Latent Class (Finite Mixture) Segments
How to find them and what to do with them

Jay Magidson
Statistical Innovations Inc.
Belmont, MA USA
www.statisticalinnovations.com

Sensometrics 2010, Rotterdam

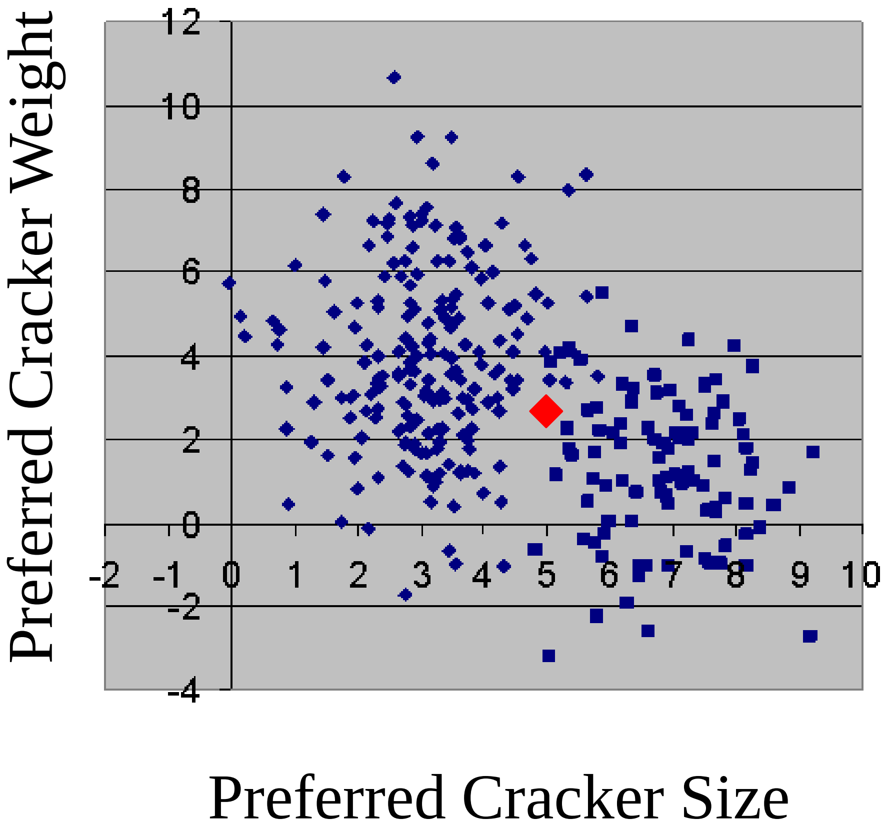
Overview of Presentation

- Graphical Introduction
- Kellogg Data Case study: cracker taste test
  - ➢ LC Cluster Models – Identify segments with different sensory preferences
  - ➢ LC Regression Models – Simultaneously segment and estimate effects of product attributes for each segment
- For each segment determine the relevant attributes and attribute interactions from possibly hundreds, with small sample size (brief discussion as time permits):
  - ➢ Penalty/regularization methods
  - ➢ PLS Regression
  - ➢ Correlated Component Regression (CCR) – New (Magidson, 2010a, 2010b)

Overview of Presentation

- **Graphical Introduction**
- Kellogg Data Case study: cracker taste test
    - ➢ LC Cluster Models
    - ➢ LC Regression Models – Segmentation based on effects of product attributes
    - ➢ Correlated Component Regression (CCR) to Select Attributes and Attribute Interactions (e.g., flavor preference depends upon texture)

