

An L-PLS Preference Cluster Analysis on French Consumer Hedonics to Fresh Tomatoes

D. Plaehn, G. Stucky and D. Lundahl

Sensometrics 2004

July 29, 2004

The Data

- Data from Centre Technique des Fruits et Legumes and the Institut National de la Recherche Agronomique (France) was provided by Pascal Schlich
- The data referred to 17 tomato varieties and included
 - Consumer Hedonic Ratings
 - Sensory attributes
 - Physical & chemical measures
 - Usage and Attitude (U&A) information
- 379 Respondents (4 dropped due to extreme incompleteness of U&A responses)

Overview

- Segmentation Analysis of Overall Liking Response Data
 - Treatment of Incomplete Data
 - Clustering Algorithms
 - Selecting Optimal Cluster Number
- L-Partial Least Squares (L-PLS) Analysis¹
 - Aggregating Data Over Clusters
 - Variable Selection and L-PLS Model Building
 - Mapping of Respondent Data into Aggregate L-PLS Solution
 - 95% Confidence Ellipsoids of Individual Scores

¹Martens, H. et al, (2003) Regression of a data matrix on descriptors of both its rows and of its columns via latent variables: L-PLSR. Computational Statistics and Data Analysis. In press.

Segmentation Analysis

Treatment of Incomplete Data

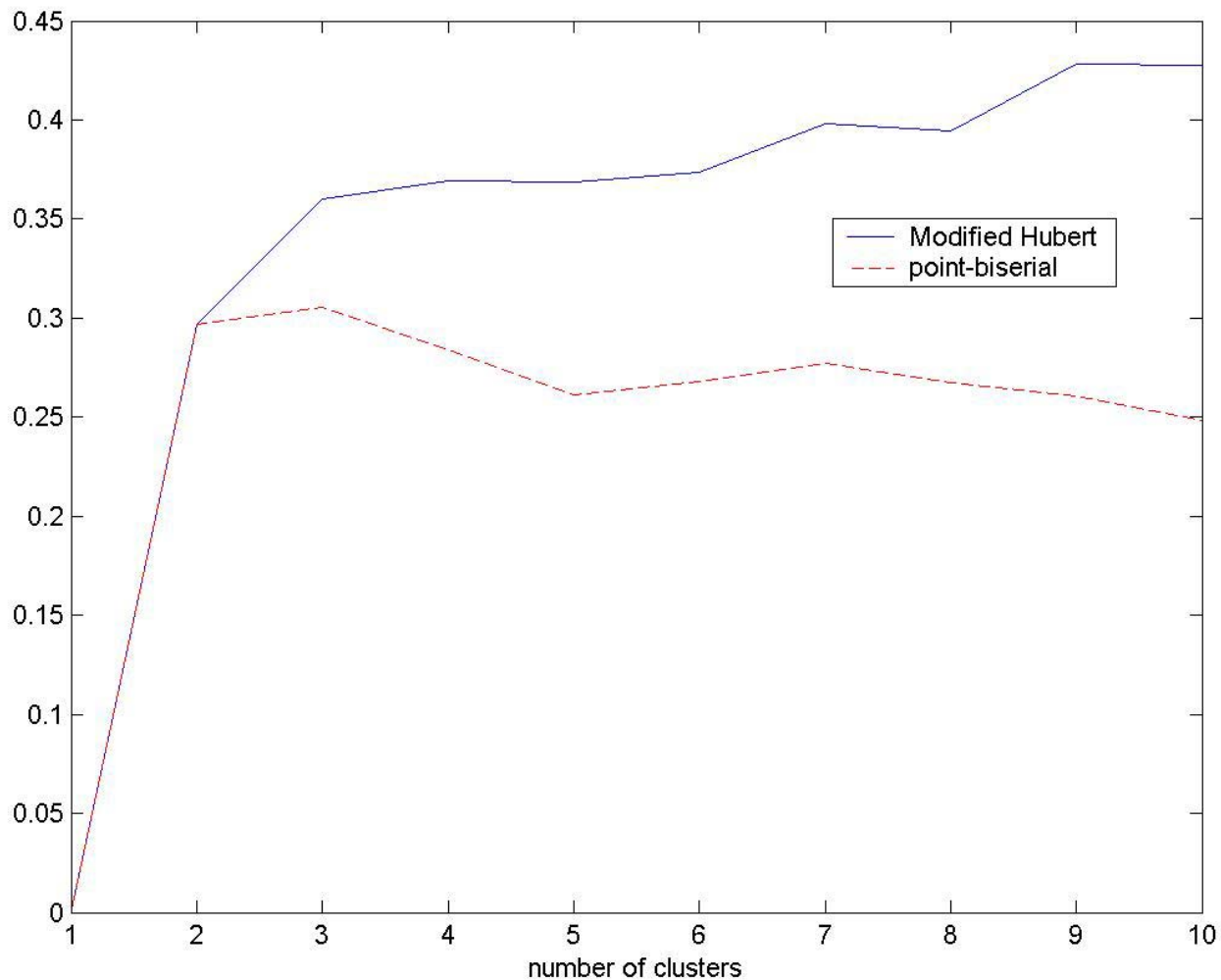
- Incomplete Block Design - 41% of the Responses on Overall Liking of Tomatoes are Incomplete
- Iterative PCA Approach to Estimate Incomplete Data
 1. Initial estimate of incomplete values (marginal means)
 2. Run PCA on complete + incomplete data estimates (estimate PCA loadings and optimal PC number based upon cross-validation)
 3. Replace incomplete data with PCA estimates ($X = T \text{ scores} \times P^T \text{ loadings}$).
 4. Go to 2 and repeat until loadings “stabilize” and optimal PC number does not change.

