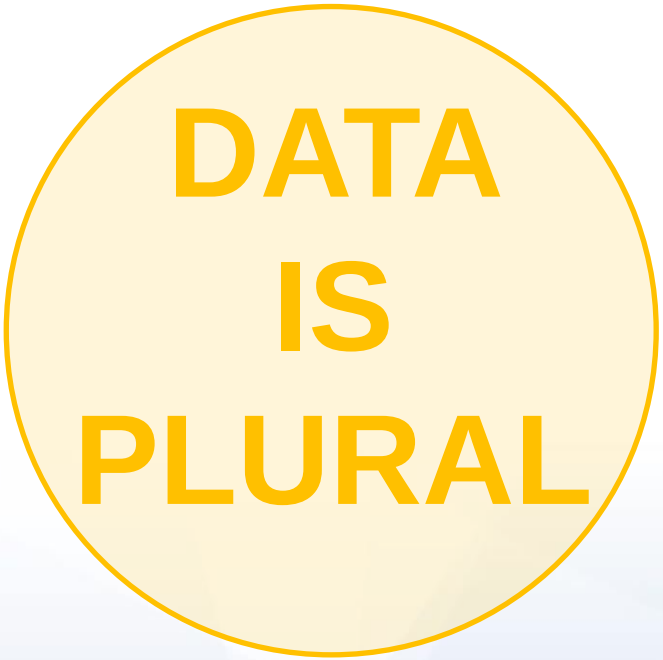
RGÖSSL CONSULTING

Einführung in die Multivariate Datenanalyse

Beispielfolien

Multivariate Datenanalyse
Inhalt

- Einleitung und Zielsetzung
- Deskriptive und grafische multivariate Verfahren
- Multivariate lineare Regressionsmodelle
- Multiresponse Optimierung
- Principal-Component-Analysis (PCA)
- PLS-Analysen
- Multivariate Kalibrierung
- Übungen



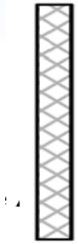
RGÖSSL CONSULTING

- **Multivariate Statistik: n Beobachtungen an $p > 1$ Variablen $\rightarrow (n \times p)$ Matrix $\mathbf{X}$**
- **Systematisierung I**
  - R-Verfahren mit Fokus auf Variablen
    - Funktionale Zusammenhänge und Korrelationen der Variablen
    - Redundanz der Variablen
    - Dimensionsreduktion
  - Q-Verfahren mit Fokus auf Beobachtungen
    - Ähnlichkeiten bzw. Unterscheide der Beobachtungen
    - Gruppierungen (Cluster)
    - Klassifizierungen
- **Systematisierung II**
  - Strukturfindende Verfahren
    - Identifikation und Beschreibung von Datenstrukturen
    - Multivariate deskriptive Statistik
  - Strukturprüfende Verfahren
    - Bewertung der Datenkompatibilität mit vorgegebenen Strukturen (Modellen)
    - Multivariate induktive Statistik

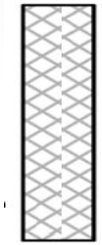
RGÖSSL CONSULTING

- **Datenbasis – $(n \times p)$ Datenmatrix $\mathbf{X}$**

Multivariate Daten
$n < p$



Univariate Daten
$n \gg p = 1$



Bivariate Daten
$n \gg p = 2$

- **Fehlende Beobachtungen**
- **Modellierungen $\mathbf{Y} = \mathbf{f}(\mathbf{X})$**
  - $\mathbf{Y}$-Variable und $\mathbf{X}$-Variable nicht unabhängig (Multikollinearität innerhalb $\mathbf{Y}$ bzw. $\mathbf{X}$)
  - Nur wenige $\mathbf{X}$ erklären $\mathbf{Y}$ („80/20-Regel“ – „Pareto-Prinzip“)
  - Modelle oft von $\mathbf{X}$-Wechselwirkungen dominiert
  - Multiresponse Optimierung