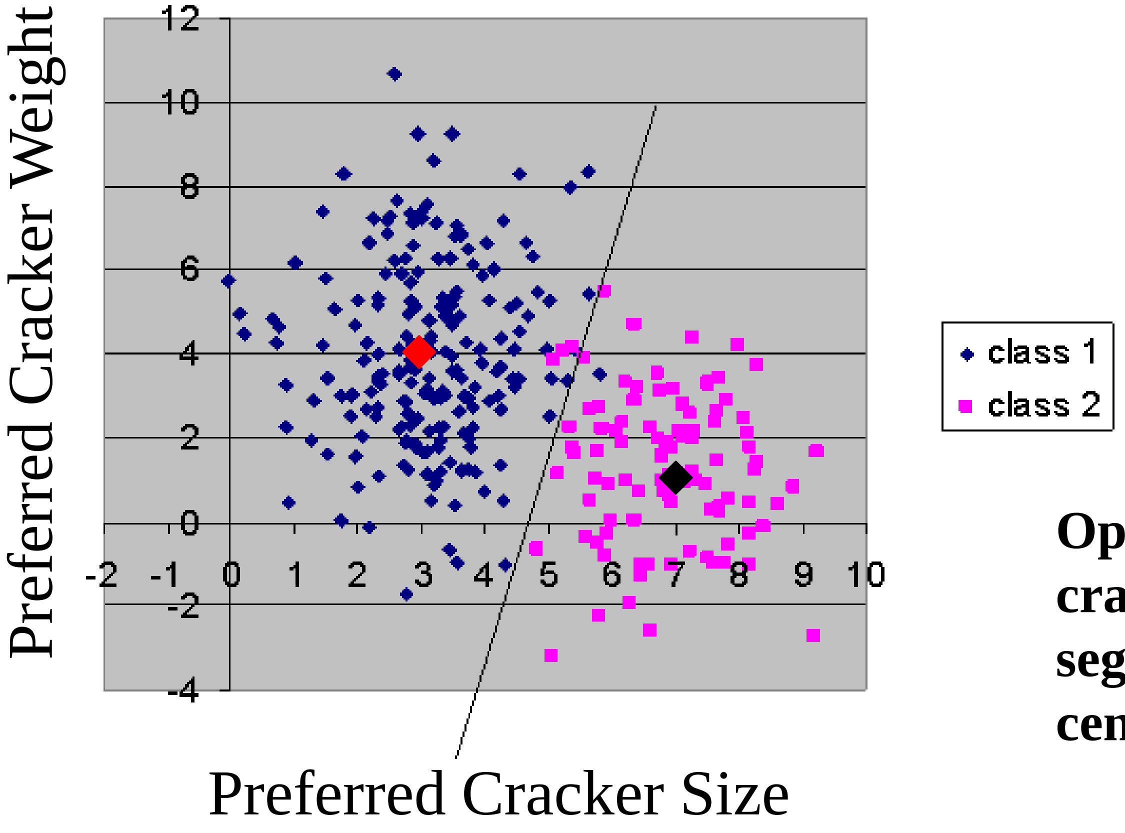
Idealized Example: Simulated data with 2 segments



Respondents in each segment (class) specify their preferred size and weight for crackers.

Mistakenly assuming a single homogeneous population, a single sub-optimal cracker can be developed with attributes at the centroid ◆ .