Segmentation Analysis

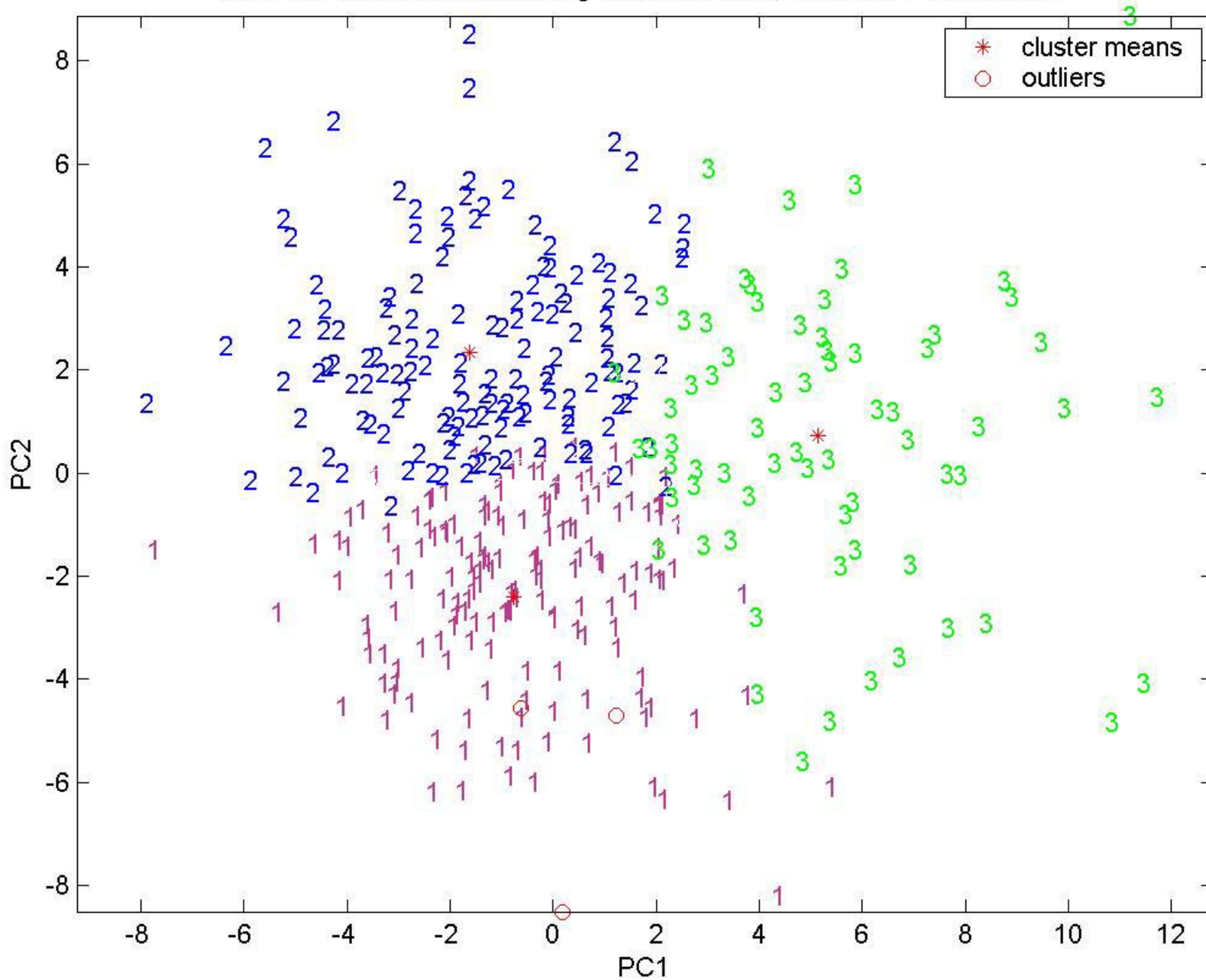
Clustering Algorithm

- For a given number of clusters, g , first the k-means solution is obtained in the form of a classification,, and cluster means (weights), W .
- The k-means solution is improved through an iterative algorithm that maximizes the Calinski-Harabasz (CH) index for a given g .
 - $CH(g) = \text{trace}(\mathbf{B}) / (g-1) / \text{trace}(\mathbf{L}) / (n-g)$ where n is the number of data points. The total variation, \mathbf{T} , of the data is the sum of the between-cluster variation, \mathbf{B} , and the within-cluster variation, \mathbf{L} .
 - By adding or subtracting points from the various k-means clusters the routine tries to increase the CH index.
 - Note: K-means tries to minimize \mathbf{L} .

