# Multivariate Analysis in Food Science

Simplifying complex food data through dimensionality reduction techniques, multivariate analysis provides crucial insights for various applications. It helps in understanding intricate relationships in **sensory analysis** (taste, texture, aroma), **food quality control**, **ingredient optimization**, **nutritional profiling**, **consumer preference studies**, **shelf-life prediction**, **food authentication**, and **process optimization** in food manufacturing.
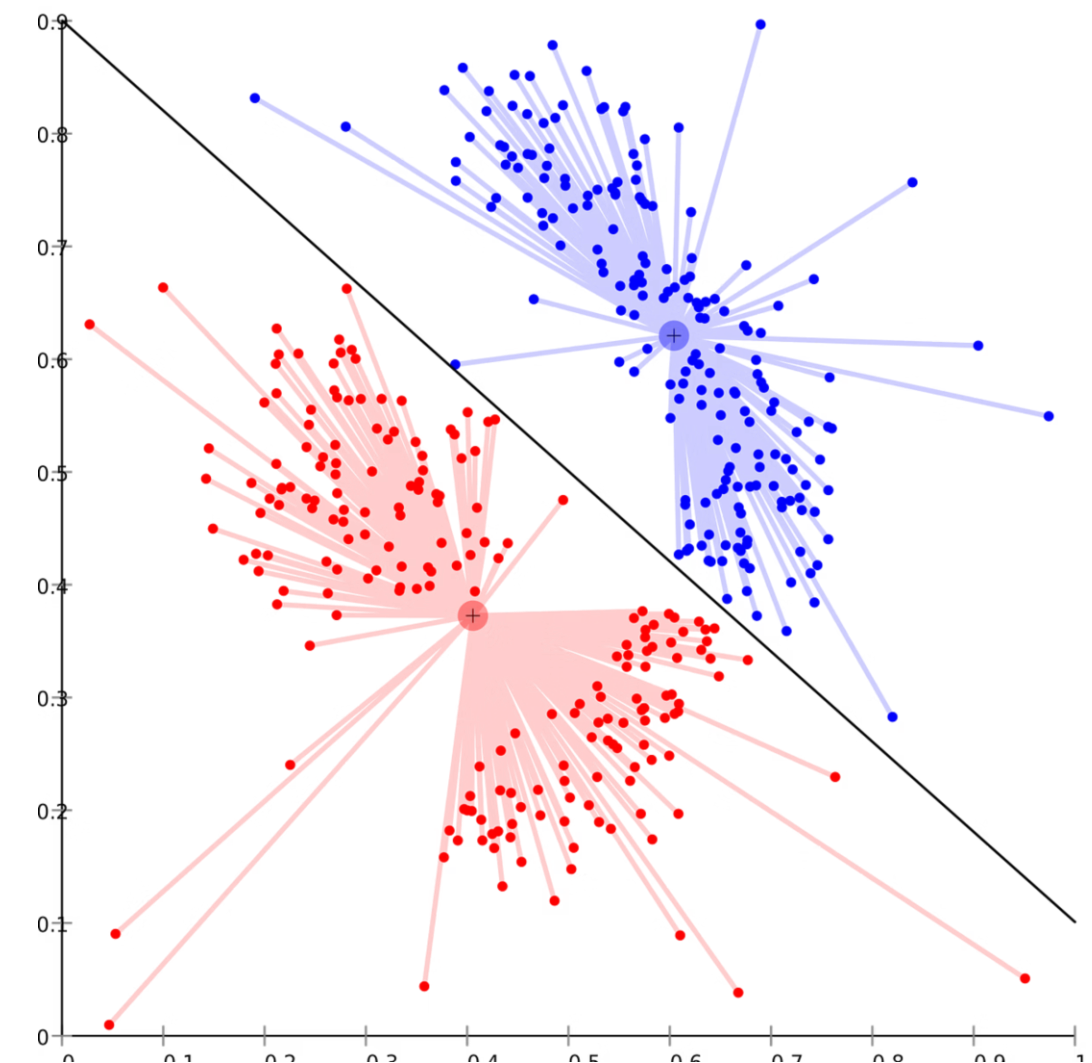
What is Multivariate Analysis?

Multivariate analysis encompasses a range of **statistical methods** designed to analyse multiple variables simultaneously, uncovering hidden relationships and patterns within complex datasets.

These techniques are essential across diverse fields, particularly in food science:

- **Sensory Analysis** – understanding taste, texture, and aroma profiles
- **Food Quality & Safety** – predicting shelf-life and authenticating ingredients
- **Product Development** – optimizing ingredients and understanding consumer preferences
- **Process Optimization** – enhancing efficiency in food manufacturing

Principal Component Analysis (PCA) stands as one of the most powerful techniques within this analytical toolkit, offering elegant solutions for data simplification.

Understanding Multivariate Data

Multivariate data involves recording **multiple measurements or variables** for each individual observation. This comprehensive approach allows for deeper insights than analyzing variables in isolation, providing a holistic view of complex systems.

Dataset

Details

Observations
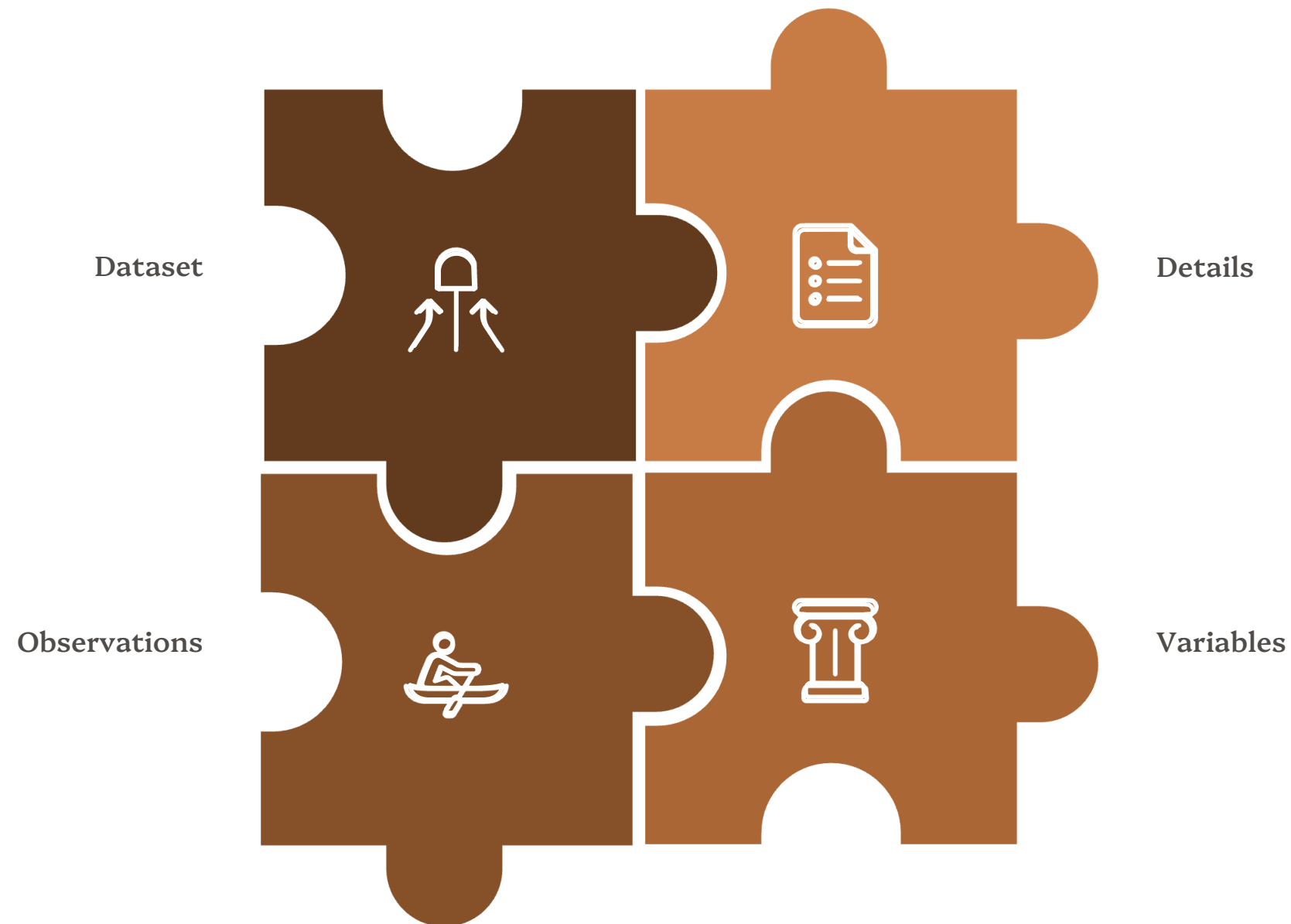
Variables

Why Multiple Variables Matter

Food quality, safety, and consumer preference are rarely determined by a single factor. Instead, they result from the complex interplay of numerous attributes. Multivariate analysis allows us to uncover these crucial interactions, leading to more robust product development and quality control strategies.

The Problem with Many Variables

While multivariate data offers rich insights, managing and analysing high-dimensional datasets presents several significant challenges in food science applications.

Visualization Challenge

Visualizing complex sensory profiles (e.g., taste, texture, aroma across many attributes) or hundreds of chemical markers in food becomes extremely difficult to plot and understand when exceeding three dimensions.

Computational Complexity

Analyzing extensive datasets from food quality control, ingredient optimization, or shelf-life prediction requires substantial computational resources due to the exponential growth in calculations with more variables.

Interpretation Difficulty

Identifying meaningful patterns and relationships from vast data on consumer preferences, nutritional profiles, or process parameters becomes overwhelming as the sheer number of interacting variables increases.

What Multivariate Analysis Does

In food science, multivariate analysis acts as a powerful bridge, transforming overwhelming datasets from sources like sensory panels or quality control into actionable insights. It systematically unravels the inherent complexity of food data to reveal underlying structures and relationships crucial for product development and optimization.

Complex Food Data Inputs

Begins with high-dimensional datasets, such as extensive sensory analysis profiles (taste, texture, aroma), detailed nutritional compositions, or spectroscopic data from quality control, often too complex for intuitive interpretation.

Multivariate Techniques Applied

Applies methods like Principal Component Analysis (PCA) to identify key relationships between food attributes, reduce data complexity, and preserve crucial information for areas like ingredient optimization or shelf-life prediction.

Simplified Insights for Food Science

Delivers distilled, understandable data, enabling better decision-making in areas such as consumer preference studies, food authentication, or process optimization in food manufacturing, by uncovering hidden trends and drivers.

Real-World Example: Food Science Applications

Imagine a food company gathering extensive data on its products: **sensory attributes (taste, texture, aroma), chemical composition, processing parameters, and consumer preference scores**. Using multivariate analysis, food scientists can identify key drivers of consumer acceptance, optimize ingredient formulations for desired sensory profiles, or detect variations in product quality. This allows them to predict shelf-life, ensure food authentication, fine-tune manufacturing processes, and tailor nutritional profiles, leading to improved product development and consumer satisfaction.

Why Dimensionality Reduction?

The Challenge

In food science, data from sensory panels, quality control, or nutritional profiling often involve hundreds of interconnected variables (e.g., taste attributes, chemical compounds, processing parameters), making comprehensive analysis and interpretation extremely complex.

The Solution

Techniques like PCA (Principal Component Analysis) identify a condensed set of new variables (principal components) that capture the most significant information from these numerous food-related variables, effectively reducing data dimensionality.

