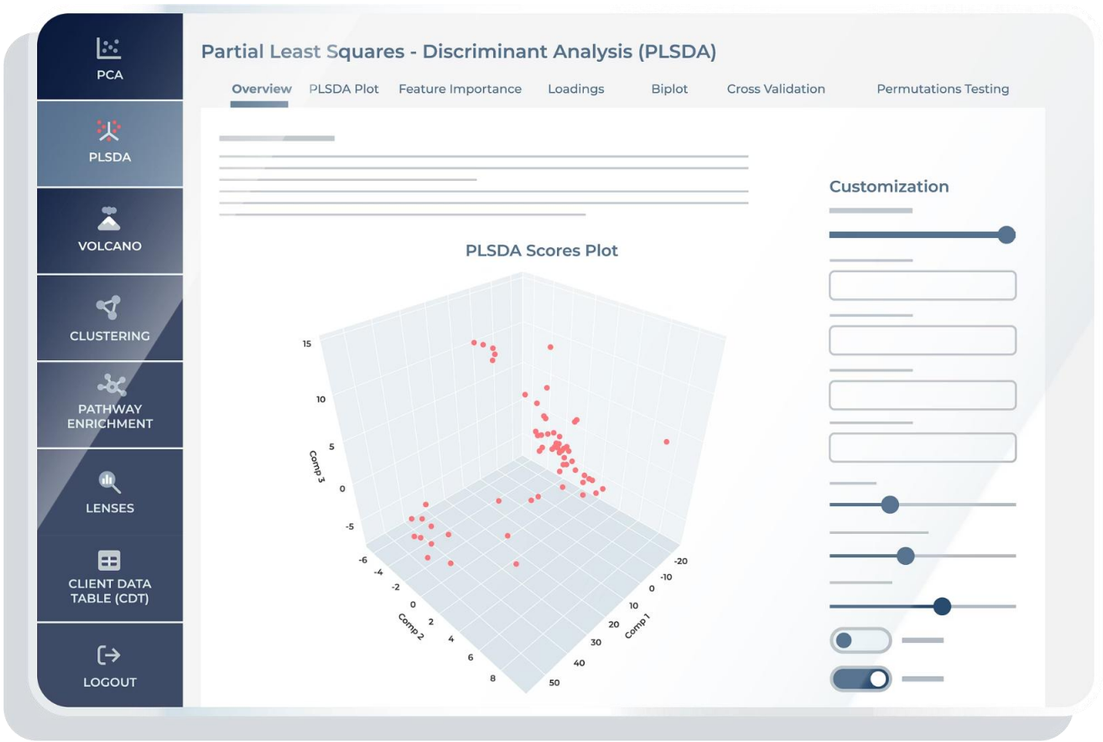
PLS-DA: Categorical Assignment

PLS-DA (Discriminant Analysis) is utilised when the outcome variable (Y) is **categorical**, aiming to classify samples into predefined groups, e.g., 'organic vs. conventional', 'origin A vs. origin B', or 'fresh vs. spoiled'. It transforms the categorical response into a numerical format to maximise separation between these distinct classes.

Visualising PLS vs. PLS-DA in Action

To better understand the distinct applications, let's look at how PLS Regression and PLS-DA visually represent their respective outcomes: continuous predictions versus categorical classifications.





PLS-DA vs PCA: Why PLS-DA for Classification?

PCA: Exploratory Analysis

Principal Component Analysis maximises variance in the dataset **without considering class labels**. Excellent for exploration but not optimised for classification tasks.

PLS-DA: Targeted Classification

PLS-DA maximises **covariance between predictors and class membership**, directly focusing on discrimination. Purpose-built for authentication where class differences may be subtle.

In food authentication studies where subtle compositional differences determine origin or quality, PLS-DA's supervised approach provides superior discriminatory power.

Handling Paired and Complex Designs: Multilevel PLS-DA

The Challenge

Nutritional intervention studies frequently employ paired data from cross-over designs, where individuals serve as their own controls.

1

The Solution

Multilevel PLS-DA separates treatment effects from individual variation, accounting for within-subject correlation structures.