Optimal Number of Clusters



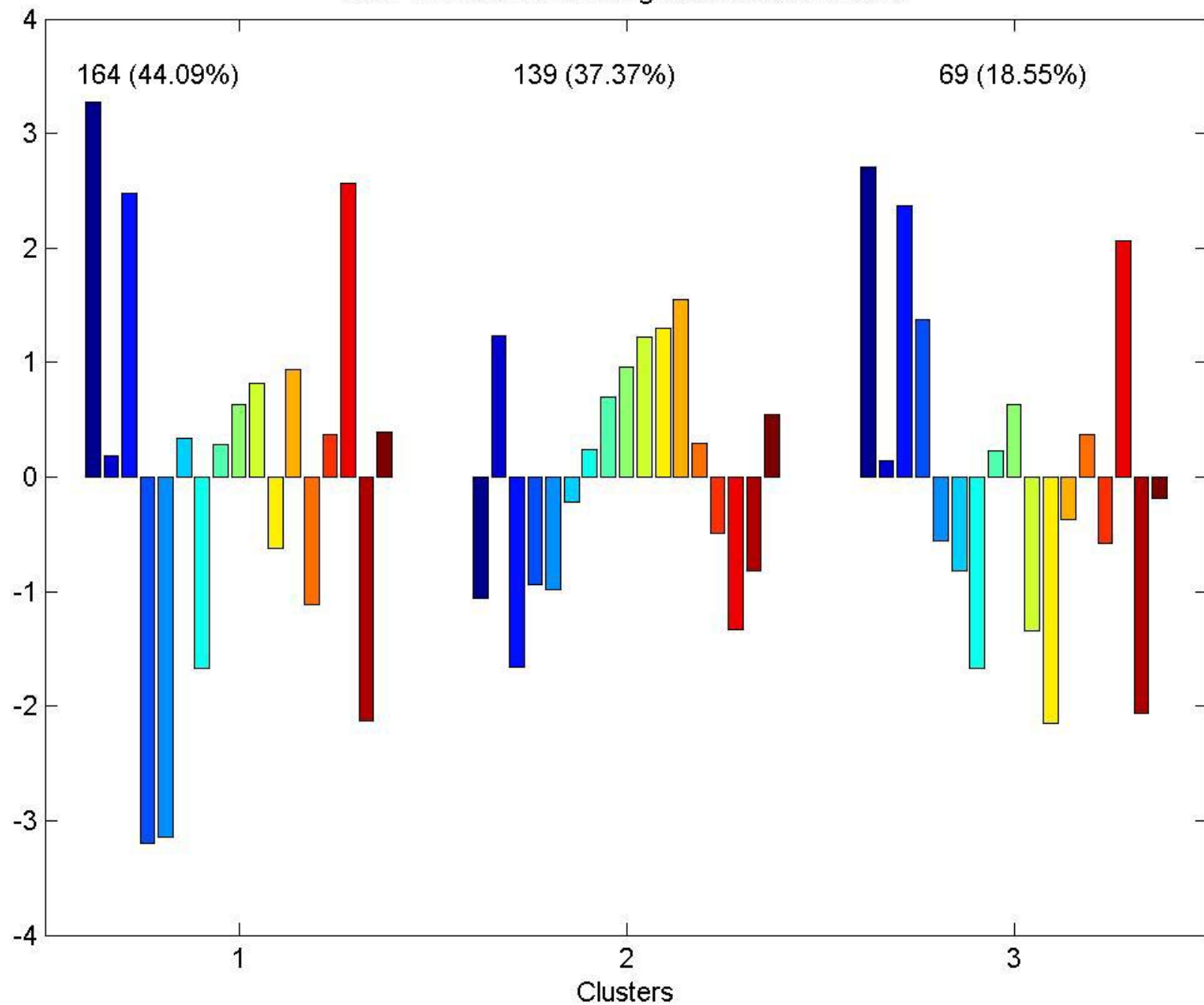
2001 Tomato Filled RMC Liking Data: 3 Cluster, K-means + WO solution



Model Data Preprocessing

- Usage & Attitude (Z) Data
 - Multiple Nominal and Ordinal Category Ratings Converted to Nominal Data (0,1)
 - Averaged Over Segments, Variable Centering and Scaling
- Descriptive, Physical & Chemical (X) Data
 - Variable Centering and Scaling
- The Liking (Y) Data
 - Average Over Segments
 - Double-Centering, Not Scaling

2001 Tomato RMC Liking Data: Cluster Means

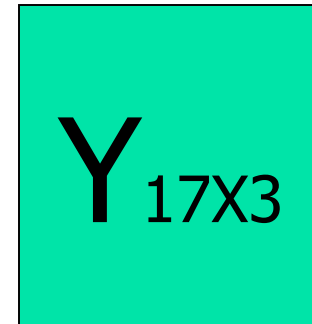
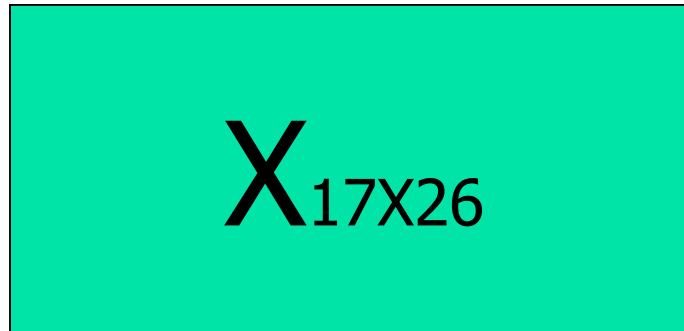