The Benefit

This simplification allows food scientists to uncover underlying patterns, predict shelf-life, optimize ingredient formulations, ensure food authentication, or understand consumer preferences more effectively by reducing noise and revealing hidden structures.
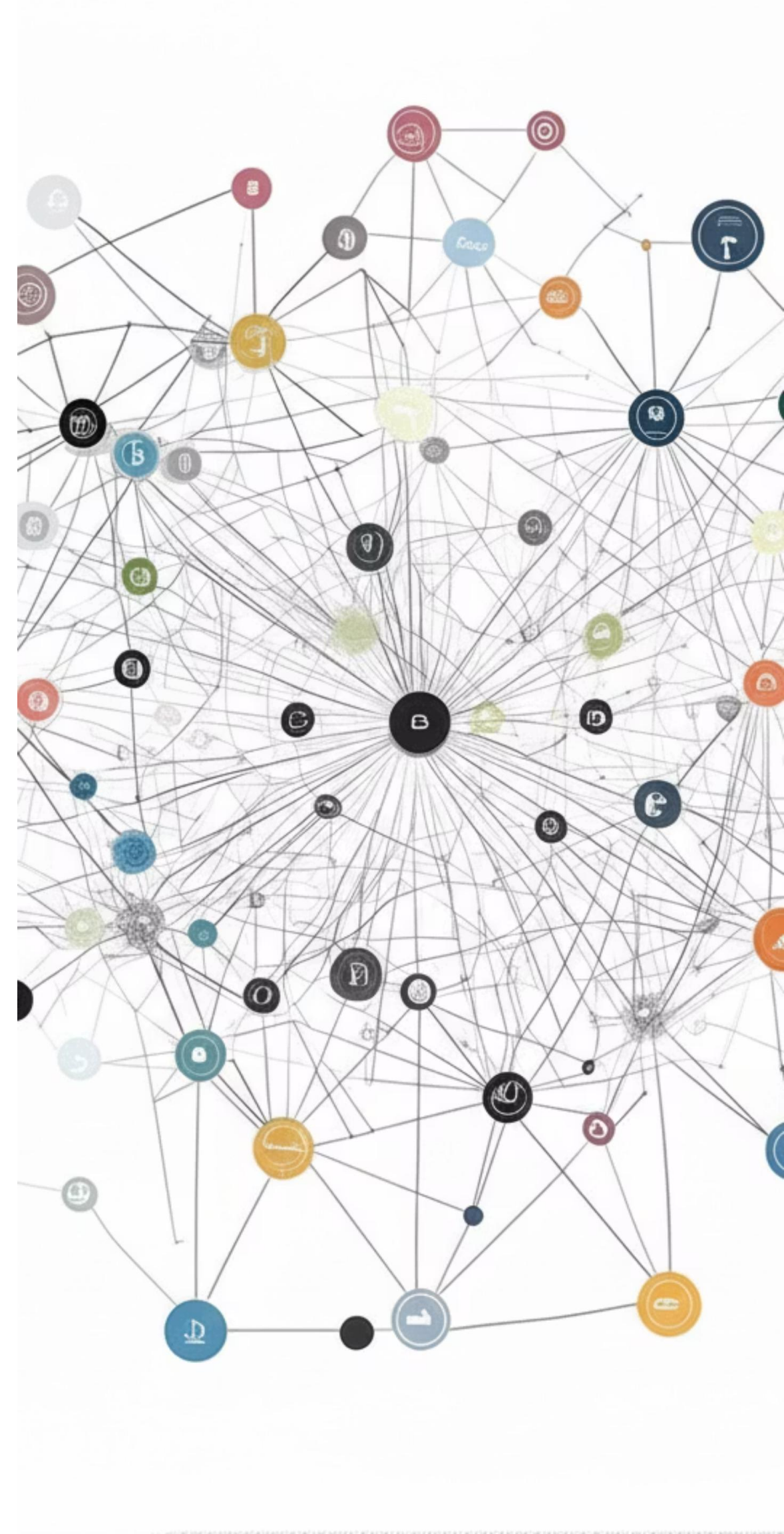
PCA in a Nutshell: A Food Science Perspective

In food science, Principal Component Analysis transforms your original measurements (e.g., sensory attributes, chemical compositions, physical properties) into **new, uncorrelated variables** called principal components.

Each principal component represents a **linear combination** of the original variables, weighted to capture maximum variance present in complex food datasets.

The Hierarchy of Components in Food Data

- **PC1** – captures the greatest variance, often related to the most dominant sensory attribute or chemical profile difference (e.g., sweetness vs. bitterness, major flavor components).

- **PC2** – captures the second greatest variance, highlighting another significant source of variation (e.g., texture differences, secondary aroma notes).

- **PC3, PC4...** – capture progressively less variance, revealing more subtle distinctions in food samples.

By selecting only the top components, food scientists can retain most crucial information for applications like sensory analysis or quality control, while dramatically reducing the complexity of high-dimensional data.

**Population Genetics
3D Principal Component Analysis (PCA)**

Legend:
- Population 1
- Population 2
- Population 3
- Population 4

Axes: PC1, PC2, PC3
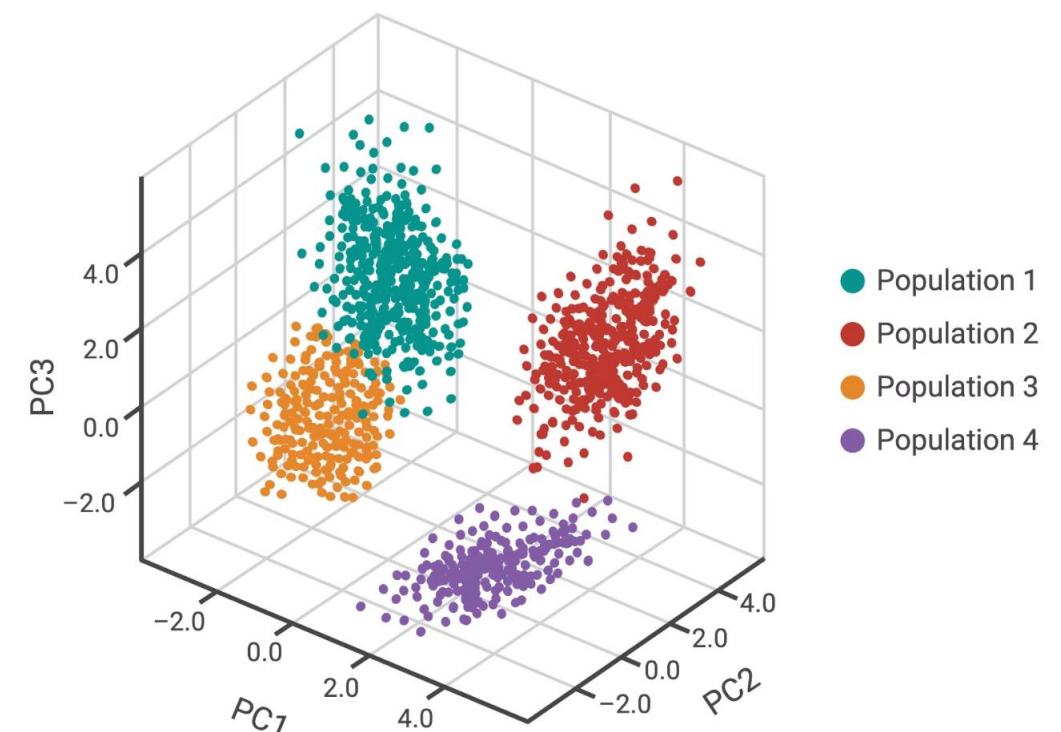
Geometric Intuition of PCA in Food Science

Understanding Principal Component Analysis (PCA) becomes intuitive when you visualize complex food science data – such as sensory attributes, nutritional profiles, or processing parameters – as a **cloud of points** in a multi-dimensional space.

Food Data Cloud

Each food product, described by numerous attributes (e.g., sweetness, bitterness, firmness, pH, moisture), becomes a point in this data cloud, revealing inherent patterns of variation.

Identify Key Drivers

PCA then finds the principal directions (axes) that capture the most significant variability within this food data, such as the primary drivers of consumer preference or quality differences.

Simplified Dimensions

These identified directions transform into new coordinate axes, allowing us to simplify the complex relationships between food attributes and gain insights into underlying structures, like distinguishing product categories or identifying key sensory drivers.
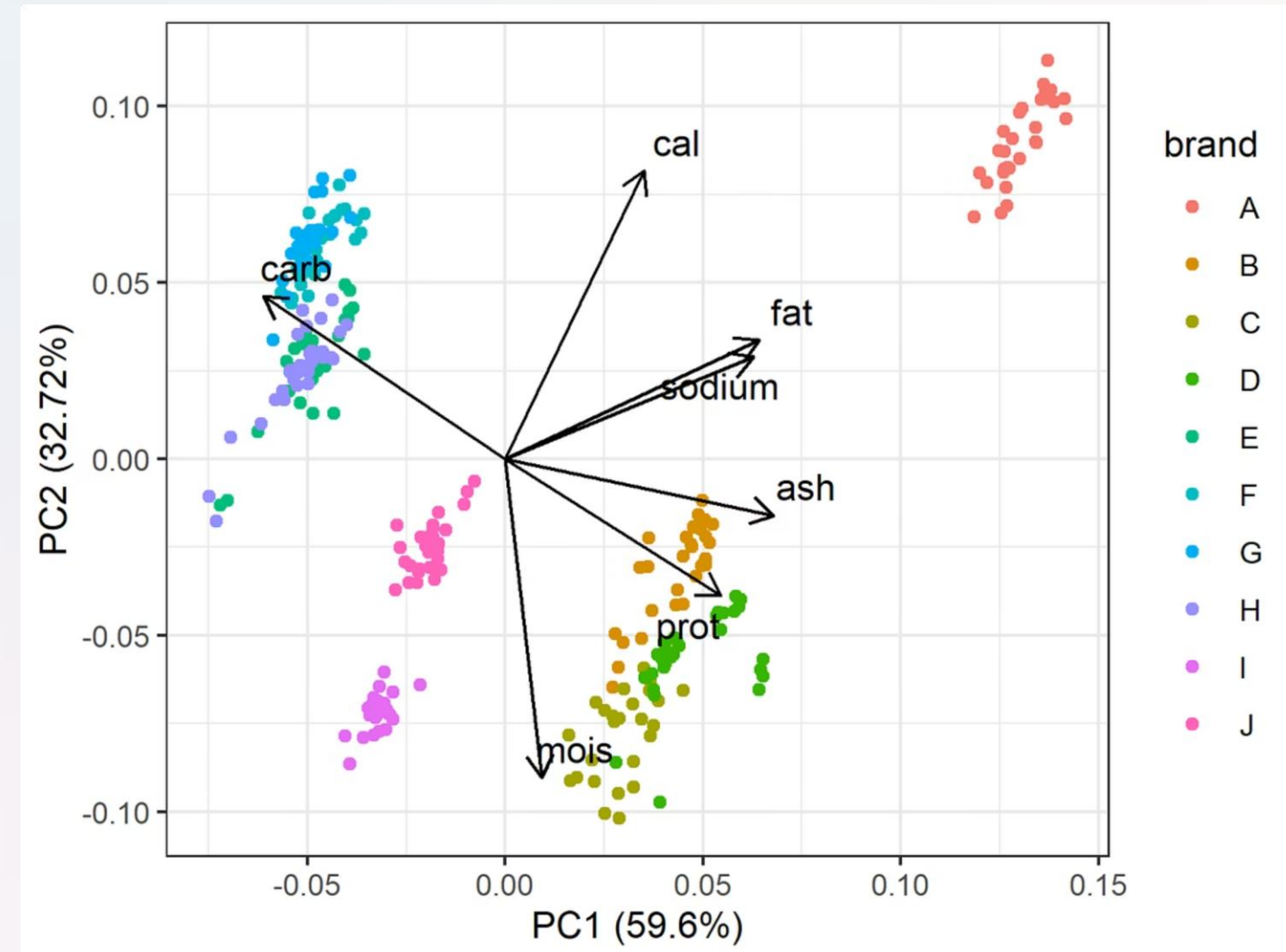
How PCA Works Mathematically in Food Science

01

Compute Covariance Matrix of Food Attributes

Measure relationships between all food attribute pairs (e.g., sensory scores, chemical components, texture measurements) to understand how they co-vary across different food samples.