2

The Benefit

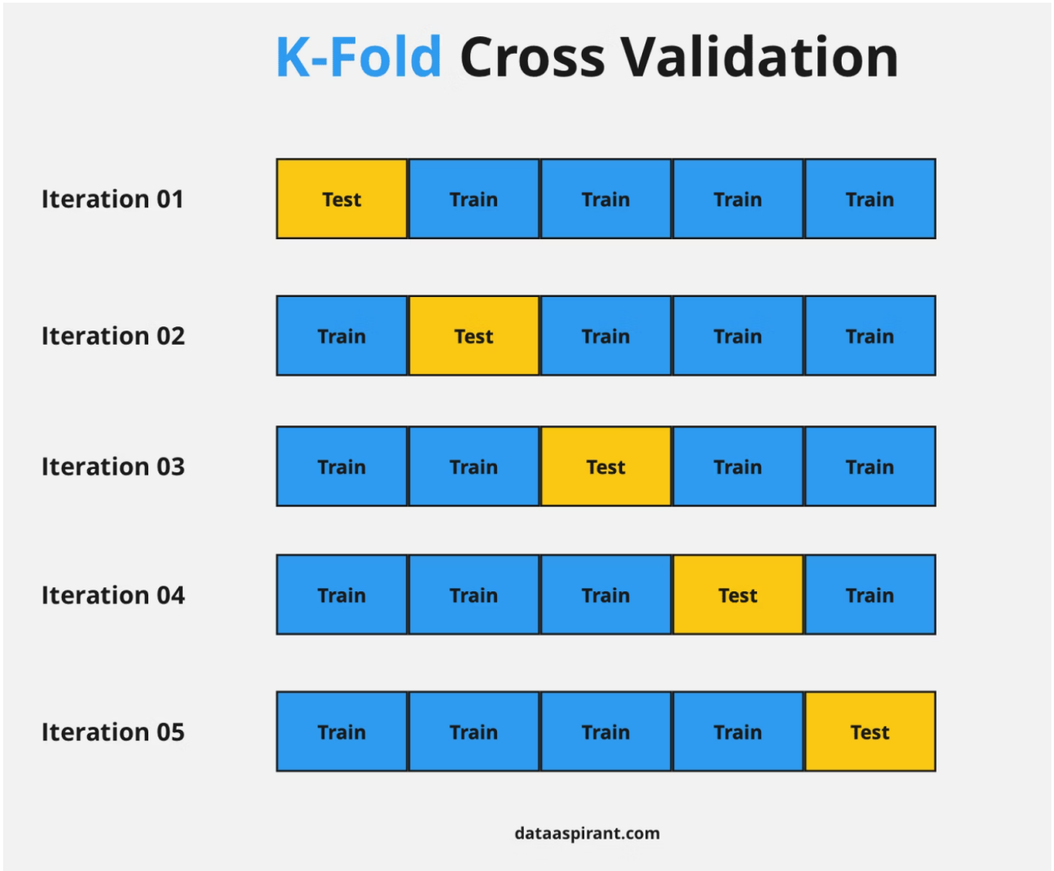
Dramatically improves statistical power and interpretability in food metabolomics and intervention studies.

3

Model Validation and Avoiding Overfitting

Validation Strategies

- **Cross-validation:** K-fold and leave-one-out approaches assess model stability
- **Permutation tests:** Confirm that classification is not due to chance
- **Independent test sets:** Evaluate true predictive performance