Idealized Example: Preferred Cracker Size & Weight



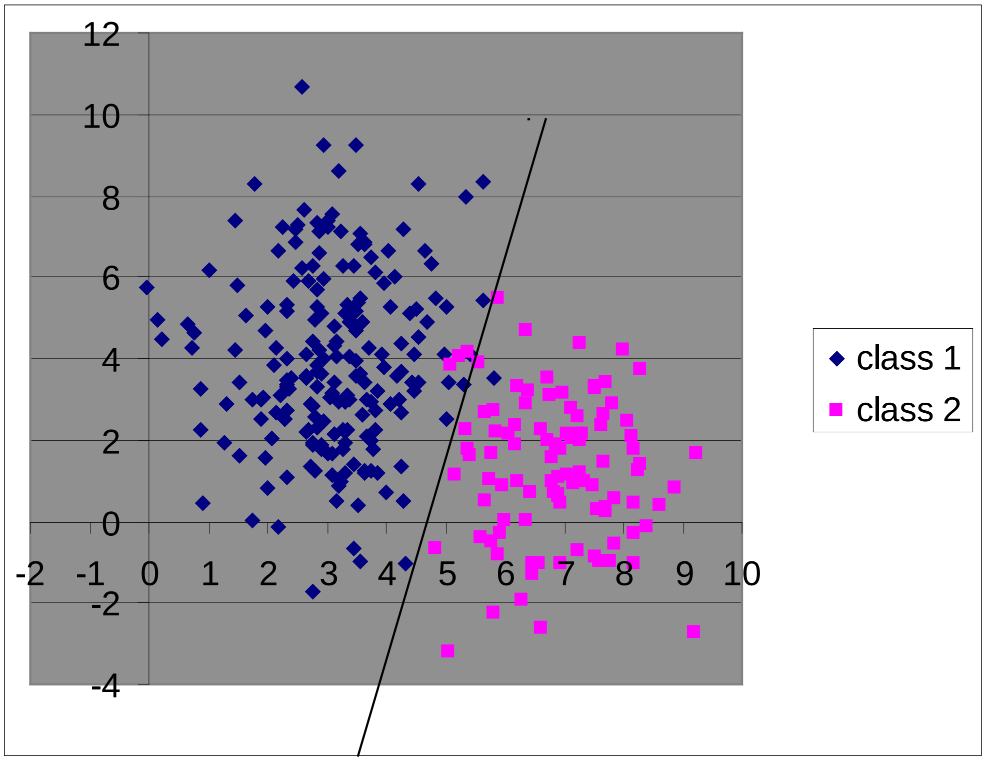
Latent Class analysis identifies 2 segments.

Within each segment the preferred cracker weight and size are independent (*local independence **).

Optimal -- develop 2 crackers, 1 for each segment, at the class centroids.

*** Class membership explains the correlation in the data.**

LC Results same as gold standard (discriminant analysis) -- only 4 cases misclassified – much better than K-means



K-means recovery:

• 24 cases misclassified;

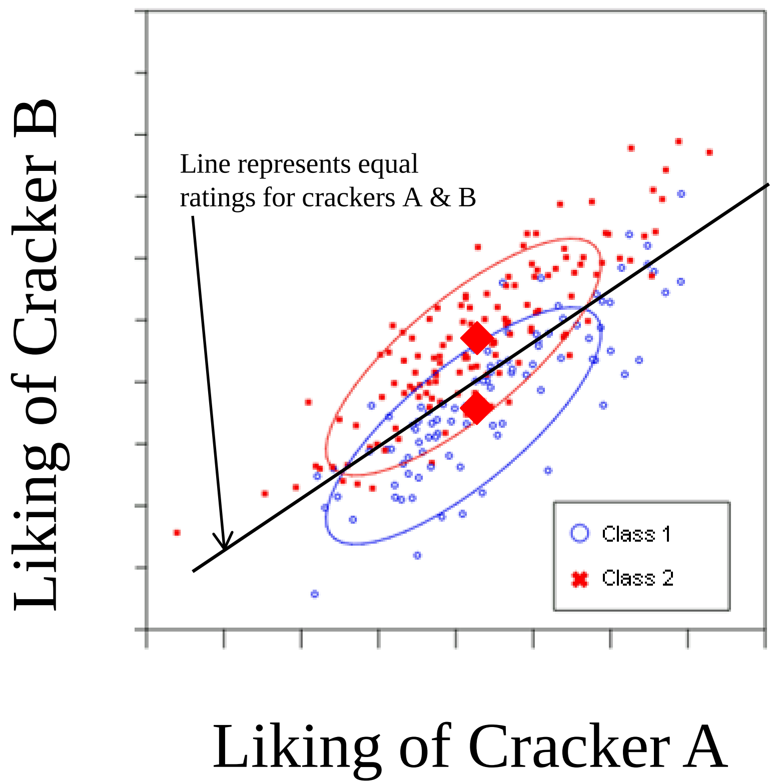
or if Z-scores are used*

• 15 cases misclassified

•Magidson and Vermunt (2002a, 2002b)