02

Eigen-Decomposition of Food Data

Perform mathematical decomposition to extract eigenvalues and eigenvectors from the covariance matrix, representing the patterns and variations in the food attribute data.

03

Interpret Principal Components

Eigenvectors become principal components (e.g., underlying sensory dimensions like 'freshness' or 'richness'); eigenvalues quantify the variance explained by each component in the overall food characteristics.

04

Project Data for Dimensionality Reduction

Project original complex food attribute data onto the new principal components, effectively reducing many variables (e.g., numerous ingredient concentrations or consumer preference scales) to fewer, more interpretable dimensions for easier analysis and visualization.

Eigenvalues and Eigenvectors Explained

The Mathematical Foundation

In PCA, **eigenvectors** define the new coordinate axes, or Principal Components (PCs), representing the directions of maximum variance within your complex food dataset. Imagine finding the primary 'lines' along which your food samples spread out most significantly.

Each eigenvector has a corresponding **eigenvalue**. This value quantifies the amount of variance captured by that specific Principal Component. Essentially, it tells you how much "information" or variation is explained by that particular direction.

Applying to Cheese Samples

Consider analyzing various cheese samples with attributes like saltiness, sharpness, aroma, and texture.

*   **PC1 (largest eigenvalue):** This eigenvector might point towards "overall intensity." Cheeses that vary most in their combined saltiness, sharpness, and potent aroma would align strongly with this direction. Its large eigenvalue indicates it captures the most significant differences among cheeses.

*   **PC2 (second largest eigenvalue):** This eigenvector could represent a "texture dimension," differentiating cheeses based on their firmness versus creaminess. Its smaller eigenvalue means it explains less variation than PC1 but still highlights an important distinguishing factor.

Key takeaway: Larger eigenvalues correspond to more important Principal Components, highlighting the most dominant characteristics that differentiate your food products.

Data Preparation and Standardization

In food science, data often comes from varied sources with different scales and units – think pH (0-14), moisture content (0-100%), or sensory sweetness scores (1-10). Without proper handling, variables with inherently larger numerical ranges would disproportionately influence PCA, masking the true impact of other important but smaller-scale attributes.

Standardization, particularly using the z-score method, addresses this by transforming each variable to have a mean of zero and a standard deviation of one. This places all attributes on an equal footing, ensuring each contributes fairly to the analysis.

$$z = \frac{x - \mu}{\sigma}$$

Where:

- x is an individual data point
- μ is the mean of the variable
- σ is the standard deviation of the variable

CRITICAL: Without standardization, PCA results can be heavily biased towards variables with larger numerical scales, obscuring genuine relationships among food attributes.

Step 1: Linear Combinations of Variables

The Foundation of Principal Components

Each Principal Component (PC) is a new variable formed by combining your original food attributes in a **weighted sum**, known as a linear combination.

$$PC_k = w_1X_1 + w_2X_2 + \ldots + w_nX_n$$

Here, X represents the original attributes (e.g., sweetness, acidity), and w are the weights (loadings) derived from the eigenvectors. These weights quantify each attribute's contribution to that specific PC.

"Overall Palatability" in Food Science
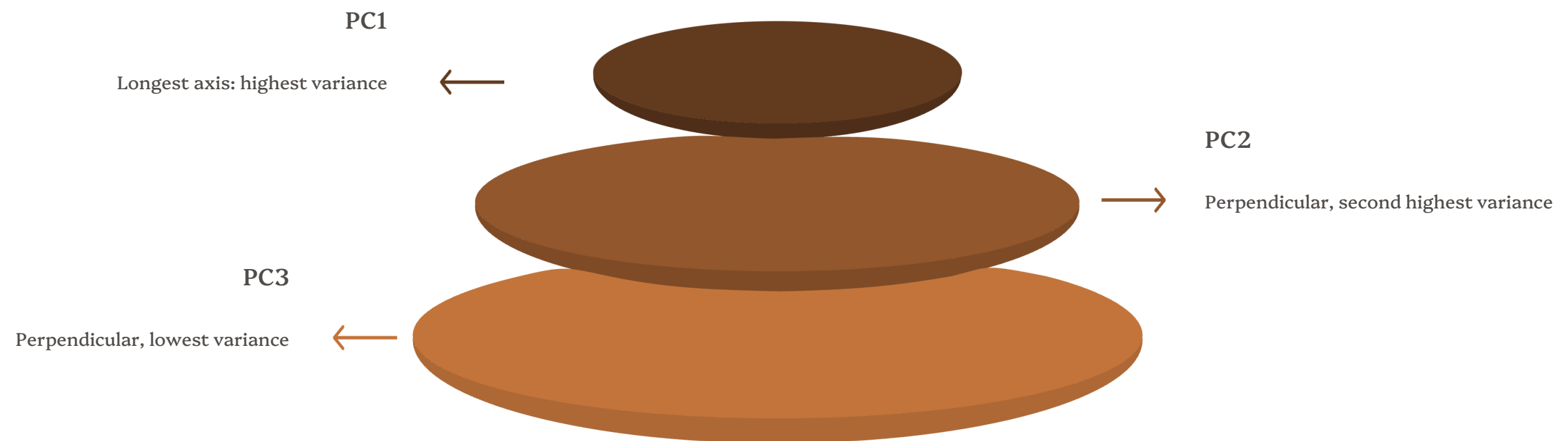
Imagine a PC1 for a food product defined as:

```
PC1 = 0.5×sweetness + 0.4×sugar + 0.3×liking - 0.2×bitterness
```

This PC could represent an "overall palatability" dimension, effectively summarizing several sensory traits into a single, interpretable score for each food item, allowing for quick comparisons.

□ **Weight Interpretation:** Positive weights mean an attribute increases with the PC; negative weights indicate it decreases as the PC increases.

Step 2: Orthogonality - Independent Dimensions

A fundamental property of Principal Components (PCs) is their **orthogonality**, meaning they are statistically uncorrelated and mathematically perpendicular to one another. This ensures that each PC captures a unique, independent source of variation present in your original food data.

PC1

Longest axis: highest variance ←

PC2

→ Perpendicular, second highest variance

PC3

Perpendicular, lowest variance ←

PC1: Flavor Intensity

Captures the largest variation, differentiating cheeses by their overall sensory strength (e.g., strong vs. mild).

PC2: Texture Dimension

Perpendicular to PC1, it explains the next most variance, distinguishing cheeses based on physical characteristics (e.g., firm vs. creamy).

PC3: Aging Characteristics

Orthogonal to both PC1 and PC2, it captures the remaining variance, highlighting nuances related to maturation (e.g., aged vs. fresh notes).

Step 3: Projecting Data onto Principal Components

Once the Principal Components (PCs) are defined as new axes, each original food sample, with its many attributes, is transformed by being **projected** onto these new, lower-dimensional components. This process assigns a unique "score" to each sample for every PC, effectively translating complex raw data into a more manageable and interpretable format.

High-Dimensional Data

Projection Flow

PC1 & PC2 Scores
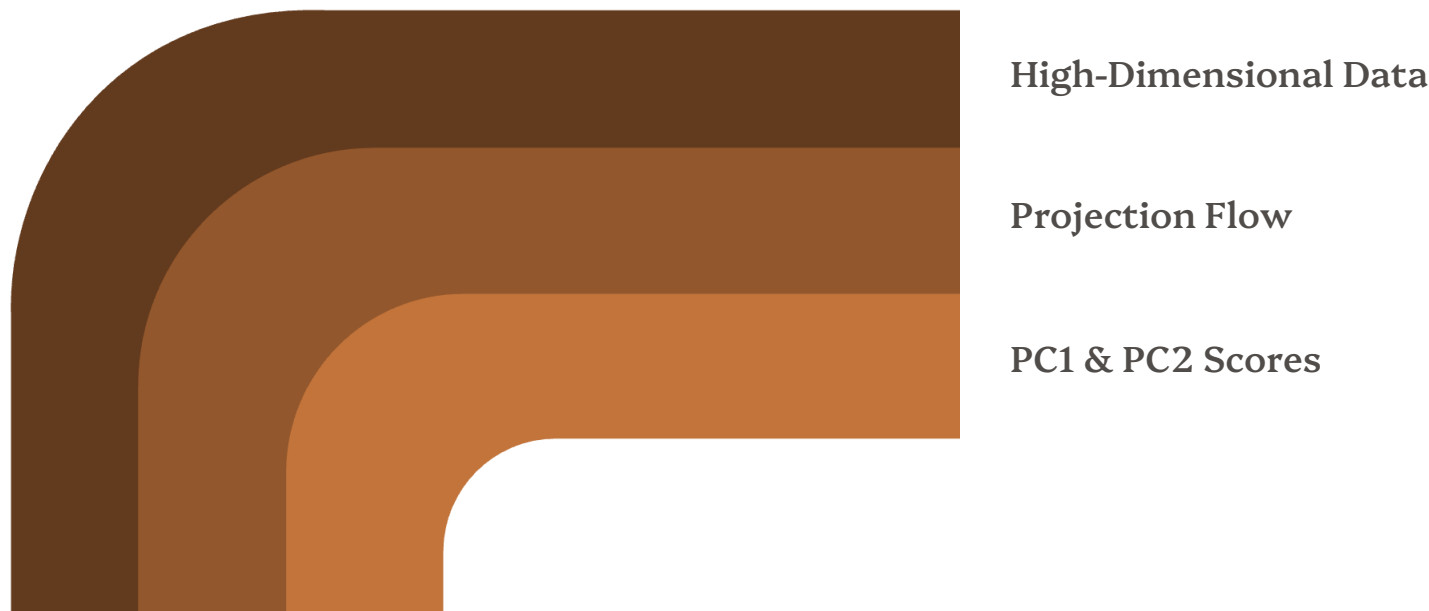
Understanding the Scores

Each PC score represents how strongly a particular food sample aligns with that specific Principal Component. A positive score on PC1, for instance, might indicate a strong presence of the characteristics captured by PC1 (e.g., "overall intensity"), while a negative score might indicate the opposite.

These scores effectively condense the information from numerous original variables into a few meaningful numbers, making it easier to compare and classify food products based on their dominant characteristics.

Cheese A Example

Consider a specific cheese sample, Cheese A, defined by several sensory attributes:

- Sweetness = 7
- Sugar = 12%
- Texture = 8
- Bitterness = 2

After projection onto the newly defined Principal Components, Cheese A's characteristics are summarized by its PC scores:

- **PC1 Score** = **4.2**
- **PC2 Score** = **-1.5**

These scores now represent Cheese A's position in the reduced-dimensional PC space.

These generated PC scores are the foundation for subsequent data visualization, clustering, and further analysis, allowing food scientists to uncover patterns and relationships that would be obscured in the original, high-dimensional data.