RGÖSSL CONSULTING

- **Multivariate lineare Regression, MANOVA und MANCOVA**
  - (Lineare) Zusammenhänge zwischen $p > 2$ Einflußvariablen und $q > 2$ Responses
  - Simultane Konfidenzintervalle und multiple Testverfahren
  - Regression: Multiresponse-Optimierung → Mehrdimensionale Prozeßfenster
- **Principal Component Analysis (PCA – Hauptkomponentenanalyse)**
  - Orthogonale Transformation der Originalvariablen → Unkorrelierte Principal Components
  - Wenige Komponenten repräsentieren Großteil der Datenvariabiltiät
  - Dimensionsreduktion
- **Clusteranalyse**
  - Konstruktion in sich homogener Gruppen von Beobachtungen (Cluster)
  - $Var_{InnerhalbCluster} < Var_{ZwischenCluster}$
  - Partitionierende Verfahren ↔ Hierarchische Verfahren
- **Diskriminanzanalyse**
  - Klassifizierung von beobachteten und neuen Beobachtungen in definierte Klassen
  - Ableitung und Bewertung von Diskriminanzfunktionen („Klassierungsregeln“)
  - Grundidee vieler „Machine Learning“ Ansätze

- **Zielsetzung analog zum univariaten Fall**
  - Deskriptive Auswertung und Bewertung von beobachtetem Datenmaterial
  - Tabellarische und graphische Aufbereitung
  - Verdichtung der Daten auf charakteristische Größen – Kenn/Maßzahlen
- **Numerische Methoden**
  - Lagemaßzahlen
  - Streuungsmaßzahlen
  - Korrelationsmaßzahlen
- **Grafische Methoden**
  - Multivariate Scatter-Plots und Box-Plots
  - PCA Biplots
  - Multi-Vari-Charts
  - Scaled-Value-Plots

- **Basis: $(n \times p)$ Datenmatrix $\mathbf{X}$**
- **Analyse-Sichtweisen**
  - Spalten (Variablen): Maßzahlen zu Lage, Streuung und Korrelation
  - Zeilen (Beobachtungen): Maßzahlen zu Ähnlichkeit und Abstand
- **Überblick**

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \Rightarrow \textbf{Dist(X)} = \begin{pmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & 0 \end{pmatrix}$$

$$\Downarrow$$

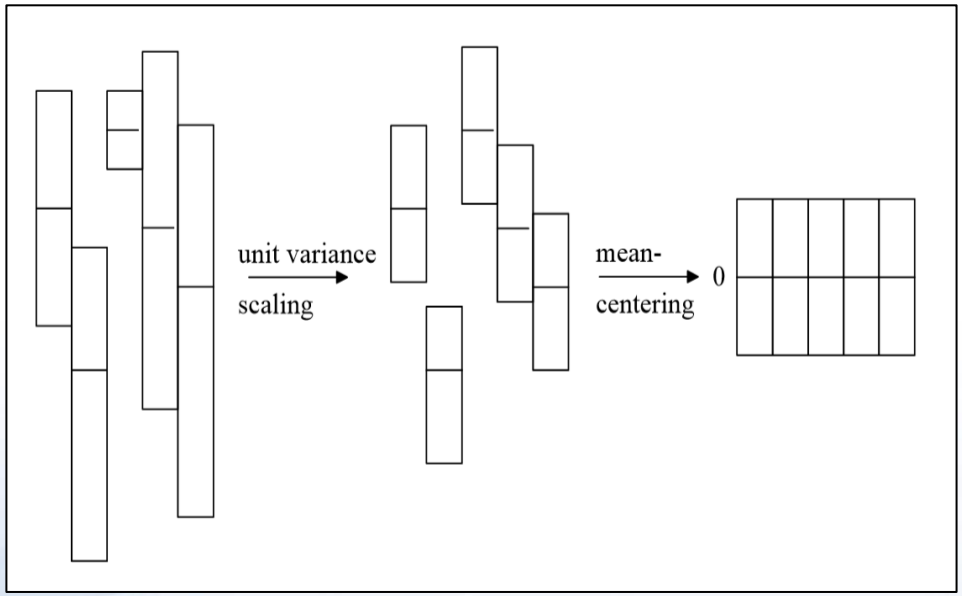
$$\bar{\boldsymbol{x}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p)$$
$$\hat{\boldsymbol{\sigma}} = (\hat{\sigma}_1, \hat{\sigma}_2, \ldots, \hat{\sigma}_p)$$