***LC results not affected by linear transformations of variables -- thus, LC model provides same results (4 misclassified) if Z-scores used instead of original metric.**

Real-world Data: Liking Ratings of Crackers A and B



Again, suppose there are 2 segments

Segments (classes) equal on liking of Cracker A

Class 2 higher on liking of Cracker B –

- **Class 1 prefers Cracker A over B**
- **Class 2 prefers Cracker B over A**

Local dependence **-- positive correlation remains within both classes.**

In real world some respondents give high ratings for all crackers while others tend to give lower ratings for all -- they like (dislike) all crackers or tend to use higher (lower) ratings ('response style').

Research Questions Addressed Here

1. For each of these data examples, how can Latent Class Modeling identify **meaningful** segments?

2. What techniques can assist in determining the most relevant attributes, and attribute levels for each segment?

Brief History of Latent Class Modeling

- LC proposed originally by Lazarsfeld (1950) as part of Latent Structure Analysis for dichotomous variables
- Maximum likelihood algorithm developed for nominal variables by Goodman (1974) (Now known as EM algorithm)
- Program advances: extension to many variables of differing scale types, approaches for handling *local dependence*, etc. – Latent GOLD (Vermunt and Magidson, 2000), Latent GOLD Choice (2003)
- Latent GOLD v 4.0 (2005) added continuous factors
  - e.g., factor mixture model, random effects models
- Latent GOLD v 4.5 (2008) added general syntax language

Modern Definition of Latent Class Modeling

“The basic idea underlying latent class (LC) analysis is a very simple one: some of the parameters of a postulated statistical model differ across unobserved subgroups. These subgroups form the categories of a categorical latent variable (called ‘latent classes’) … Outside the social sciences, LC models are often referred to as finite mixture models.”

Vermunt, J. and Magidson, J. Latent Class Analysis. *Encyclopedia of Social Science Research Methods*, Sage Publications, 2003