Choosing How Many Components to Keep

1

Scree Plot (Elbow Method)

Visually inspect a plot of eigenvalues in descending order. The "elbow" point, where the slope of the curve sharply decreases, often indicates the optimal number of components to retain before the remaining components contribute little to explaining variance.
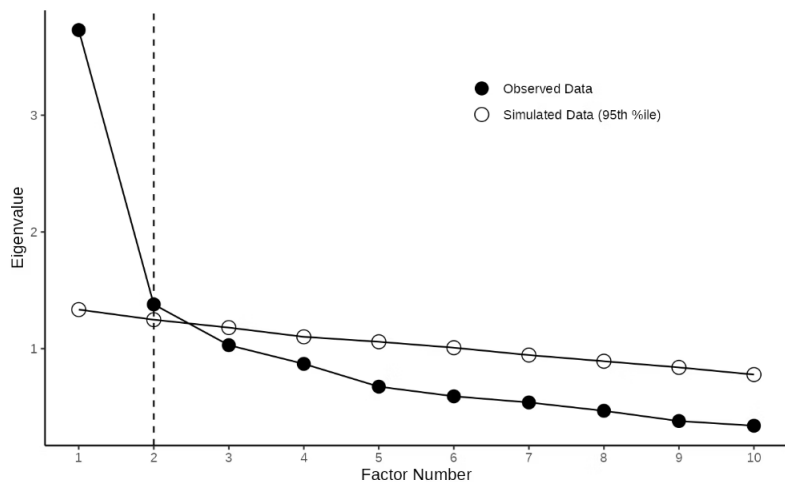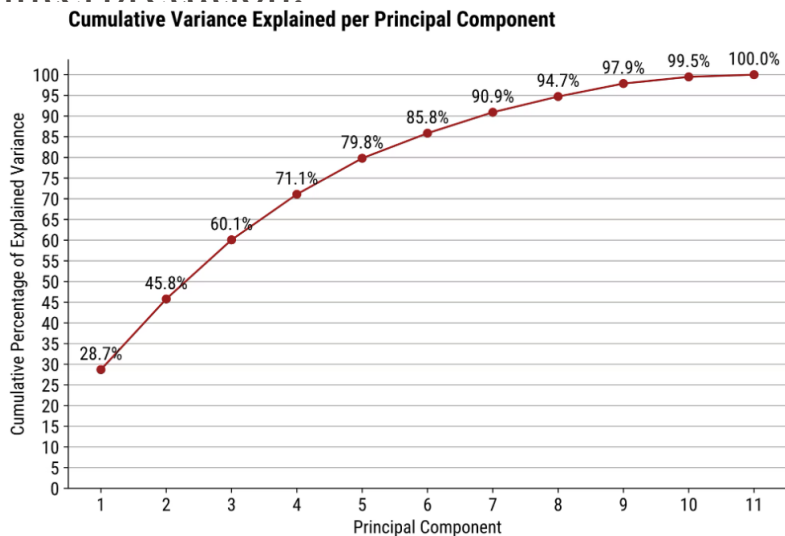

2

Cumulative Variance Explained

Select the minimum number of principal components that collectively explain a significant portion of the total variance in the original data. Common thresholds in food science are typically between **70-90%**, ensuring enough information is retained for meaningful interpretation.
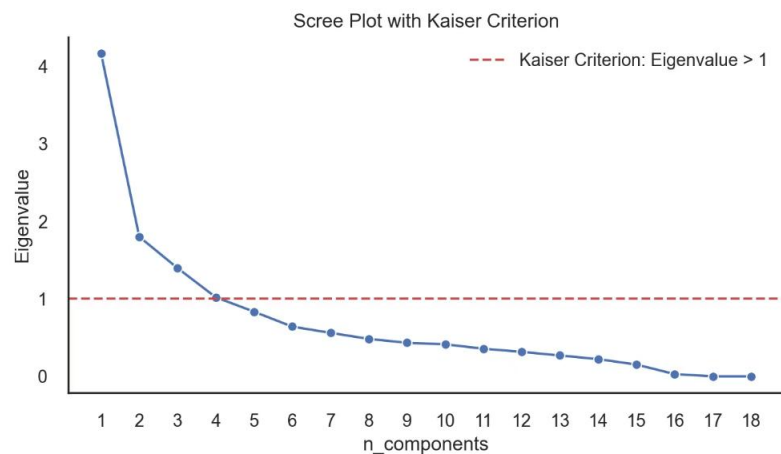

3

Kaiser Criterion

A simple rule of thumb: keep only those principal components with an eigenvalue **greater than 1**. Each such component explains more variance than a single original standardized variable, suggesting it captures meaningful information.

Interpreting PCA Results: Loadings and Scores

Loadings: Unveiling Component Meaning

Loadings are coefficients that quantify the contribution of each original food attribute (e.g., pH, sweetness, texture) to a specific Principal Component (PC). High positive or negative loadings indicate a strong relationship. For instance, if PC1 has high positive loadings for sweetness, sugar content, and consumer liking, it reveals PC1 represents a **"sweetness/preference" dimension**. Loadings help us understand **WHAT** each PC truly represents.

Scores: Positioning Food Samples

Scores are the coordinates of each individual food sample in the new, reduced Principal Component space. After projecting the original data onto the PCs, each sample gets a score for every PC. These scores can be plotted to visualize relationships between samples, identify clusters of similar products, or pinpoint outliers. Scores show us **WHERE** each food sample falls within these newly defined dimensions, aiding in product comparison and differentiation.

Figure 2: Biplot of PCA

PCA scores
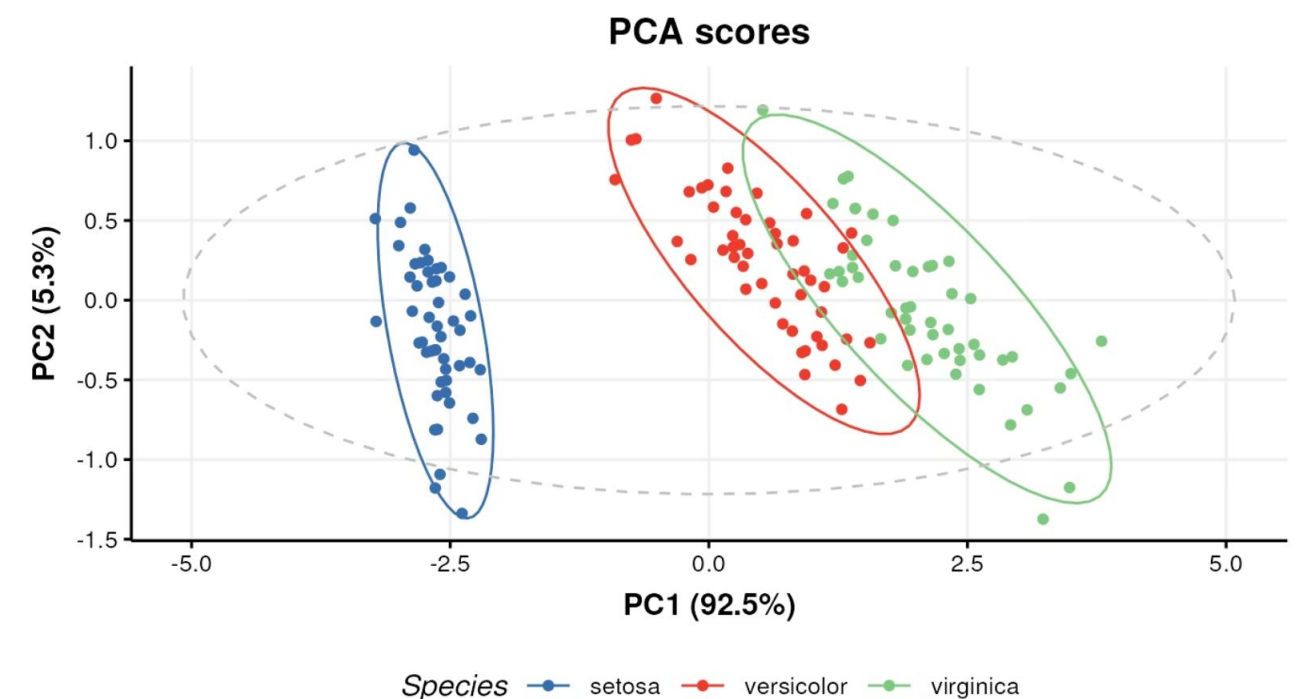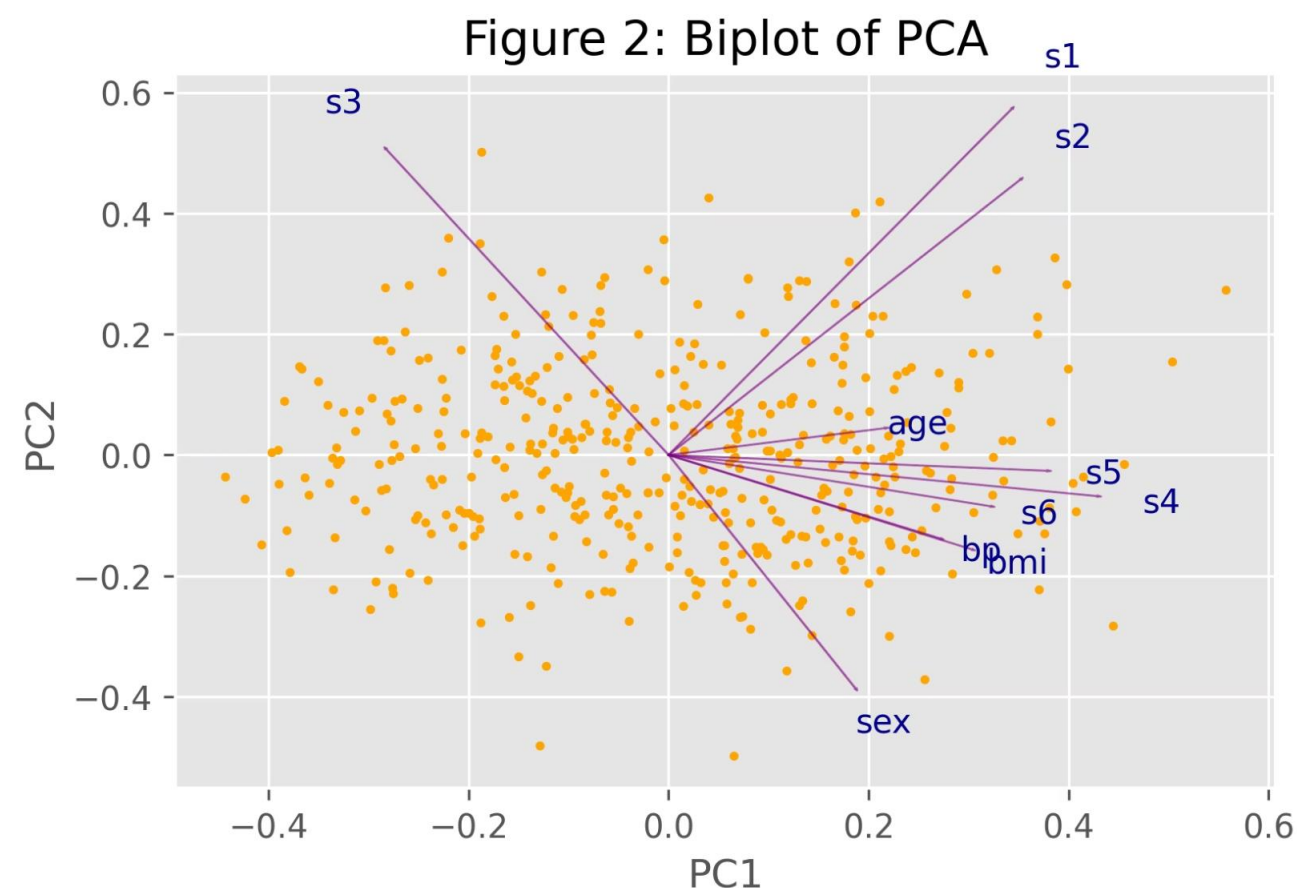
Understanding the Covariance Matrix

What is Covariance?