$$\textbf{Cor(X)} = \begin{pmatrix} 1 & \hat{\rho}_{12} & \cdots & \hat{\rho}_{1p} \\ \hat{\rho}_{21} & 1 & \cdots & \hat{\rho}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\rho}_{p1} & \hat{\rho}_{p2} & \cdots & 1 \end{pmatrix}$$

...

RGÖSSL CONSULTING

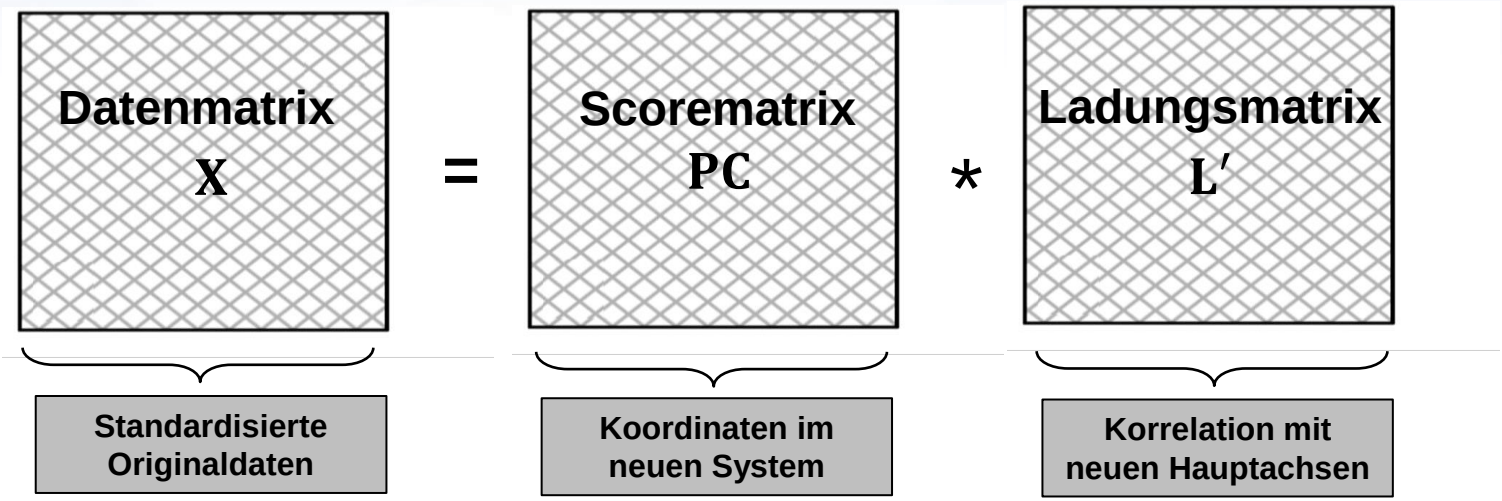
- **Principal-Component-Analysis (PCA) - Hauptkomponentenanalyse**
  - Lineare Koordinatentransformation
  - Dimensionsreduktion
- **Datenbasis: Standardisierte Originaldaten**