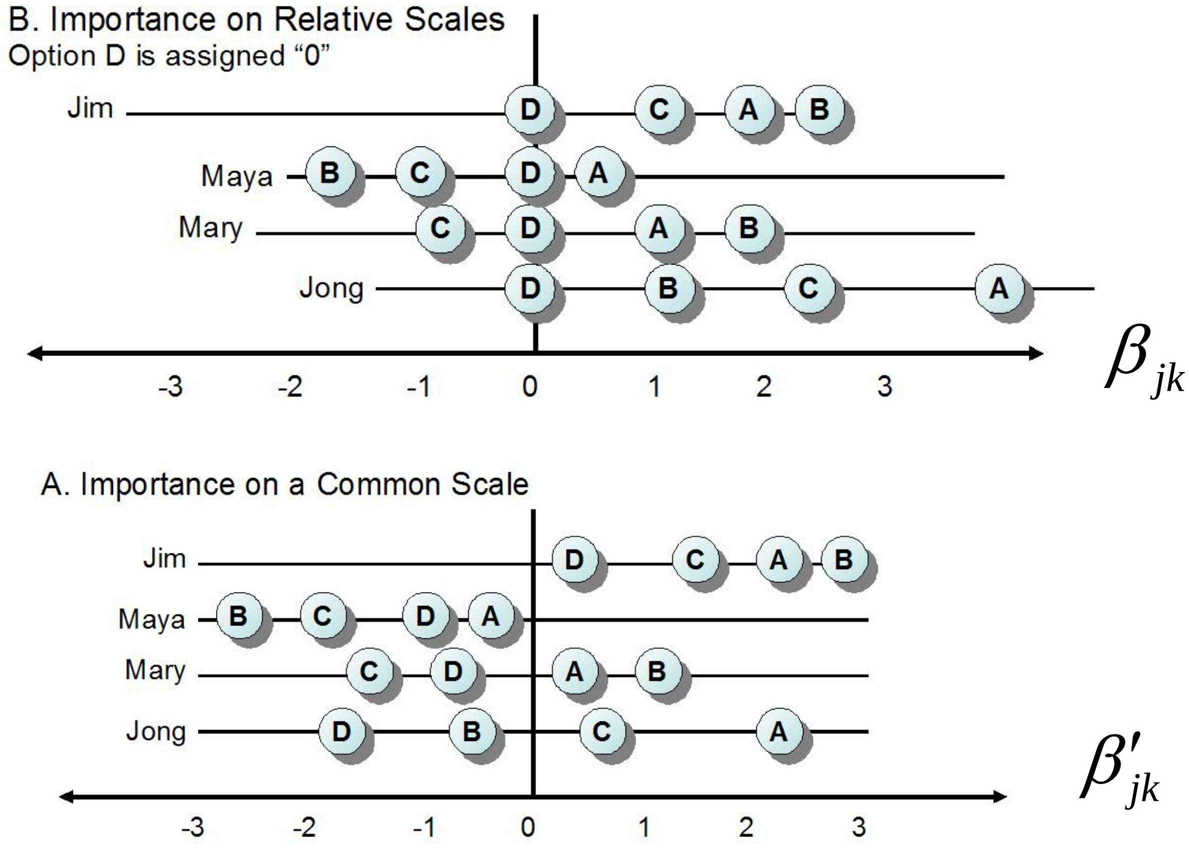
Latent Class Methods* also Can be Used to Explain Heterogeneity with Ranking (Full, or Partial such as MaxDiff / Best-Worst) Data

Relative scale (from ranking data) may be converted to absolute scale by adding appropriate *class-specific constants obtained using additional information from ratings – Magidson, et. al. 2009*

Mary judges attribute D as ***more*** important than C, but in absolute terms she does not consider either to be very important (Figure A).

For Jim, D is ***less*** important than D, and both important.

Given only their rankings, it is tempting, but not valid, to infer that Mary considers D to be more important than Jim



*Data fusion model developed using syntax version of LG Choice

General Latent GOLD Model

- Latent GOLD based on a simple probability structure from which most important LC models are derived

$$P(Y \mid Z) = \sum_x P(X \mid Z) P(Y \mid X, Z)$$

- Y is a set of dependent (endogenous) variables
- Z is a set of independent (exogenous) variables – predictors of Y, predictors of X ('covariates')
- X is a set of nominal/ordinal latent variables

- Y density is a weighted sum of class-specific exponential family densities (multinomial, Poisson, normal)
  - Estimates are obtained by maximizing the appropriate likelihood function

Mixed mode data: choosing the appropriate probability density function P(y) for each dependent variable

- nominal: multinomial
- ordinal: restricted multinomial
- counts: Poisson / binomial
- continuous: (multivariate) normal

- **Discrete choice data*** – first choice only, full ranking, partial ranking (best/worst "MaxDiff")

***Requires Latent GOLD Choice program**

Overview of Presentation

- Graphical Introduction
- **Kellogg Data Case study: cracker taste test**
  - ➢ LC Cluster Models
  - ➢ LC Regression Models – Segmentation based on effects of product attributes
  - ➢ Correlated Component Regression (CCR) to Select Attributes and Attribute Interactions (e.g., flavor preference depends upon texture)

Application of latent class models to food product development: a case study

For demo program, tutorials, and articles including Popper, Magidson, and Kroll (2004) article see website

http://statisticalinnovations.com/products/popper.pdf

Background

- Food manufacturers need to understand the taste preferences of their target consumers
- Taste preferences are rarely homogenous – different preference segments exist
- Latent class (LC) modeling can be used to determine ***meaningful segments*** and has many advantages over traditional clustering algorithms (e.g. hierarchical clustering, K-means)
- LC models also offer ways to separate out respondent heterogeneity due to:
  - differences in relative preference for one product over another
  - differences in average liking across all products

Background

- To guide food developers, important to ***relate a segment's taste preferences to the underlying sensory attributes*** of the product category (taste, texture, etc.)
- Some latent class models (LC regression/LC choice) allow attribute information to be used directly to predict liking, and thus used in forming segments, which can lead to more actionable results.