In food science, it quantifies how two attributes, like sweetness and consumer liking, change together. It reveals if their trends are synchronous or independent across different samples.

Positive Covariance

Both variables tend to increase or decrease in tandem. For example, higher sugar content often correlates with increased sweetness perception and consumer preference in a beverage.

Negative Covariance

As one variable increases, the other tends to decrease. For instance, increased bitterness might lead to decreased consumer acceptance and lower overall palatability for some products.

The Covariance Matrix

This square matrix organizes all pairwise covariances between food attributes. It's a foundational step in PCA, capturing the complete inter-relationship structure of your food data.

Covariance Matrix Heatmap
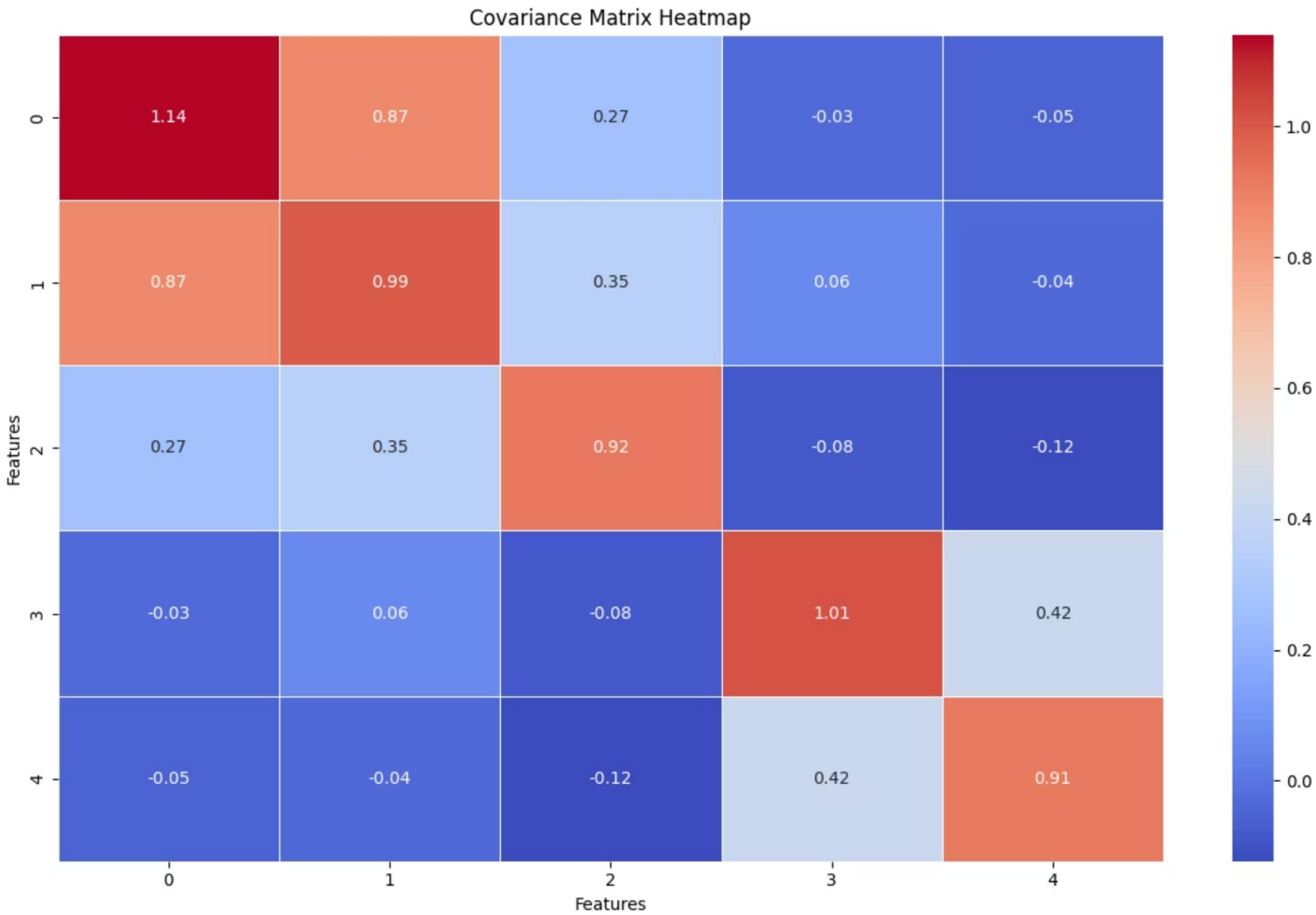
Visual Example: PCA for Sensory Data

Biplot

Label
- 🔵 Loadings
- ⚫ PC scores
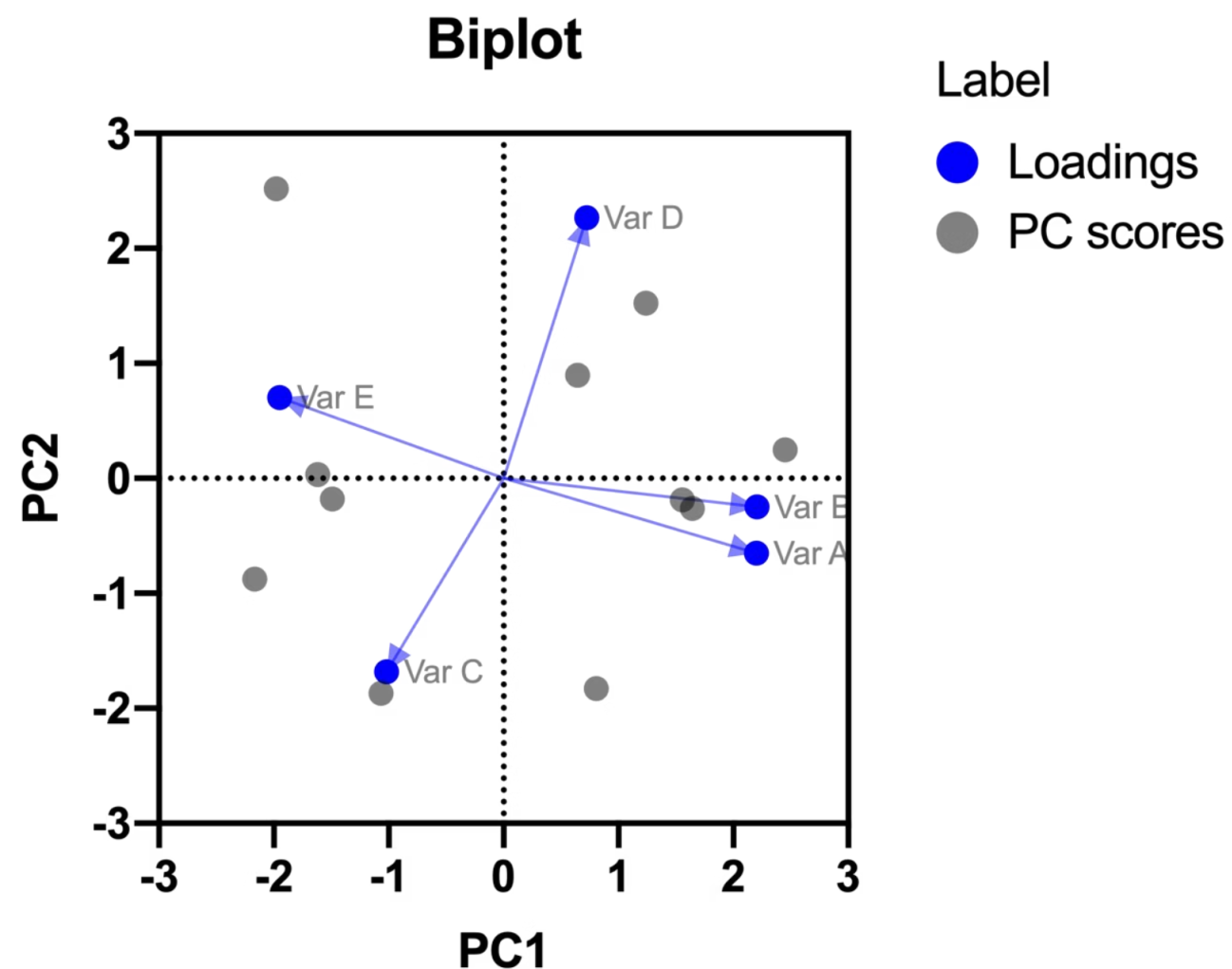
## Understanding PCA in Sensory Data

Consider various food samples assessed across multiple sensory attributes (e.g., sweetness, bitterness, texture). PCA performs the following:

1. **PC1** – identifies the primary direction of variation, capturing the greatest differences in sensory profiles among food samples.

2. **PC2** – finds the orthogonal axis capturing the second most significant variation, which might relate to other distinct sensory characteristics or ingredient contributions.

🗒 **Key Insight:** If a major characteristic (e.g., overall flavor intensity or a dominant ingredient) accounts for most of the variance, PCA can effectively reduce complex multi-dimensional sensory data into a more interpretable, lower-dimensional representation.

PCA in Food Science: Practical Applications
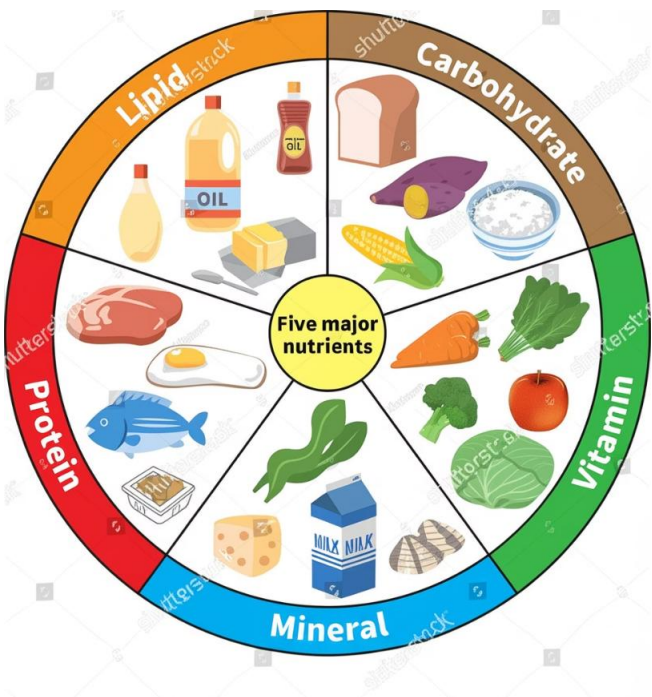
Sensory Analysis
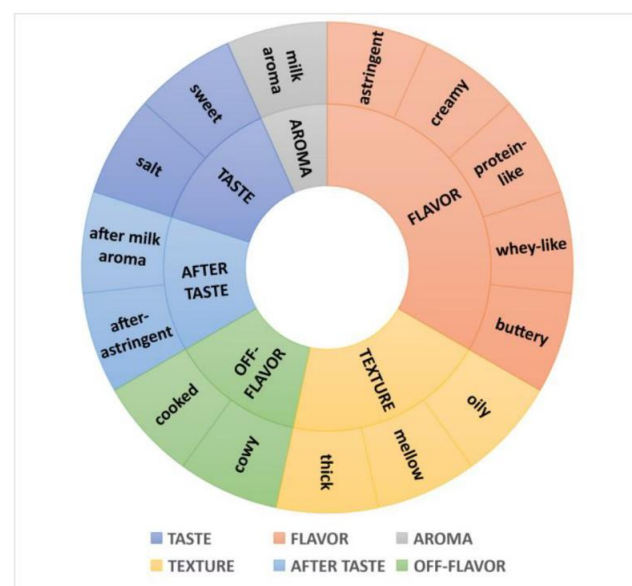
Quantifying complex sensory attributes of food products to identify key drivers of perception (taste, texture, aroma).