Empirische Korrelationsmatrix

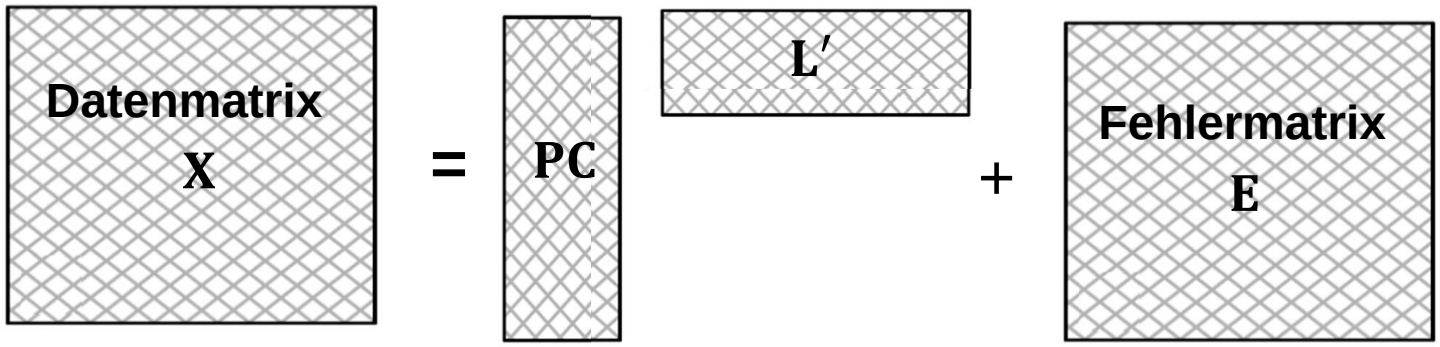
$$\frac{1}{n-1}(X'X)$$

- **Koordinatentransformation formal**



- Originaldaten sind Linearkombinationen der Scores $\boldsymbol{X} = \underbrace{\boldsymbol{PC}}_{\text{„Scores“}} \cdot \overbrace{\boldsymbol{L'}}^{\text{„Loadings“}}$
- Neues Koordinatensystem mit orthogonalen Hauptachsen
- Richtungen der neuen Hauptachsen sind Eigenvektoren von $\frac{1}{n-1}(\boldsymbol{X'X})$

- **Dimensionsreduktion formal**
  - Wenige PC Komponenten erklären Hauptteil der Variabiltät der Originaldaten
  - Einschluß von Fehlermatrix E



„Loadings“

$$X = \underbrace{PC}_{\text{„Scores“}} \cdot \overbrace{L'} + E\} \text{ „Error“}$$

- **Anteil erklärter Variabilität der ersten p Hauptkomponenten**
  - $\sum_{k=1}^{p} \lambda_k \Big/ \sum_{i=1}^{m} \lambda_i$ mit λ_* als Eigenwerten von $\frac{1}{n-1}(\boldsymbol{X'X})$
  - Screeplot

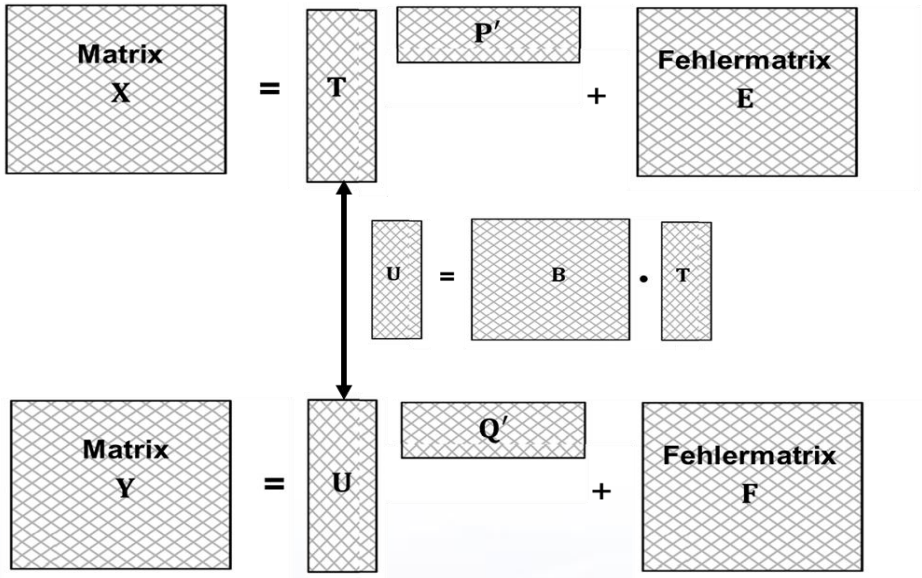
...