Aggregate Data

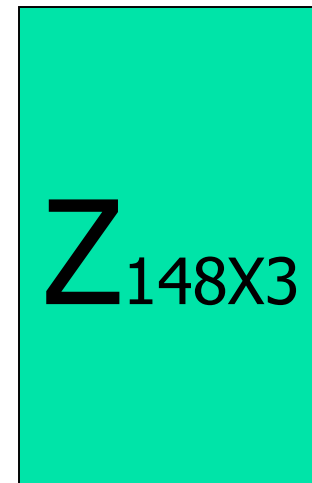
Descriptive and Analytical

Clusters

Tomatoes

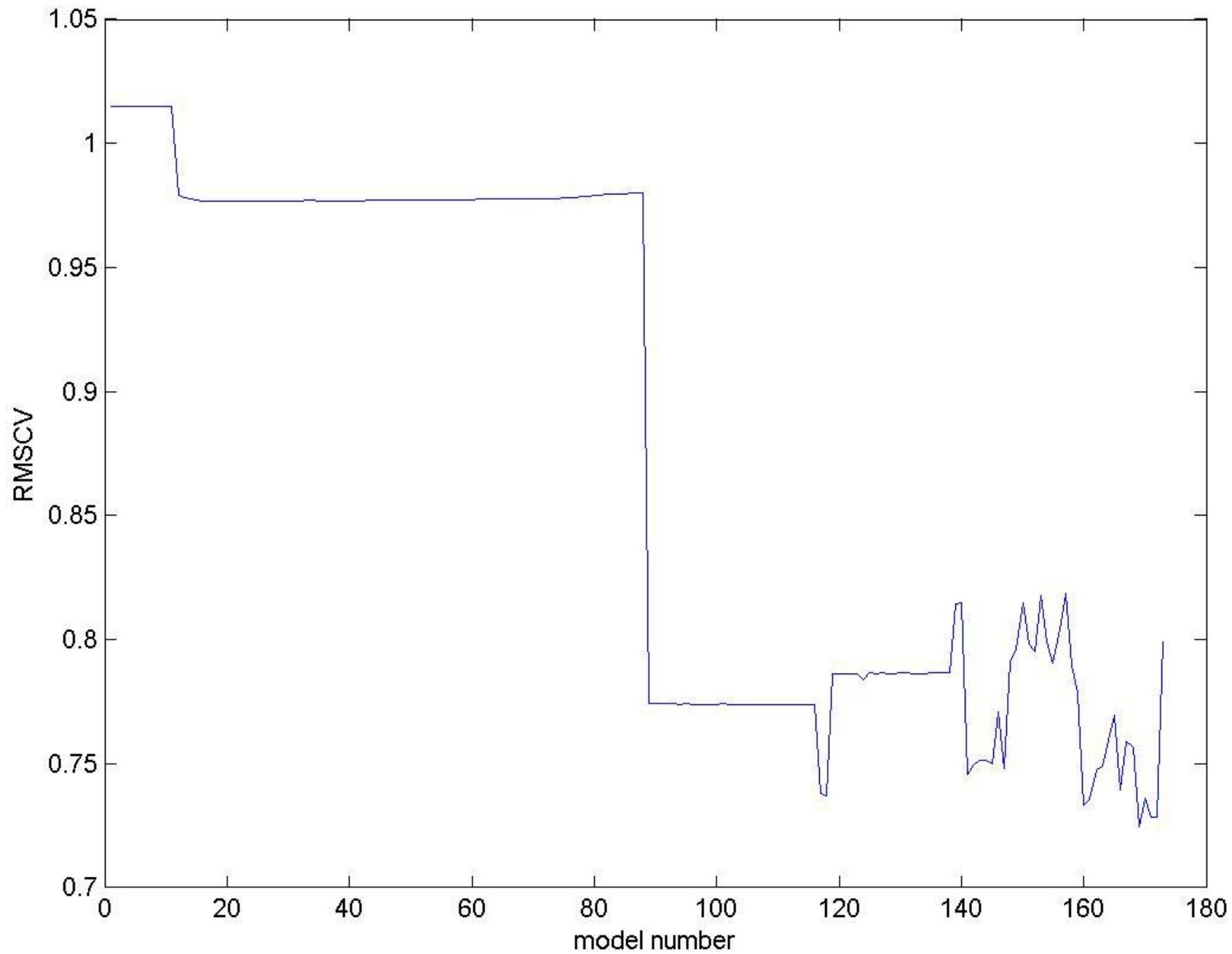


U&A/Demo



L-PLS Model: Variable Selection

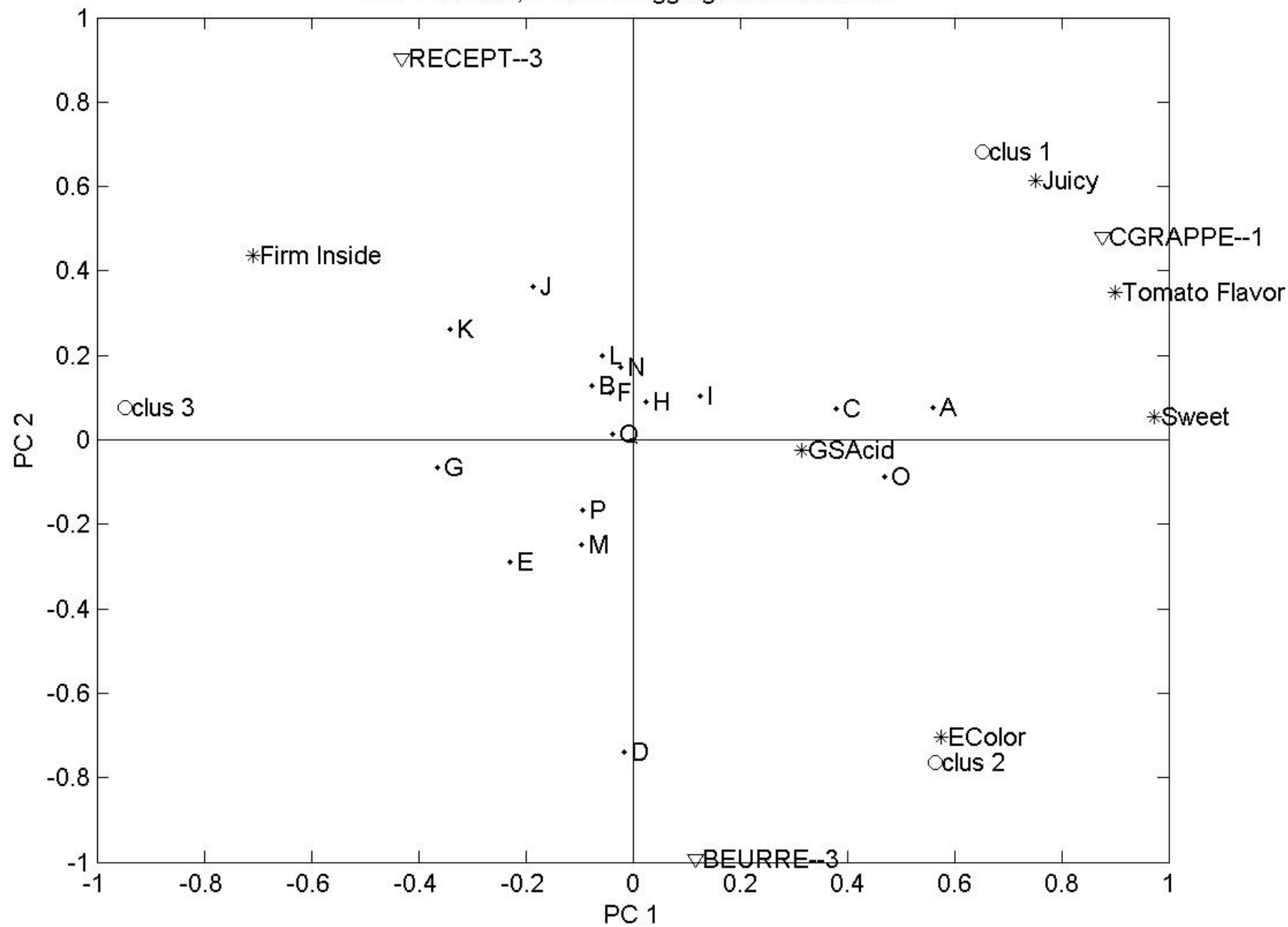
- **Significant variables (interactions) were determined using a variable selection process.**
 - **Initial Ranking:** Full cross-validation and estimation of validation error for the 26 (X variables) by 148 (Z variables) interactions.
 - **Forward Stepwise Addition:** Build rough models by progressively adding in the X and Z variables in the top interactions and select the best of these based on RMSCV (cross-validation error)
 - **Backward Stepwise Deletion:** Alternately, back out terms to improve the model (reduce RMSCV).