Food Quality Control

Monitoring production processes to ensure consistent product characteristics and detect deviations from quality standards.

Nutritional Profiling

Identifying patterns in nutrient composition data to optimize recipes and assess the nutritional impact of food products.

Shelf-Life Prediction

Analyzing changes in food properties over time to accurately forecast product freshness, stability, and safety.

Limitations and Considerations in Food Science PCA

Linear Assumptions

PCA assumes linear relationships between food properties (e.g., sensory attributes, chemical compositions) and maximizes variance. This linearity may not fully capture complex non-linear interactions in taste profiles or ingredient functionality, potentially oversimplifying real-world food structures.

Scaling Sensitivity

PCA is sensitive to the measurement scales of different food science variables. For instance, combining sensory scores (e.g., 1-9 hedonic scale) with chemical concentrations (e.g., mg/L) without prior standardization can bias results towards variables with larger numerical ranges, thus requiring proper preprocessing.

Categorical Data Handling

Direct application of PCA to categorical food science variables, such as food types, processing methods, or consumer demographics, is not straightforward. Preprocessing methods like one-hot encoding or using alternative techniques (e.g., Multiple Correspondence Analysis) are necessary to integrate such data effectively.

Capturing Nonlinear Patterns

Kernel PCA offers an extension to classical PCA, enabling the capture of complex nonlinear patterns inherent in food science datasets. This is particularly useful for analyzing intricate relationships in sensory perception, ingredient interactions, or consumer preference studies where responses are often non-linear.

Summary: Why PCA Matters

∞

Simplification

Transforms complex multivariate food data (e.g., chemical composition, sensory attributes) into interpretable components like flavor profiles or quality markers.

1st

Foundation

A foundational technique for advanced food science applications such as sensory analysis, food quality control, and ingredient optimization.

100%

Essential

Mastering PCA unlocks deeper insights from high-dimensional food data, crucial for consumer preference studies, shelf-life prediction, and food authentication.

PCA empowers food scientists to visualize, interpret, and model complex food data, focusing on the key patterns that determine product quality, consumer acceptance, and process efficiency.

Carrot Aroma Analysis: E-Nose vs. GC-MS

Exploring the comparative advantages of E-Nose and GC-MS technologies for comprehensive volatile compound analysis in stored carrots.

E-Nose Technology: Hairpin-DNA Sensors

1   **What is an E-Nose?**

A device that mimics the human sense of smell, designed to detect and identify volatile compounds and gas mixtures.

2   **How hpDNA Sensors Work**

They use gold nanoparticles and specific "hairpin" DNA sequences that reconfigure upon contact with the analyte, generating a measurable signal.

3   **Utility in Food Analysis**

It offers a rapid, economical, and simple alternative to traditional GC-MS for monitoring food quality and freshness.

4   **Sensor Array**

The system employs an array of 8 different hpDNA sequences immobilized on quartz crystal microbalances for detailed multi-analyte detection.

1 CAGC
2 TTGG
3 CGGG
4 TAAGT
5 CCCGA
6 ATAATC
7 AAGTA
8 CTGCAA
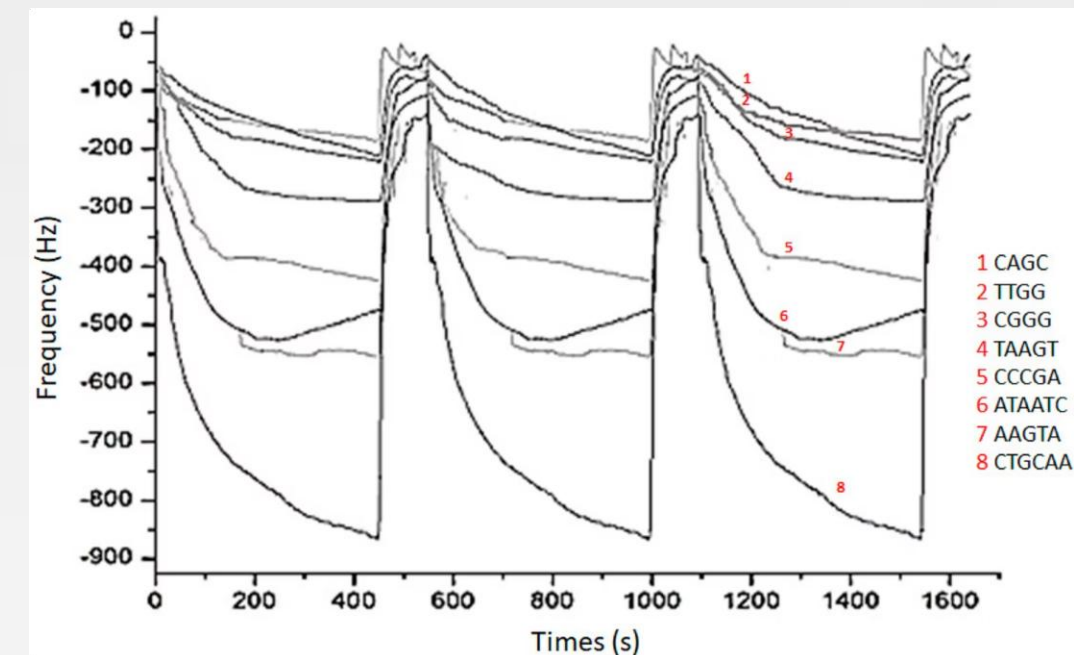
Experimental Study: General Scheme

Carrot Samples

Fresh blanched carrots prepared for the experiment.

Storage at Different Temperatures

26 days at -18°C, 4°C, 25°C, 40°C to simulate various conditions.

Parallel Analysis

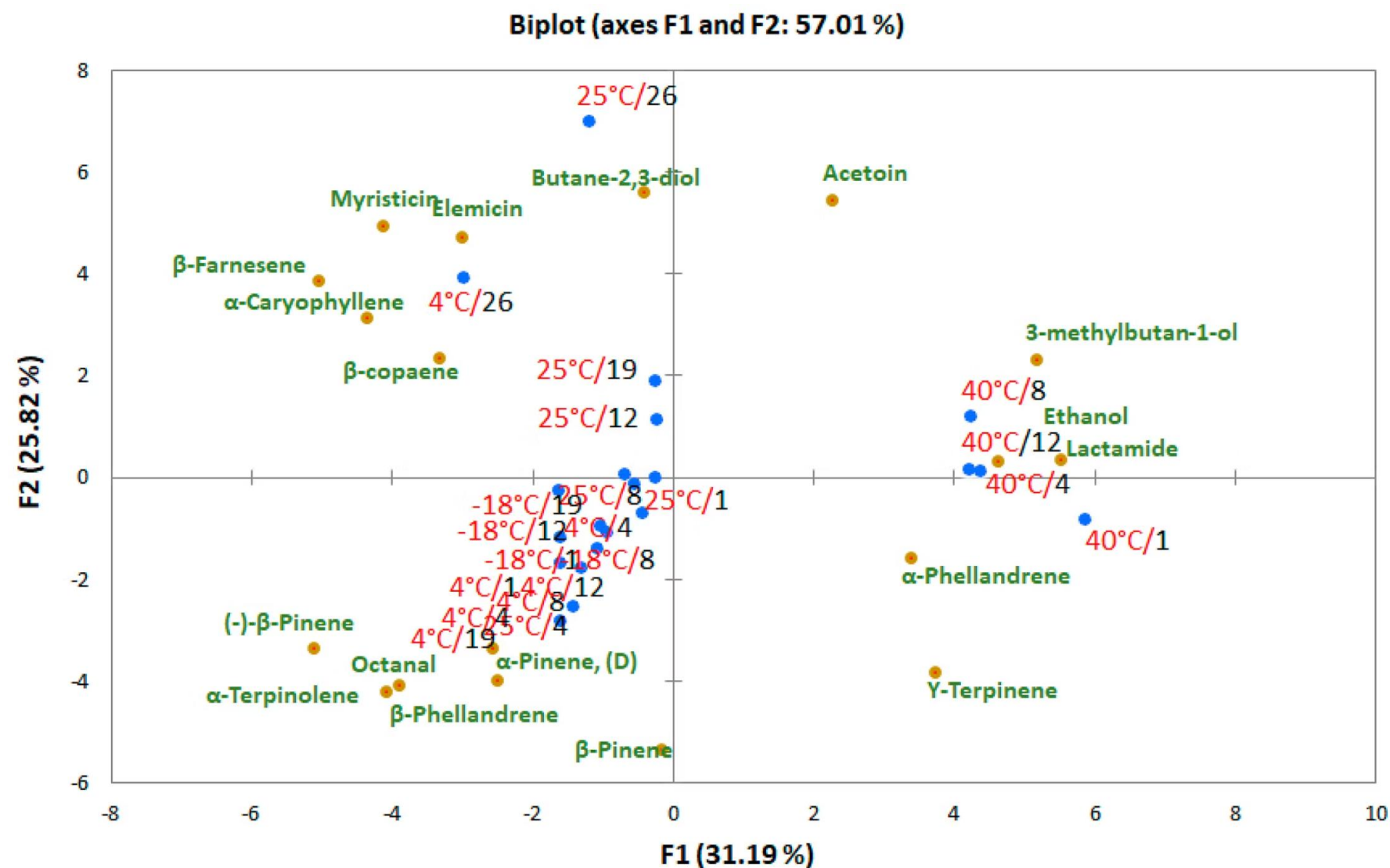E-Nose (hpDNA sensors) and GC-MS for the detection of volatile compounds.

Data Processing

Principal Component Analysis (PCA) to reduce data complexity.

Results and Discrimination

Identification of the correlation between compounds and storage conditions.

PCA of GC-MS Data: Sample Discrimination

Biplot (axes F1 and F2: 57.01 %)

PC1 and PC2 explain 57.01% of total variance (PC1: 31.19%, PC2: 25.82%)