The Case Study

- **Products**: 15 crackers
- **Consumers**: n=157 (category users)
  - evaluated all products over three days
  - 9-point liking scale (dislike extremely➔like extremely)
  - completely randomized block design balanced for the effects of day, serving position, and carry-over

LC Segmentation Models -- 2 Kinds

- Cluster – Each class represents a grouping of cases that are similar in their responses to selected segmentation (dependent) variables (e.g., liking ratings on each of the 15 crackers).

- Regression – Each class represents a grouping of cases that are similar in their regression coefficients. Predictors in regression will be the cracker attributes (can also include interactions).

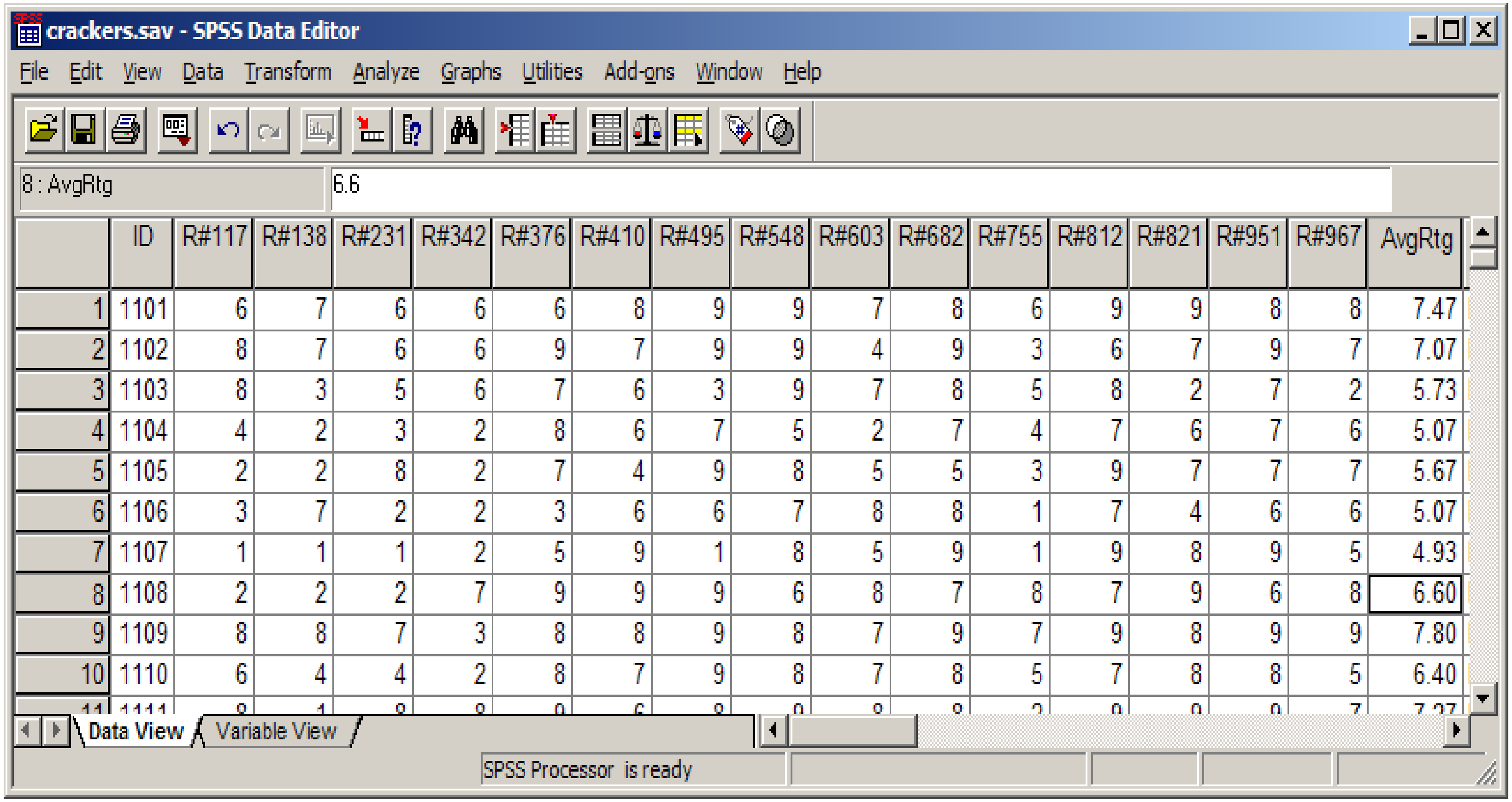
Objectives

- To determine if consumers could be segmented according to their liking ratings of the crackers
- To estimate and compare alternative models
  - LC Cluster model
  - LC Regression model with a random intercept (nominal factor + one continuous factor)
- For the regression models, to identify and interpret segments in terms of the sensory attributes that drive liking for that segment
- *Sparse regression methods* for determining most relevant attributes and interactions for each segment