RGÖSSL CONSULTING

- **Acronym PLS**
  - Wold (1975): „Partial Least Squares“
  - Wold / Mertens (1984): „Projection to Latent Structures“
- **Erweiterung der PCA-Analyse auf zwei Variablen-Gruppen**
  - Unabhängige Variable $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_p \rightarrow$ Matrix $\mathbf{X}$
  - Abhängige Variable $\mathrm{Y}_1, \mathrm{Y}_2, \ldots, \mathrm{Y}_q \rightarrow$ Matrix $\mathbf{Y}$
- **Anwendungsbeispiele**
  - Quantitative-Structure-Activity-Relationship (QSAR) Modelle
  - Preference Mapping
  - Multivariate statistische Prozeßkontrolle (MSPC)
  - Multivariate Kalibrierung
- **Grundtypen**
  - **PLS1**: MEHRERE unabhängige Variable und EINE abhängige Variable
  - **PLS2**: MEHRERE unabhängige Variable und MEHRERE abhängige Variable

Grundprinzip

- Zerlegung von $\mathbf{X}$ in Scorematrix $\mathbf{T}$, Loadingmatrix $\mathbf{P}$ und Fehlermatrix $\mathbf{E} \rightarrow \mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E}$
- Zerlegung von $\mathbf{Y}$ in Scorematrix $\mathbf{U}$, Loadingmatrix $\mathbf{Q}$ und Fehlermatrix $\mathbf{F} \rightarrow \mathbf{Y} = \mathbf{U}\mathbf{Q}' + \mathbf{F}$



- Verbindung beider Informationen durch Zusammenhang der Scores $\mathbf{U} = \mathbf{B} \cdot \mathbf{T}$
- Bestimmung von $\mathbf{B}$ unter Nebenbedingungen:
  - Fehlermatrix $\mathbf{F} \rightarrow$ Min!
  - Korrelation $\mathrm{Cor}(\mathbf{X}, \mathbf{Y}) \rightarrow$ Max!

RGÖSSL CONSULTING

- **Multivariate Analyse Konsumentenbefragung Lebensmittelindustrie**

Research Article

SCI

Received: 23 April 2012 Revised: 4 September 2012 Accepted article published: 25 September 2012 Published online in Wiley Online Library:

(wileyonlinelibrary.com) DOI 10.1002/jsfa.5932

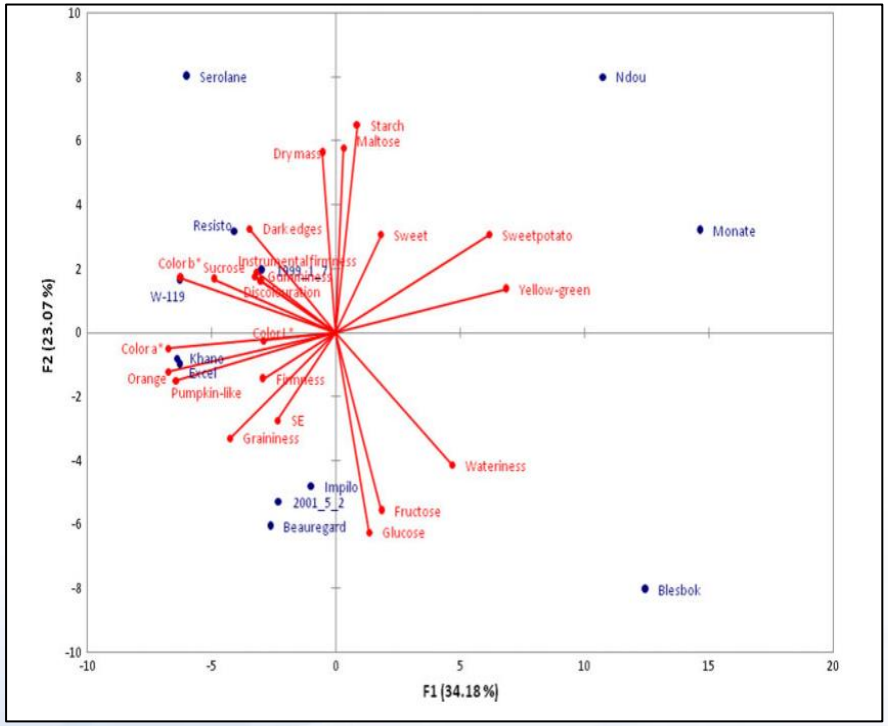
The use of sensory attributes, sugar content, instrumental data and consumer acceptability in selection of sweet potato varieties