Clear separation of samples based on storage temperature
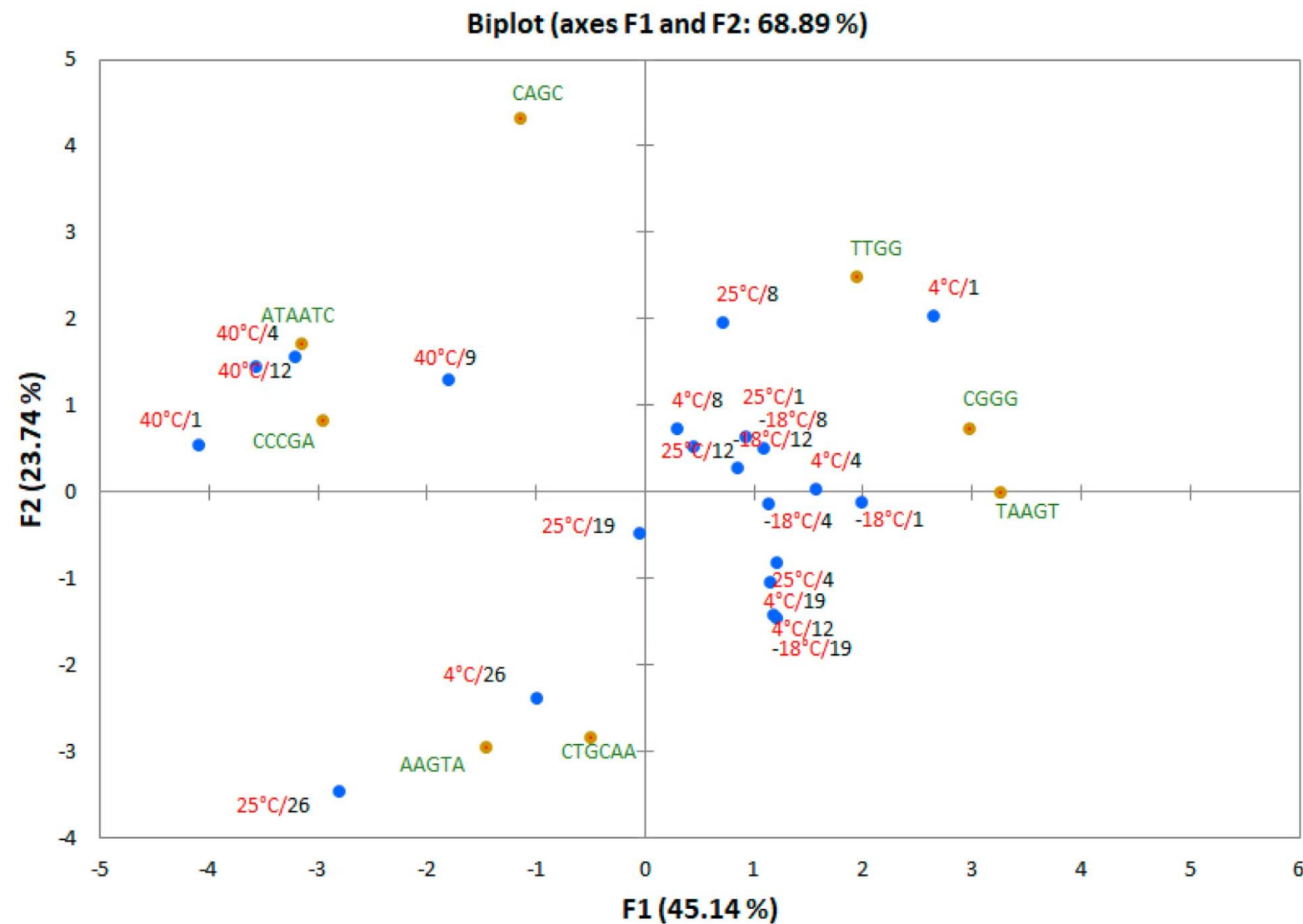
Samples at 40°C clearly separated on PC1

Samples at -18°C and 4°C grouped together (stable conditions)

Samples at 25°C show temporal evolution: days 1-8 vs days 12-26

PC1 discriminates for acetoin, lactamide, ethanol (fermentation)

PC2 discriminates for terpenoids (α-pinene, γ-terpinolene, β-pinene)

E-Nose Data PCA: Comparison with GC-MS

Biplot (axes F1 and F2: 68.89 %)

PC1 and PC2 explain 68.89% of the total variance (higher than GC-MS!)

Comparable or superior discrimination to traditional GC-MS

All samples at 40°C separated in the 4th quadrant

Samples at -18°C and 4°C grouped (1st and 2nd quadrant)

Samples at 25°C for 19 and 26 days separated (3rd quadrant)

All 8 hpDNA sensors contribute to discrimination

Loadings show that pentamers and hexamers have superior performance

Conclusions and Future Perspectives

E-Nose Exceeds Expectations

The hpDNA-based E-Nose discriminates samples with comparable (or superior) efficacy to GC-MS.

Operational Advantages

E-Nose advantages: reduced costs, rapid analysis (8 min), ease of use, no specialized personnel required.

Long-term Reliability

Excellent stability and reproducibility for 3 months of continuous use.

Drift Resistance

No drift issues with high water content samples.

Pioneering in Food Analysis

First study demonstrating the use of hpDNA for solid food matrix analysis.

Industry Potential

Promising application for industrial quality control and shelf-life monitoring.

Innovative Monitoring

Innovative tool for monitoring gases released by vegetables during storage and ripening.

Identifying Cannabis Retailers Through Cannabinoid Analysis

A groundbreaking study demonstrates how advanced chemical analysis combined with multivariate statistics can identify hemp retailers without knowing strain, cultivation methods, or geographic origin. Using HPLC-MS/MS analysis of nine cannabinoids across 161 samples from four Italian retailers, researchers achieved 92% classification accuracy.

The Challenge of Hemp Classification

Traditional Limitations

Current hemp classification relies primarily on THC and CBD concentrations. However, research shows that strains with similar THC/CBD ratios can have vastly different physiological effects on humans.

Italian law classifies Cannabis sativa L. as fiber-type (THC <0.2% w/w) or prohibited drug-type (THC >0.6% w/w), but this binary approach overlooks the plant's chemical complexity.

A More Comprehensive Approach

Cannabis sativa L. contains numerous bioactive compounds including flavonoids, terpenes, steroids, and cannabinoids. The main cannabinoids include neutral forms (THC, CBD, CBC, CBG, CBN, CBDV) and acidic forms (THCA, CBDA, CBGA).

This study analyzed all nine major cannabinoids to create a comprehensive chemical fingerprint for each retailer.

TABLE OF CANNABINOIDS
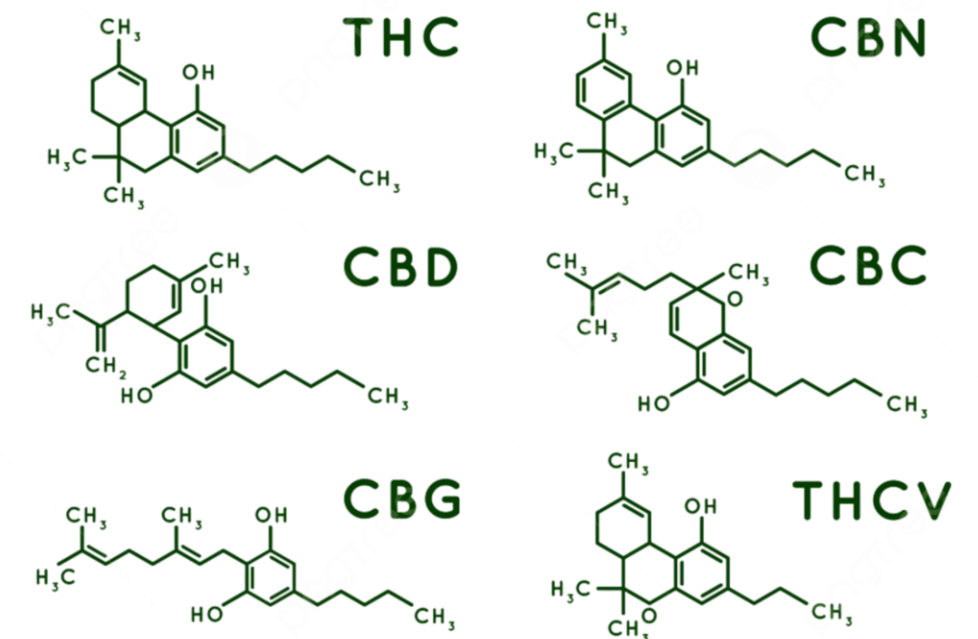
Research Design and Sample Collection

01

Sample Acquisition

161 hemp samples purchased from four Italian retailers (labeled A, B, C, D) across three regions: Lombardy (Milan and Mantova), Lazio (Pomezia), and Abruzzo (Tortoreto). Samples stored at room temperature in original packaging.

02

Sample Preparation

Each sample homogenized through trituration, crushing, and sieving (1mm). Fine powder (10mg) extracted with ethanol using ultrasonic bath for 30 minutes, followed by centrifugation and filtration.

03

HPLC-MS/MS Analysis

Samples analyzed using high-performance liquid chromatography coupled with tandem mass spectrometry. This method prevents cannabinoid decomposition that occurs with traditional GC heating methods.
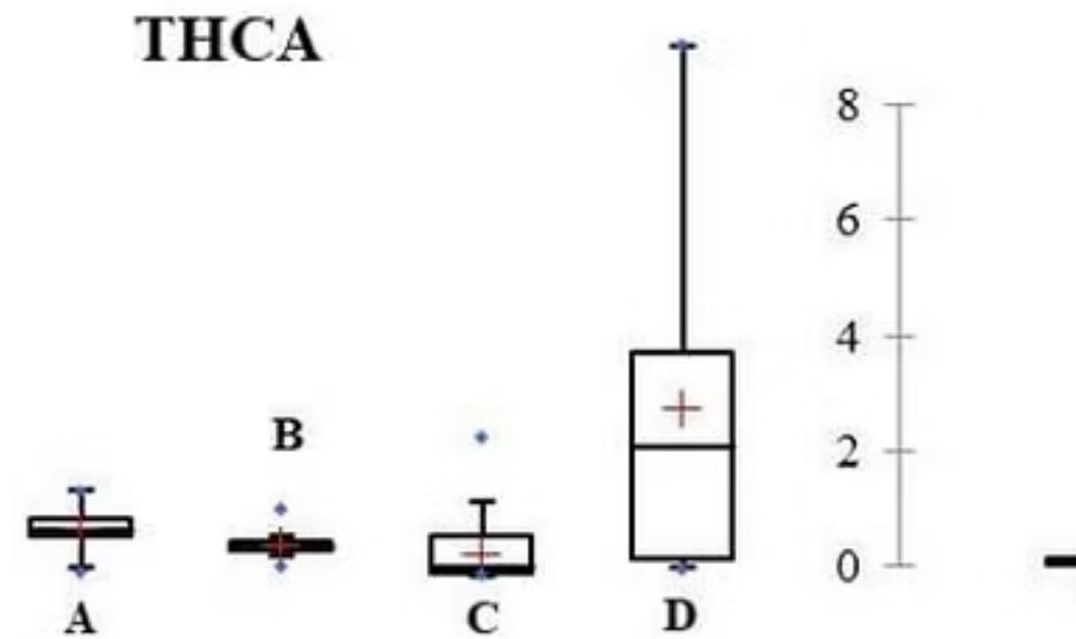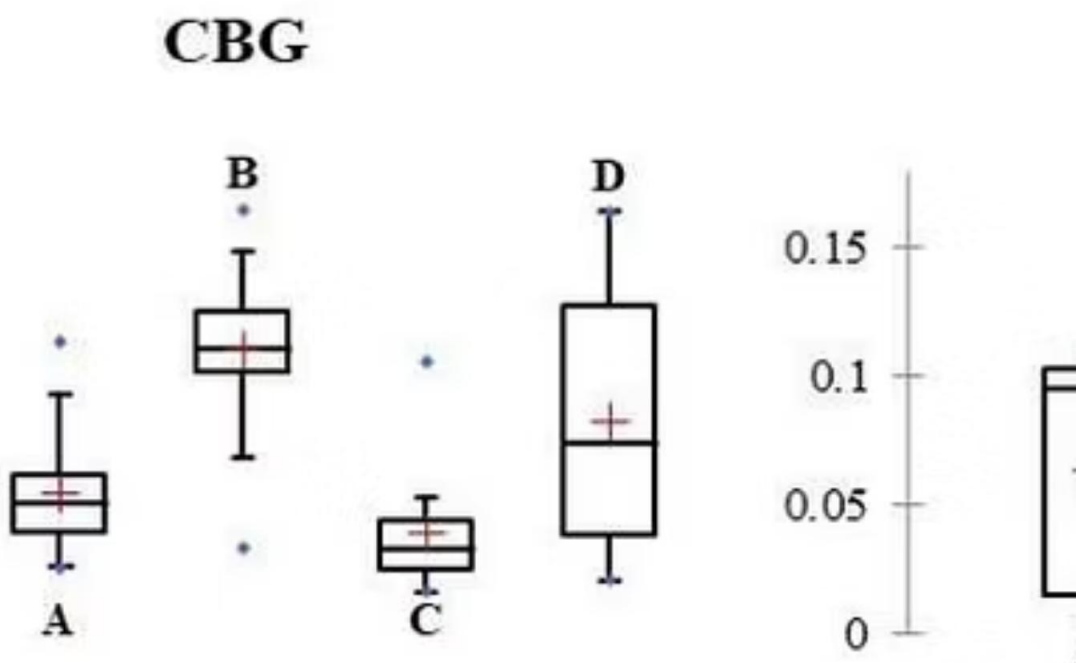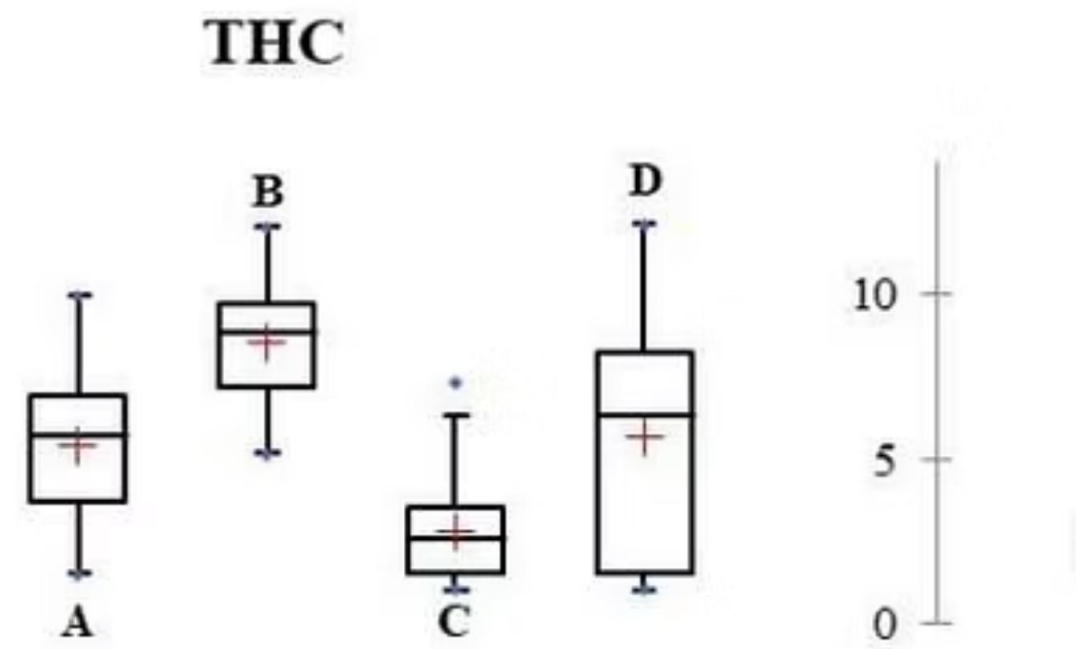
04