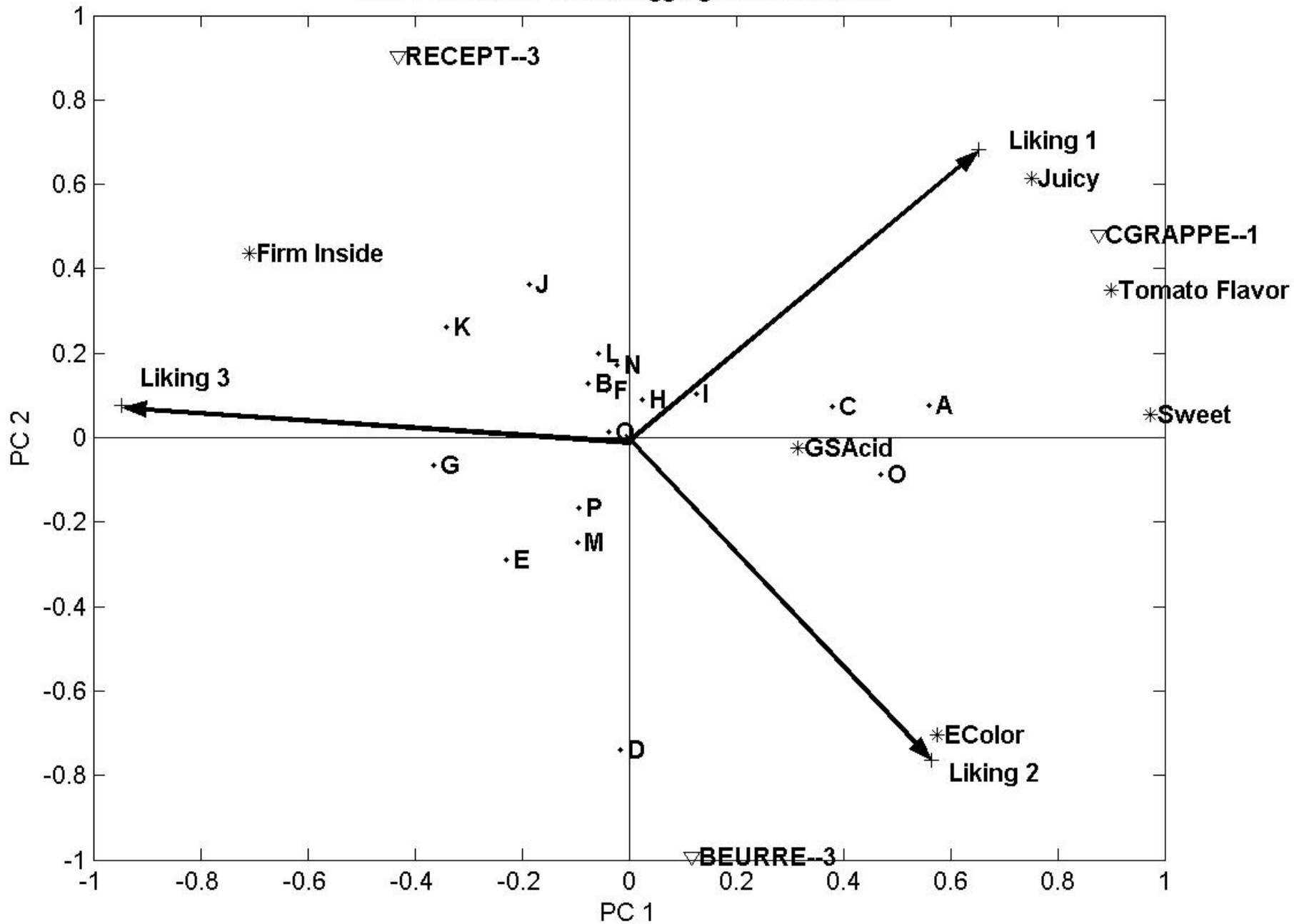
L-PLS Results: Final Model

- The Final L-PLS Model Variables
 - X-variables
 - Sensory (4): Firm Inside, Juicy, Sweet, Tomato Flavor
 - Physical & Chemical (2): Ecolor, GSAcid
 - Z-variables (UA) (3): RECEPT (level 3), BEURRE (level 3), CGRAPPE (level 1).
- Three X PCs, two Z PCs
- RMSEC = 0.3427 and RMSCV = 0.6272
- $R^2_{val} = 90.7\%$ and $R^2_{val} = 68.9\%$

2001 Tomato, 3 Cluster Aggregate LPLS Model



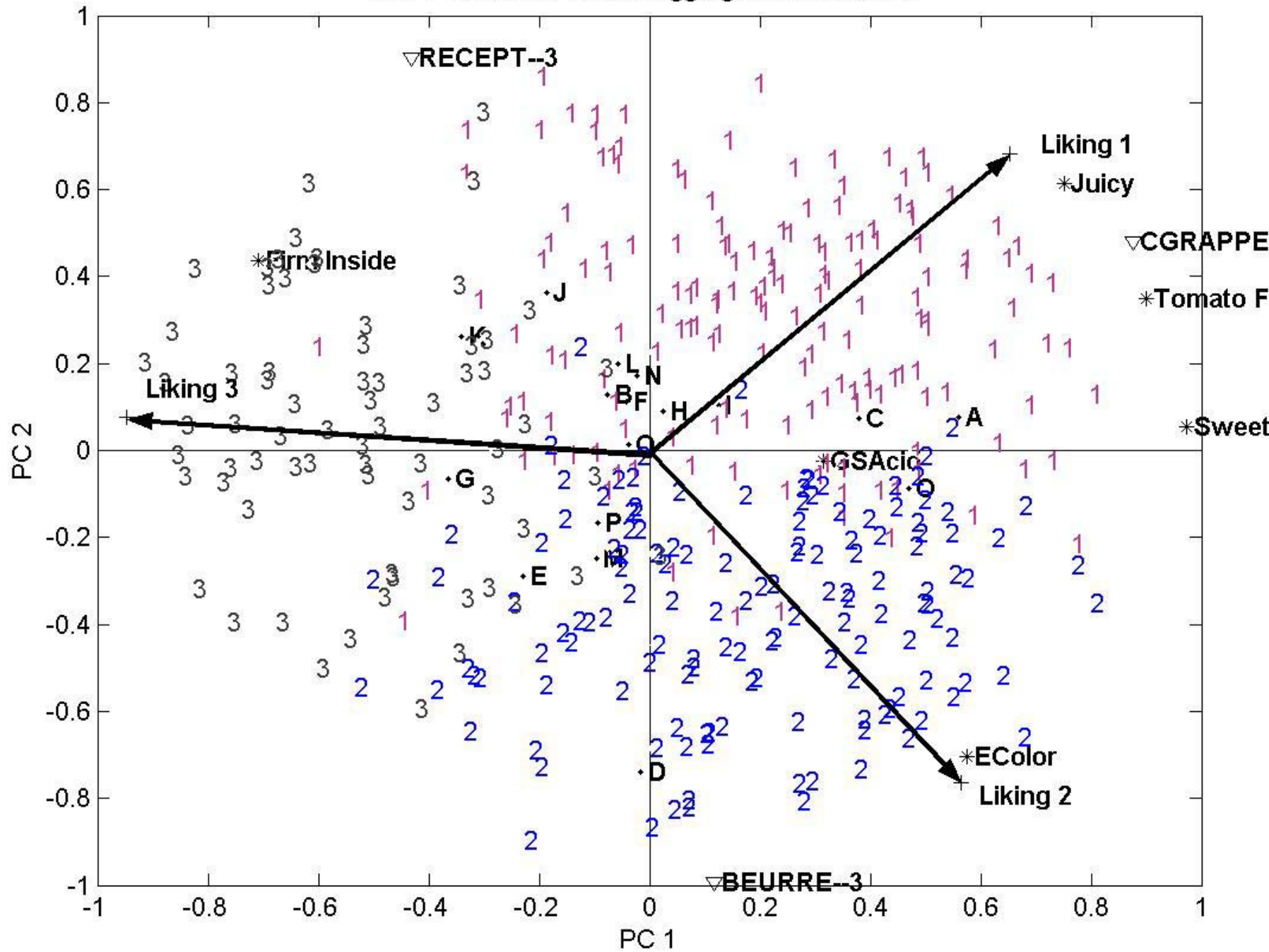
2001 Tomato: 3 Cluster Aggregate LPLS Model