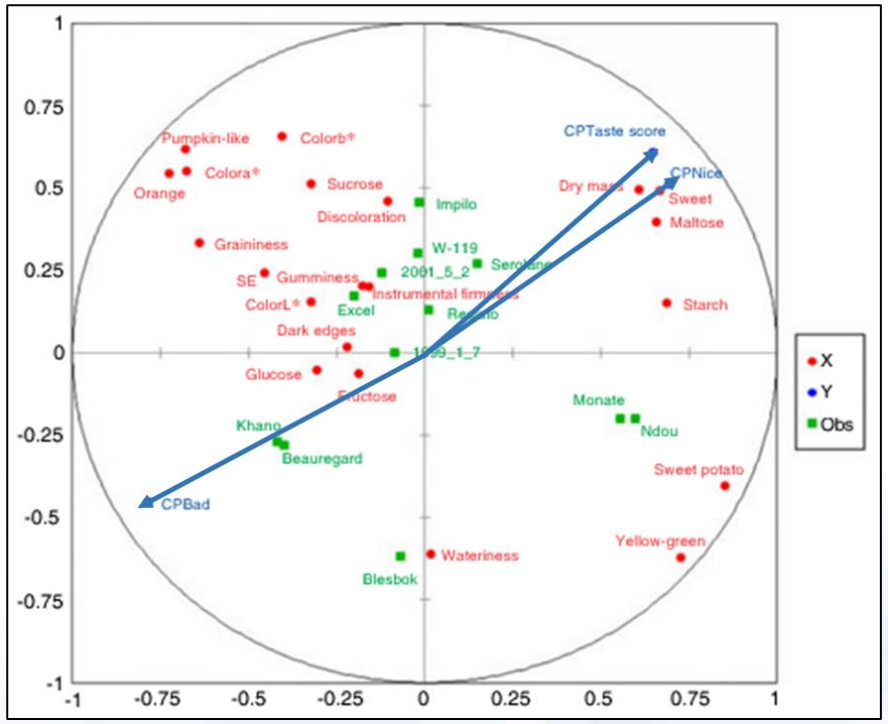
Sunette M Laurie,[a*] Mieke Faber,[b] Frikkie J Calitz,[c] Erika I Moelich,[d] Nina Muller[d] and Maryke T Labuschagne[e]

- **Datenbasis**
  - Chemisch-Physikalische Eigenschaften
  - Sensorische Daten
  - Konsumenten-Akzeptanz
- **Zielsetzung**
  - Multivariate Analyse Konsumenten-Akzeptanz für 12 Süßkartoffel-Varianten
  - Zusammenhang chemisch-physikalische Eigenschaften ↔ Sensorische Daten
  - Zusammenhang chemisch-physikalische und sensorische Daten ↔ Akzeptanz

RGÖSSL CONSULTING

- **Interpretieren Sie den PCA- und den PLS Biplot der Süßkartoffel-Varianten!**
  - Wie hängen chemisch-physikalische Eigenschaften und sensorische Daten zusammen?
  - Welche chemisch-physikalische und sensorische Daten beeinflußen Akzeptanz?