Statistical Analysis

Data processed using univariate analysis (ANOVA), unsupervised PCA, and supervised PLS-DA algorithms to identify patterns and classify retailers.

Cannabinoid Profile Analysis Results

Box and whisker plots revealed consistent patterns across all retailers: high concentrations of CBD and CBDA, with lower levels of other cannabinoids. Retailer D showed notably higher THCA and CBGA levels. CBN concentrations were particularly elevated in retailers A and B, potentially indicating oxidation and inflorescence quality markers.

<0.35%

THC Content

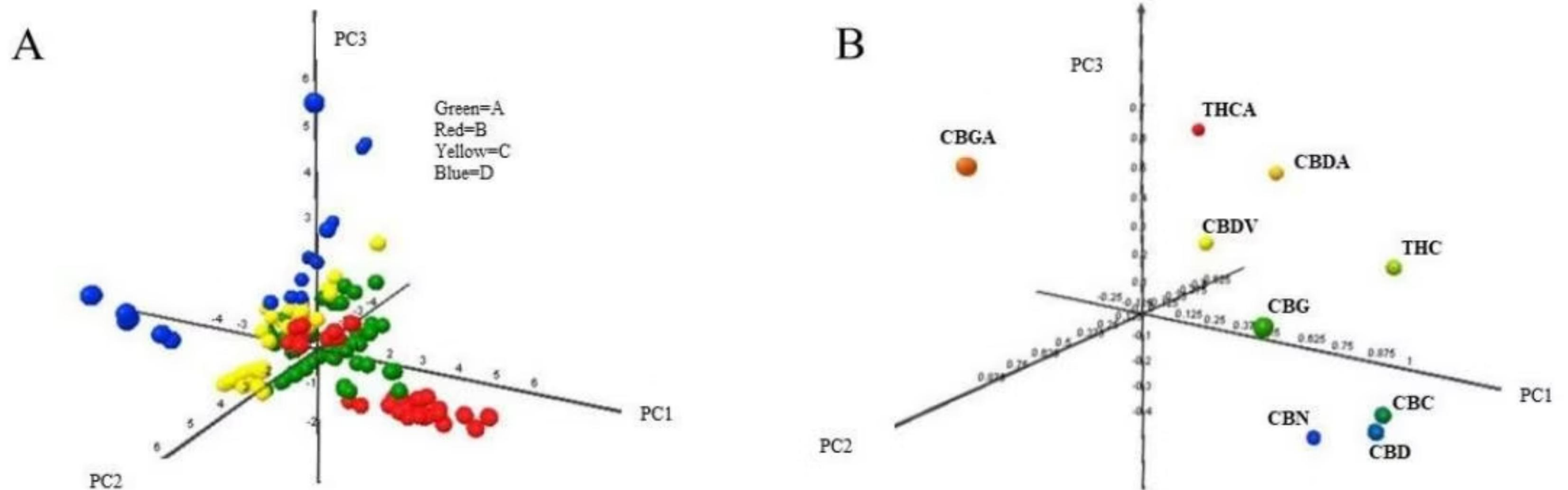Average THC concentration across all retailers remained well below legal thresholds

9

Cannabinoids Analyzed

Six neutral and three acidic cannabinoids measured per sample

161

Total Samples

Hemp samples analyzed from four distinct Italian retailers

Multivariate Analysis Reveals Hidden Patterns

Principal Component Analysis (PCA) revealed that univariate analysis alone could not differentiate retailers due to overlapping cannabinoid compositions. However, PCA scores plot showed partial separation, with PC1 accounting for 42.7% variance, PC2 for 15.8%, and PC3 for 14.9% (total 73.4%).

Cannabinoid Correlations

Positive correlations found between neutral cannabinoids (THC, CBD, CBC). CBDV showed no correlation with other compounds. Weak correlation between CBG and CBGA due to decarboxylation processes.

PLS-DA Performance

Supervised analysis achieved 100% explained variance with only 5% calibration error and 6% cross-validation error using eight model components. This demonstrated excellent discrimination capability.

Classification Accuracy

92% of samples correctly classified in both fitting and cross-validation. Retailers C and D showed highest sensitivity (95-100%), while A and B had some misclassifications to retailer C.

Implications for Forensic Science

Revolutionary Identification Method

This research proves that comprehensive cannabinoid profiling combined with multivariate statistical analysis can successfully identify hemp retailers without requiring information about Cannabis strains, cultivation methods, geographic origin, or storage conditions.

The study reinforces that categorizing hemp using only THC and CBD is inadequate. The synergistic contribution of all nine cannabinoids provides a unique chemical fingerprint for each retailer.

Future Applications

This analytical approach offers a powerful tool for forensic purposes, quality control, and supply chain verification in the rapidly growing hemp industry. The methodology could be extended to other botanical products requiring authentication and traceability.
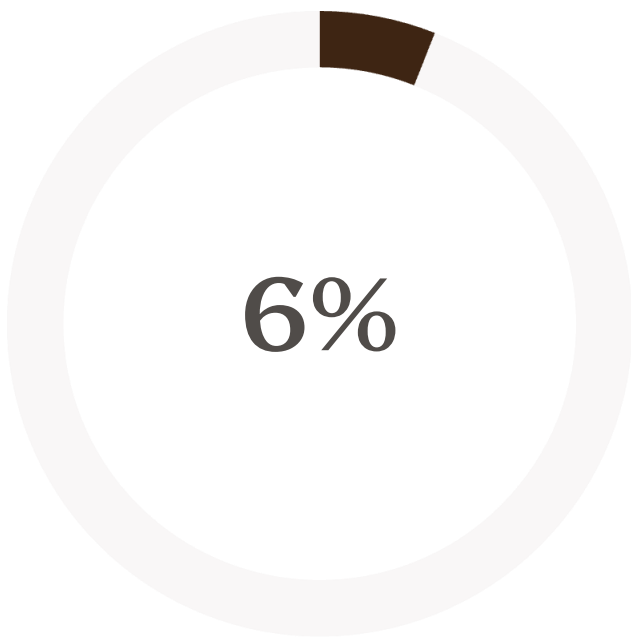
92%
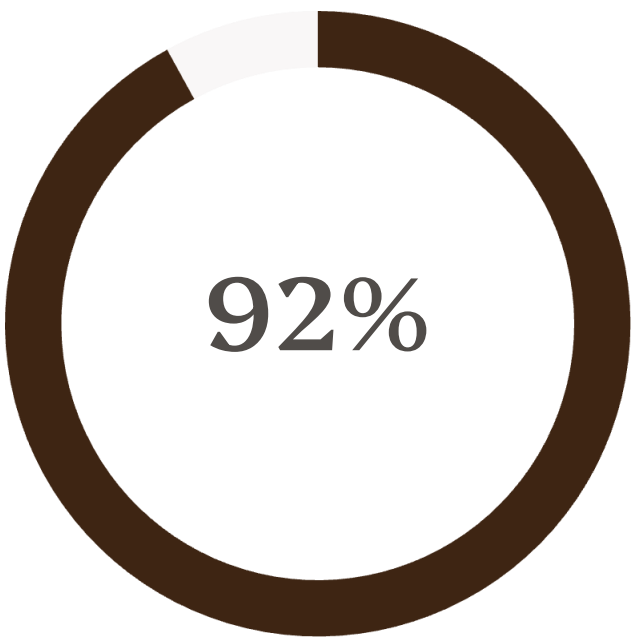
Classification Success

Samples correctly identified by retailer

100%
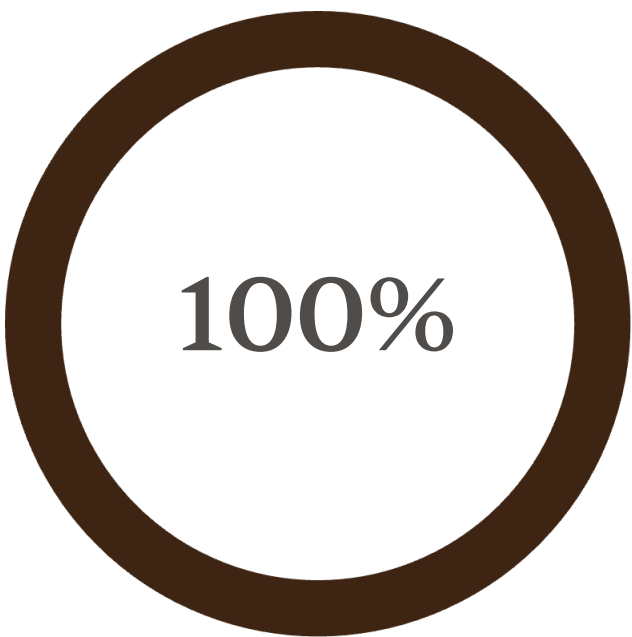
Variance Explained

By PLS-DA model components

6%

Error Rate

Cross-validation classification error