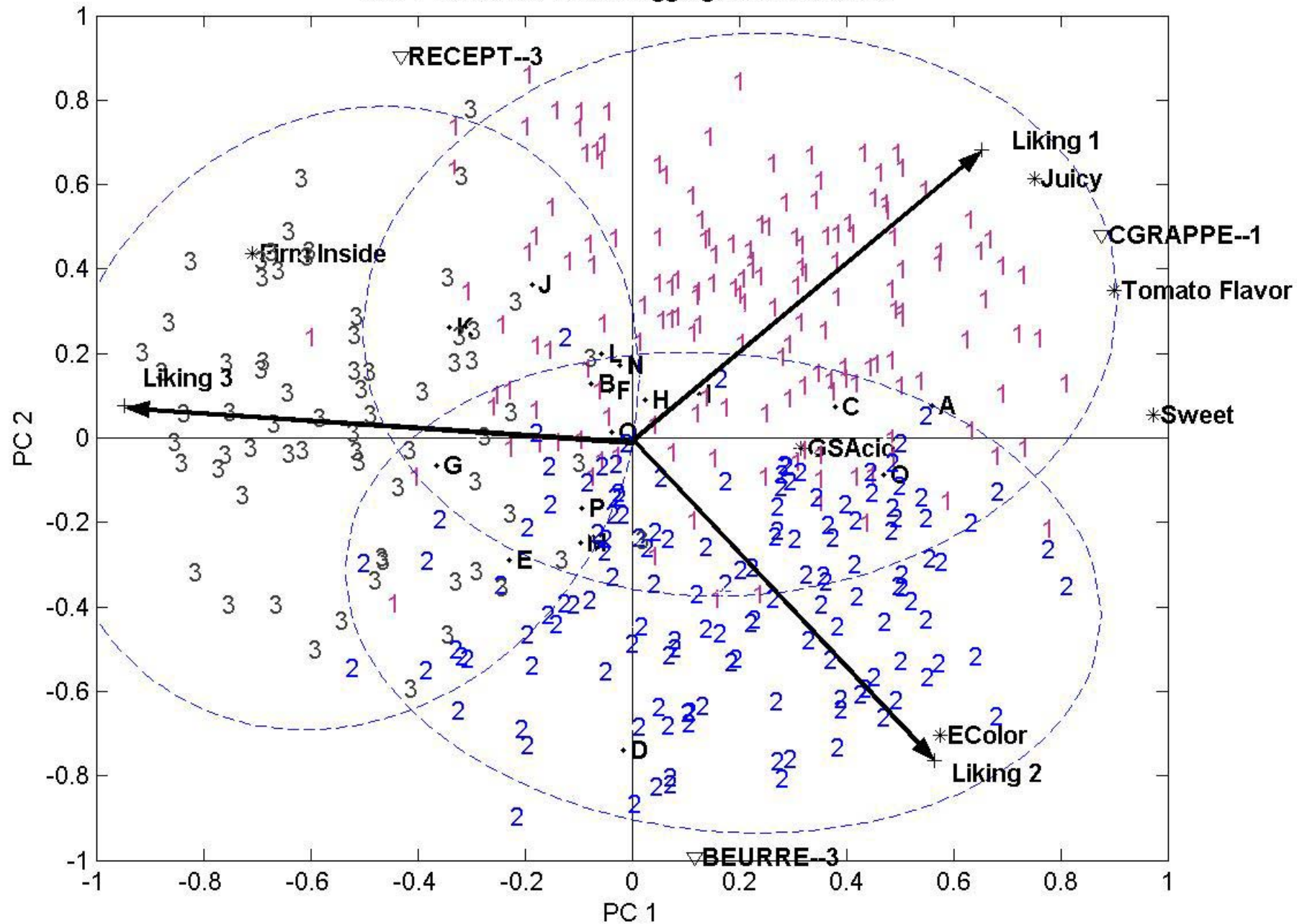
Mapping Individual Responses into L-PLS Aggregate Model

- Estimate Correlations between Aggregate L-PLS Model X Scores for Tomatoes and Individual Respondent Hedonic Ratings on Tomatoes
- Plot Correlations (Score) into L-PLS Aggregate Model Correlation Loading Maps

2001 Tomato: 3 Cluster Aggregate LPLS Model



2001 Tomato: 3 Cluster Aggregate LPLS Model



Analysis Summary

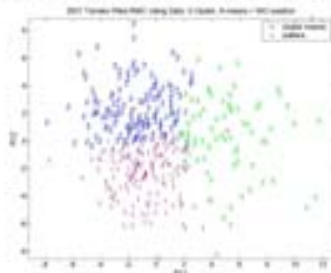
2001 Data	Segment 1	Segment 2	Segment 3
Size of Segment (%)	44.09	37.37	18.55
Tomato varieties liked most	A, C, O	A, D, O	B, J, K, L
Tomato varieties liked least	D, E, P	G, K, P	C, O
Positive drivers of liking (sensory, physical/chemical)	Juicy, Tomato Flavor	EColor	Firm Inside
Negative drivers of liking (sensory, physical/chemical)	Firm Inside	Firm Inside	Sweet, Tomato Flavor
Quadratic (curvilinear) drivers of liking (sensory, physical/chemical)	NA	NA	NA
Key demographic attitudinal, and usage characteristics	CGRAPPE (level 1) (positive)	BEURRE (level 3) (positive)	RECEPT (level 3) (positive)

An L-PLS Preference Cluster Analysis on French Consumer Hedonics to Fresh Tomatoes

D. Plaehn, G. Stucky and D. Lundahl, InsightsNow, Inc. Corvallis, OR USA

The Data

- Data from Centre Technique des Fruits et Légumes and the Institut National de la Recherche Agronomique (France) was provided by Pascal Schlich
- The data referred to 17 tomato varieties and included
 - Consumer Hedonic Ratings
 - Sensory attributes
 - Physical & chemical measures
 - Usage and Attitude (U&A) information
- 379 Respondents (4 dropped due to extreme incompleteness of U&A responses)