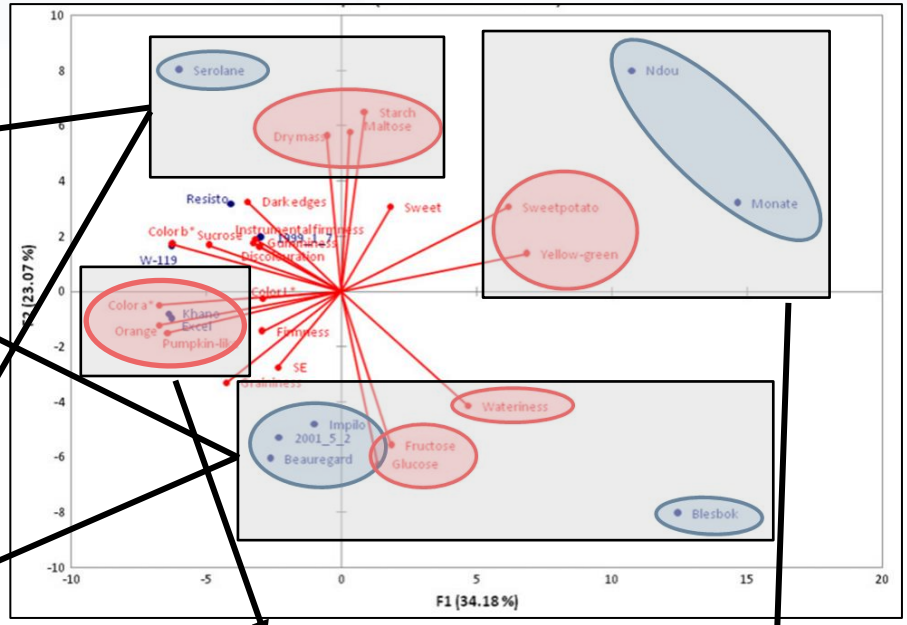


- **Verifizieren Sie diese Zusammenhänge anhand der Originaldaten!**

RGÖSSL CONSULTING

- **Chemisch-physikalische Eigenschaften ↔ Sensorische Daten**

Variety	Fructose	Glucose	Wateriness
2001_5_2	1.2	1.3	17.6
Impilo	1.4	1.6	19.8
1999_1_7	0.4	0.5	11.8
Beauregard	1.1	1.2	18.9
Blesbok	1.4	1.4	23.3
Excel	0.7	0.7	10.1
Khano	0.5	0.5	18.2
Monate	0.6	0.6	18.5
Ndou	0.8	0.6	18.4
Resisto	0.6	0.5	9.6
Serolane	0.5	ND	7.7
W-119	0.8	0.9	7.2



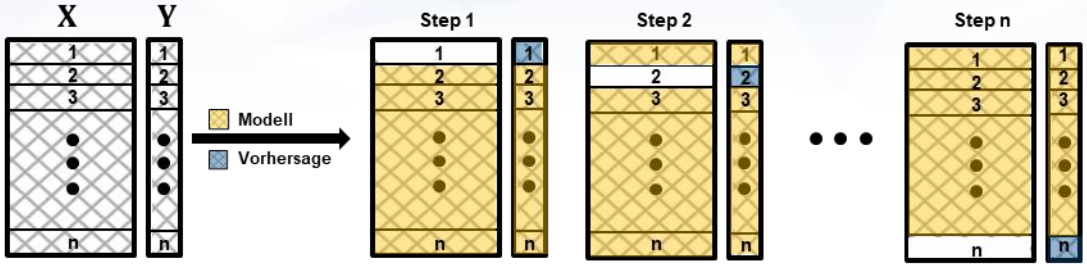
Variety	Dry mass(%)	Maltose	Starch(%)
2001_5_2	19.7	7.3	4.4
Impilo	19.4	7.7	4.6
1999_1_7	19.3	6.5	6.7
Beauregard	18.7	5.1	4.6
Blesbok	15.9	5.2	4.8
Excel	20.3	7.2	5.2
Khano	16.8	6.4	5.0
Monate	20.2	8.1	7.2
Ndou	24.6	10.1	9.8
Resisto	22.6	7.3	8.1
Serolane	22.4	11.5	8.6
W-119	23.1	6.7	8.8