Overview of Presentation

- **LC Cluster Models**
- LC Regression Models
- Correlated Component Regression (CCR) to Select Predictors and Interactions

LC Cluster Data Layout



	ID	R#117	R#138	R#231	R#342	R#376	R#410	R#495	R#548	R#603	R#682	R#755	R#812	R#821	R#951	R#967	AvgRtg
1	1101	6	7	6	6	6	8	9	9	7	8	6	9	9	8	8	7.47
2	1102	8	7	6	6	9	7	9	9	4	9	3	6	7	9	7	7.07
3	1103	8	3	5	6	7	6	3	9	7	8	5	8	2	7	2	5.73
4	1104	4	2	3	2	8	6	7	5	2	7	4	7	6	7	6	5.07
5	1105	2	2	8	2	7	4	9	8	5	5	3	9	7	7	7	5.67
6	1106	3	7	2	2	3	6	6	7	8	8	1	7	4	6	6	5.07
7	1107	1	1	1	2	5	9	1	8	5	9	1	9	8	9	5	4.93
8	1108	2	2	2	7	9	9	9	6	8	7	8	7	9	6	8	6.60
9	1109	8	8	7	3	8	8	9	8	7	9	7	9	8	9	9	7.80
10	1110	6	4	4	2	8	7	9	8	7	8	5	7	8	8	5	6.40

Ratings for each of the 15 products plus the average rating for each respondent

LC Cluster Model

- LC Cluster (Latent GOLD 4.5)
  - liking rating for each product treated as continuous (or ordinal*)
  - (a) with and b) without random intercept (i.e., with and without adjustment for response level effects)
  - under both situations, BIC (Bayesian Information Criterion) identifies a two class solution as a better fit to the data than either a one-class or three-class solutions