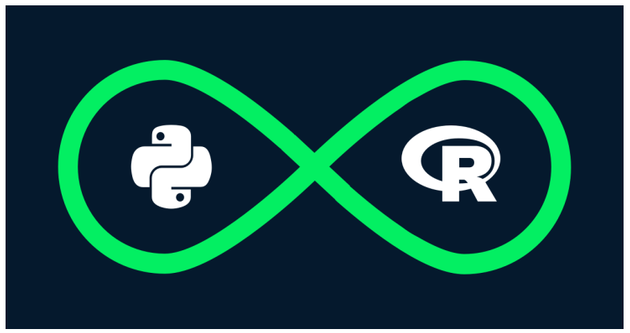
Performance Metrics

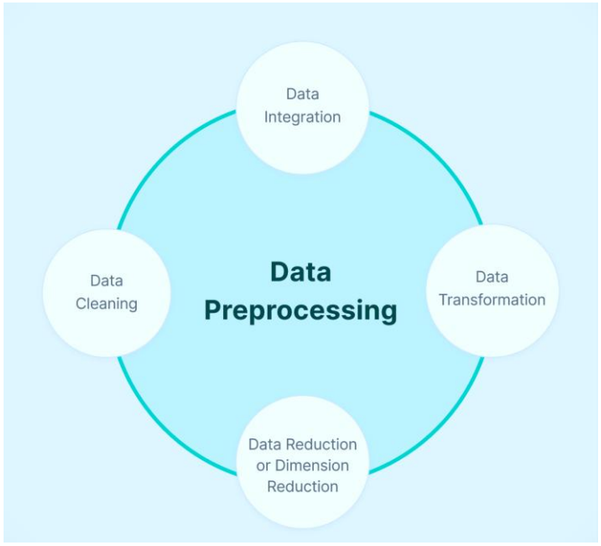
- **Classification accuracy:** Overall correctness
- **Sensitivity & specificity:** Class-specific performance
- **VIP scores:** Variable Importance in Projection identifies key discriminating variables

Rigorous validation ensures model generalisability to new, unseen food samples from production environments.

Software for PLS-DA in Food Analysis

SIMCA®





Commercial Solutions

SIMCA offers comprehensive multivariate analysis. **MetaboAnalyst** provides web-based metabolomics tools. **MATLAB PLS-DA Tool** includes free GUI.

Open-Source Options

R packages like mixOmics provide flexible, scriptable analysis. **Python libraries** offer integration with machine learning workflows.

Data Pre-processing

Critical steps include **scaling** (standardisation), **normalisation**, and handling missing data appropriately before model building.

Best Practices for Food Classification Studies

Experimental Design Excellence

Implement representative sampling strategies across production batches, seasons, and geographical regions. Ensure balanced class sizes to prevent bias.

Data Fusion Approaches

Combine multiple analytical platforms (mass spectrometry, spectroscopy, chromatography) to capture complementary information and enhance discrimination.

Contextual Interpretation

Always interpret statistical results within the context of food chemistry, production processes, and biological variation for meaningful insights.

Conclusion: PLS-DA Empowers Food Authentication and Quality Control

Robust Method

Multivariate classification tailored for complex, high-dimensional food data

Practical Integration

Ready for routine food analysis workflows

Proven Success

Authenticating origin, cultivar, and detecting sophisticated adulteration

Ongoing Innovation

Advances in sparse methods, data fusion, and accessible software tools

---

> "Integration of PLS-DA in routine analytical workflows supports safer, more transparent food supply chains, protecting consumers and legitimate producers alike."

Fingerprinting Alkaloids for Traceability in Lupins

A semi-untargeted UHPLC-MS/MS approach for comprehensive alkaloid profiling and geographical classification of Lupinus albus L. samples from four Italian regions.

FOOD CHEMISTRY | TRACEABILITY

Methodology: MRM-IDA-EPI Acquisition

Sample Preparation

- Raw L. albus seeds from Abruzzo, Lazio, Campania, and Puglia
- Ground and homogenized samples
- MeOH:H2O extraction followed by SPE clean-up
- 100 samples per region analyzed

Analytical Approach

- UHPLC-QqQ-LIT-MS/MS system
- MRM survey scan with IDA criteria
- Enhanced Product Ion (EPI) experiments
- CFM-ID for in silico MS/MS prediction

The method combined targeted analysis of 6 quinolizidine alkaloids with semi-untargeted identification of 21 additional alkaloids, enabling comprehensive alkaloid fingerprinting without requiring all reference standards.