Segmentation Analysis

Treatment of Incomplete Data

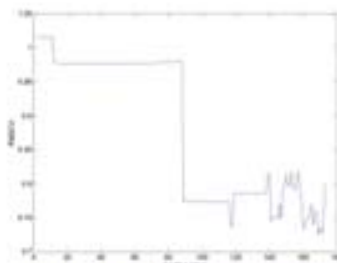
- Incomplete Block Design - 41% of the Responses on Overall Likings of Tomatoes are Incomplete
- Iterative PCA Approach to Estimate Incomplete Data
 - Initial estimate of incomplete values (marginal means)
 - Run PCA on complete + incomplete data estimate (estimate PCA loadings and optimal PC number based upon cross-validation)
 - Replace incomplete data with PCA estimate (X-Tensors = P-loadings)
 - Go to 2 and repeat until loadings "stabilize" and optimal PC number does not change

Model Data Preprocessing

- Usage & Attitude (Z) Data
 - Multiple Nominal and Ordinal Category Ratings Converted to Nominal Data (0, 1)
 - Averaged Over Segments, Variable Centering and Scaling
- Descriptive, Physical & Chemical (X) Data
 - Variable Centering and Scaling
- The Likings (Y) Data
 - Average Over Segments
 - Double-Centering, Not Scaling

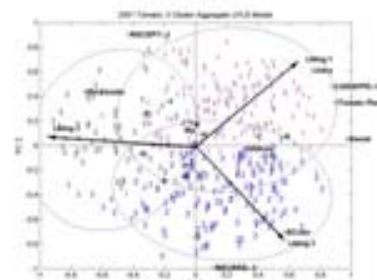
L-PLS Model: Variable Selection

- Significant variables (in interaction) were determined using a variable selection process.
 - Initial Ranking:** Full cross-validation and estimation of validation error for the 26 (X variables) by 148 (Z variables) interactions.
 - Forward Stepwise Addition:** Build rough model by progressively adding in the X and Z variables in the top interactions and select the best of these based on RMSECV (cross-validated error)
 - Backward Stepwise Deletion:** Alternatively, build out terms to improve the model (reduce RMSECV).



Mapping Individual Responses into L-PLS Aggregate Model

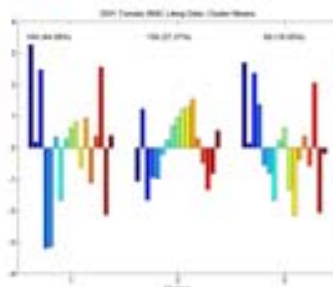
- Estimate Correlations between Aggregate L-PLS Model X Score for Tomatoes and Individual Respondent Hedonic Ratings on Tomatoes
- Plot Correlation (Score) into L-PLS Aggregate Model Correlation Loading Maps



Segmentation Analysis

Clustering Algorithm

- For a given number of clusters, e.g., find the best one solution is a function of the size of a cross-validated set, and cluster error (avg RMSECV)
- The best one solution is found through an iterative algorithm that estimates the optimal number of clusters for a given set
 - 1) Run PCA on complete + incomplete data estimate (estimate PCA loadings and optimal PC number based upon cross-validation)
 - 2) Replace incomplete data with PCA estimate (X-Tensors = P-loadings)
 - 3) Go to 2 and repeat until loadings "stabilize" and optimal PC number does not change



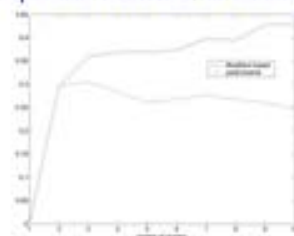
L-PLS Results: Final Model

- The Final L-PLS Model Variables
 - X-variables
 - Texture (1); Flavor Intensity, Juicy, Sweet, Total of 3 Flavor - Physical & Chemical (2) Total: 204.34
 - Z-variables (U&A) (3): RECEIPT (level 3), BELIEFE (level 3), CGRAPPE (level 1)
- Three X PCs, two Z PCs
- RMSECV = 0.3427 and RMSECV = 0.6272
- R²val = 90.7% and R²val = 68.9%

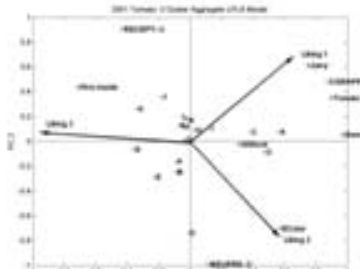
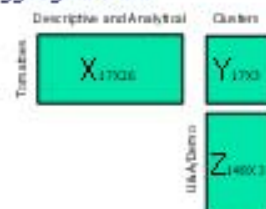
Analysis Summary

Variable	Segment 1	Segment 2	Segment 3	Segment 4
Overall Hedonic Rating	0.10	0.15	0.20	0.25
Texture (1)	0.10	0.15	0.20	0.25
Flavor Intensity	0.10	0.15	0.20	0.25
Juicy	0.10	0.15	0.20	0.25
Sweet	0.10	0.15	0.20	0.25
Total of 3 Flavor - Physical & Chemical	0.10	0.15	0.20	0.25
RECEIPT (level 3)	0.10	0.15	0.20	0.25
BELIEFE (level 3)	0.10	0.15	0.20	0.25
CGRAPPE (level 1)	0.10	0.15	0.20	0.25

Optimal Number of Clusters



Aggregate Data



References

- R.C. Dubois, "How many clusters are best?—an experiment," *Personality and Individual Differences*, 20, 645-663, 1997 (Medford/Dubois)
- J.J. Arbuthnot, *Applied Multivariate Data Analysis, Volume 1: Geometric and Multivariate Methods*, New York: Springer, 1992, pp. 550-551, 550m-basrath
- Harald Martens et al, "Regression of data matrix on descriptors of both in row and of its column variables: variable L-PLS," preprint (accepted *Geostatistical Science and Data Analysis*), 2004