Variety	Yellow-green	Sweet potato flavour	Pumpkin flavour	Orange	Colour a^*
2001_5_2	0.0	39.0	23.3	82.8	27.8
Impilo	0.3	41.2	27.9	58.9	20.5
1999_1_7	0.0	43.4	22.4	66.3	26.4
Beauregard	0.0	35.4	23.9	74.3	26.3
Blesbok	58.6	51.0	0.2	0.0	−6.9
Excel	0.0	25.8	43.1	58.9	18.1
Khano	0.0	35.2	25.6	82.9	30.6
Monate	58.4	72.0	0.2	0.0	−4.3
Ndou	75.6	76.4	0.1	0.0	−1.3
Resisto	0.0	35.3	24.7	81.6	29.8
Serolane	0.2	48.6	16.5	55.7	21.7
W-119	0.0	27.9	29.9	79.9	25.2

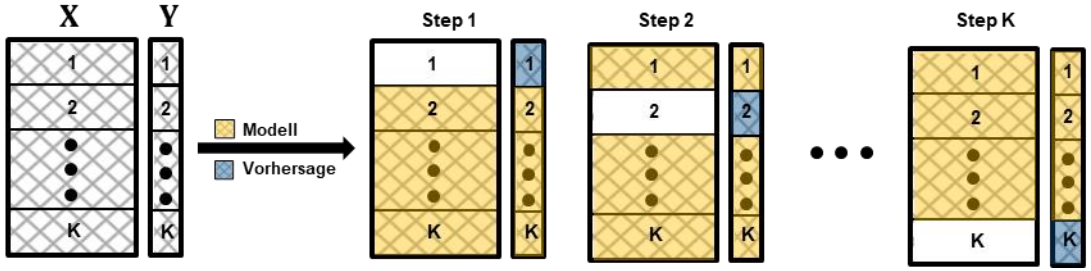
- **Multivariate Erweiterung der Inversen Regression**
  - Kalibrationsmodell mit bekannten unabhängigen und abhängigen Variablen $\mathbf{X}$ bzw. $\mathbf{Y}$
  - Matrix $\mathbf{X}$: Oftmals Spektren – NIR, HPLC, …
  - Matrix $\mathbf{Y}$: Response-Charakteristika – Konzentration, Konsumentenakzeptanz, …
  - Ziel: Vorhersage $\hat{\mathbf{Y}}_*$ der Response-Charakteristika für neue Spektren $\mathbf{X}_*$
- **Anwendungsbeispiele**
  - NIR Messungen ↔ Oktanzahl Benzin
  - GC oder HPLC Messungen ↔ Geschmack und Aroma von Whiskey oder Wein
  - HPLC oder NIR Messungen ↔ Komponenten und Metabolite bei Arzneimitteln
  - UV Messungen ↔ Huminsäuren und Sulfonate in Gewässern
- **Basis: PLS Regressionsmodelle (PLSR)**
  - $\mathbf{Y} = \mathbf{X}\mathbf{B}_{\mathbf{PLS}} + \mathbf{F}$
  - Iterative Schätzung Parametermatrix $\hat{\mathbf{B}}_{\mathbf{PLS}}$ über Zerlegungen $\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E}$ und $\mathbf{Y} = \mathbf{U}\mathbf{Q}' + \mathbf{F}$
  - NIPALS (Nonlinear Iterative Partial Least Squares) Algorithmus

RGÖSSL CONSULTING

- **LOO Full Cross-Validation**



- **K-Fold Cross-Validation**



- **Simulation Cross-Validation**