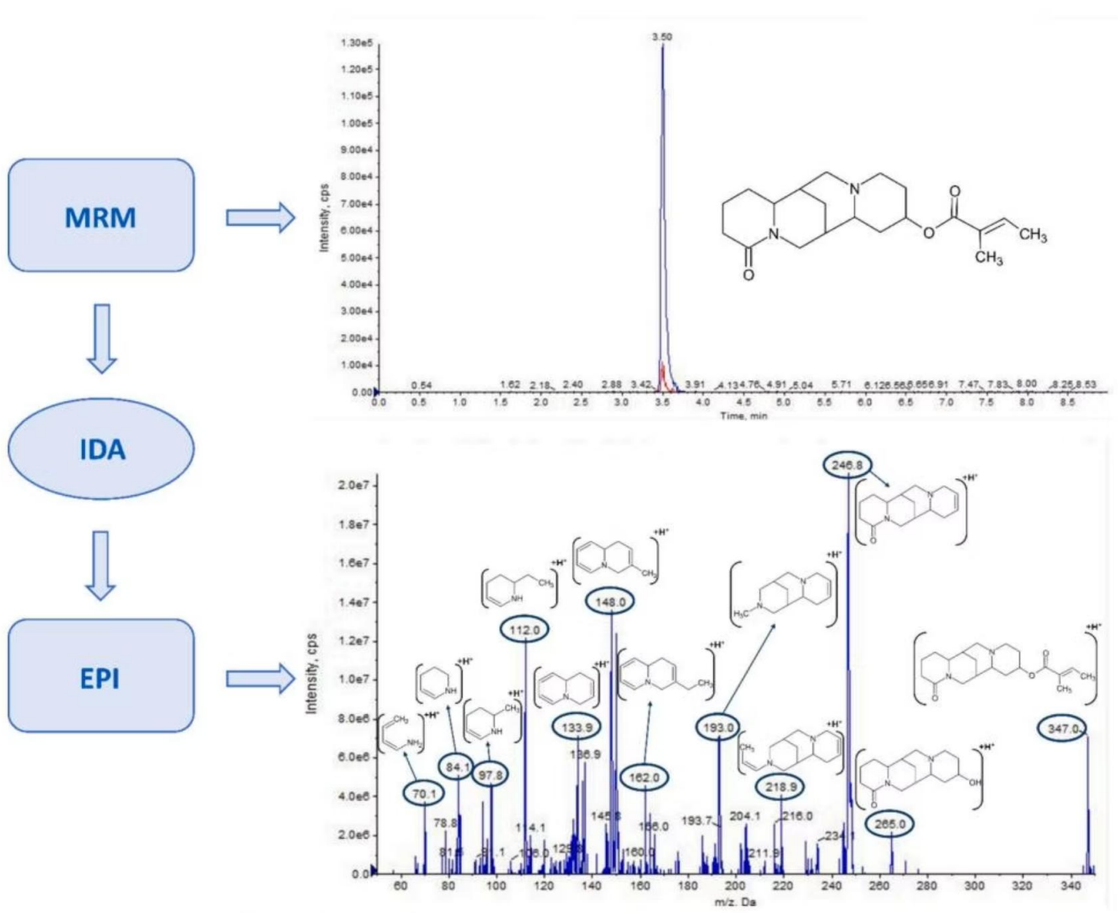


Fig. 1. Angeloyloxylupanine MS/MS spectrum putatively identified in MRM-IDA EPI-mode.

Multivariate Analysis Strategy

01

Data Preparation

Dataset of 400 observations (100 per region) with 27 alkaloid variables. Percentage conversion applied to ensure equal contribution of each variable.

02

Unsupervised HCA

Hierarchical Cluster Analysis combined with heatmap to explore data structure and natural groupings without predefined classes.

03

Supervised PLS-DA

Partial Least Squares Discriminant Analysis with 10-fold cross-validation to build predictive classification model for geographical origin.

Python libraries including pandas, numpy, sklearn, and scipy were used for all statistical processing. Features were standardized by removing mean and scaling variance to unity.

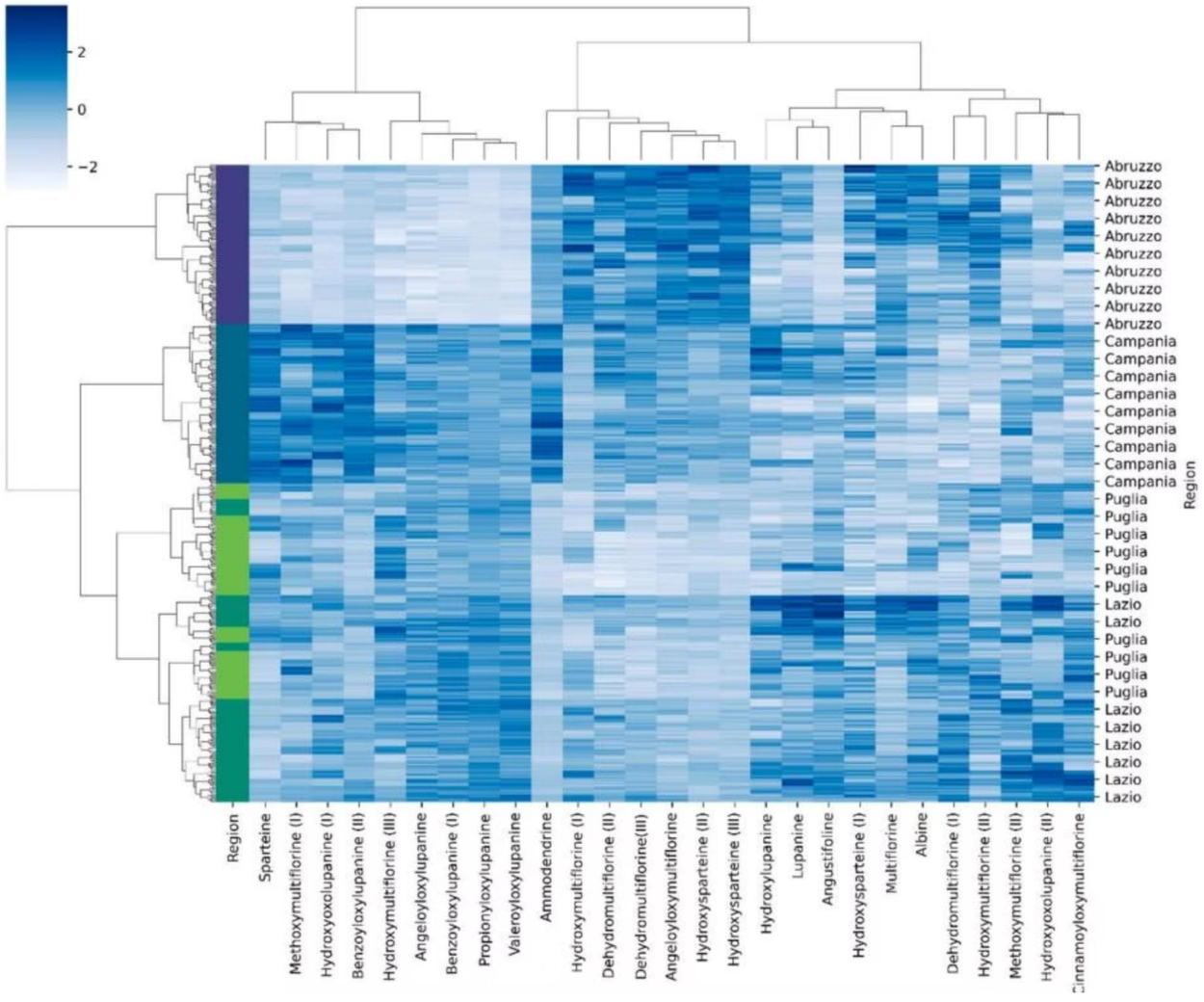
Hierarchical Clustering Results

The heatmap revealed distinct clustering patterns across the four Italian regions. Abruzzo and Campania samples showed perfect regional grouping, indicating homogeneous alkaloid profiles. Lazio and Puglia exhibited mixing in clusters, suggesting less region-specific profiles.

Three main alkaloid clusters emerged, each contributing differently to regional differentiation and demonstrating the complex chemical variability influenced by geographical and environmental factors.

Key Findings

- Perfect clustering for Abruzzo and Campania
- Mixed patterns for Lazio and Puglia
- Three distinct alkaloid clusters identified
- Environmental influence confirmed



PLS-DA Classification Performance

58%

Targeted Approach

Average sensitivity using only 6 standard alkaloids - insufficient for accurate classification

98%

Semi-Untargeted

Average accuracy using all 27 alkaloids - dramatic improvement in classification

100%

Abruzzo Samples

Perfect classification achieved with semi-untargeted approach

The semi-untargeted method using all 27 alkaloid features resulted in significantly improved PLS-DA performance. Key discriminant alkaloids included dehydroxymultiflorine (III), hydroxysparteine (III), ammodendrine, angeloyloxymultiflorine, and benzoxyloxylupanine (II). Notably, no single targeted feature showed distinct contribution - successful classification resulted from the synergistic effect of alkaloids identified through the semi-untargeted method.

Conclusions and Impact

Methodological Innovation

First application of MRM-IDA-EPI for comprehensive lupin alkaloid profiling, enabling identification without all reference standards

Geographical Traceability

Successfully distinguished samples from four Italian regions with high accuracy using alkaloid fingerprints

Food Safety Applications

Provides valuable tool for product traceability, quality assessment, and consumer information

> The integration of semi-untargeted methods with multivariate chemometrics proved essential for comprehensive geographical classification, demonstrating that targeted approaches alone are insufficient for capturing the full chemical variability of lupin alkaloid profiles.