*for simplicity, equations illustrate *continuous scale type*

LC Cluster Model with T Product Ratings

$$Y_t = \alpha_t + \beta_{xt} + \varepsilon_t \qquad \text{fixed intercepts}$$

where: Y_t is the rating for product t, for respondents $i=1,2,\ldots,N$

α_t is the intercept associated with product t

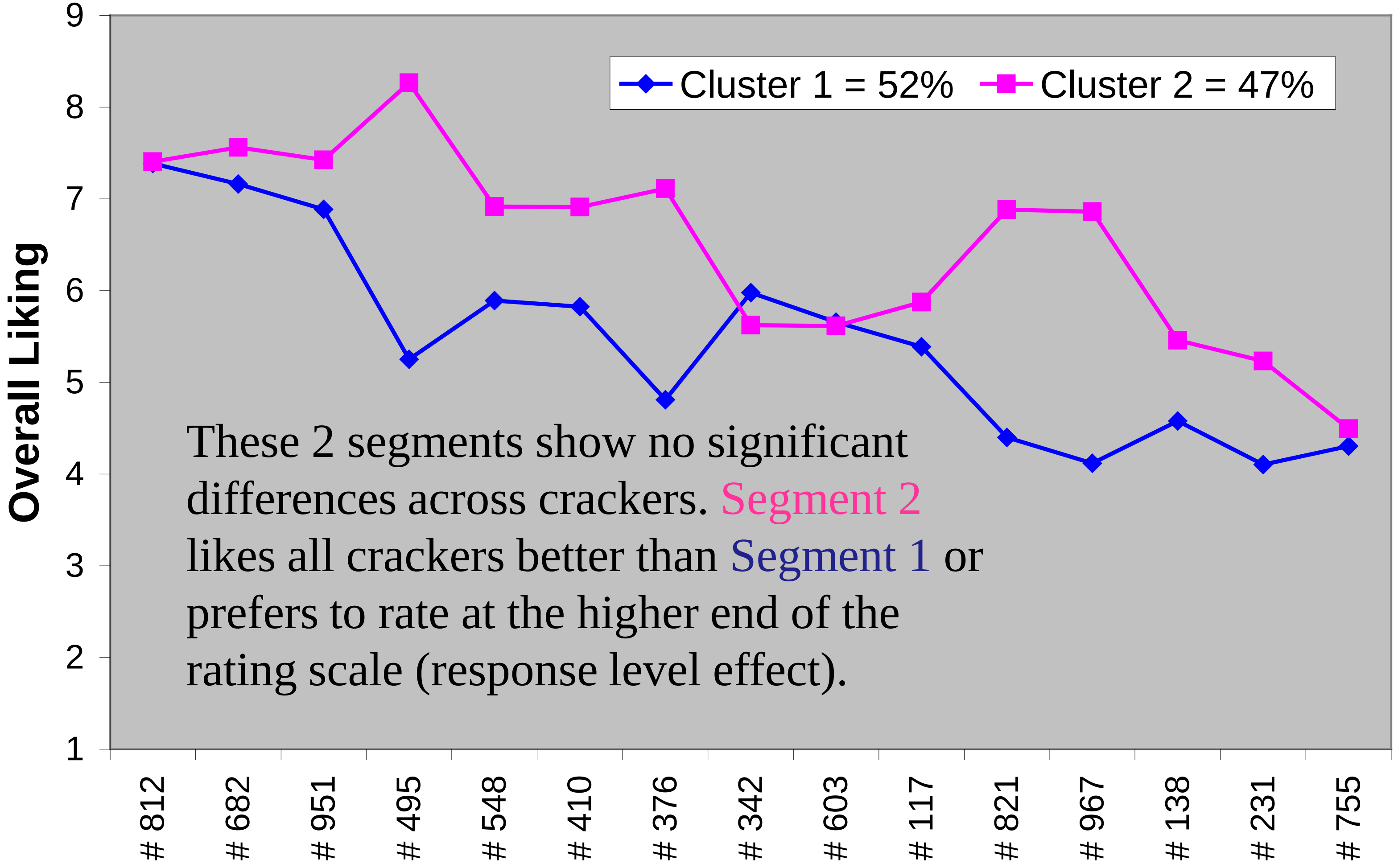
β_{xt} is the effect for product t for cases in latent class x

ε_t is random error assumed to be normally distributed (class-independent error variances)

Effect coding is used for parameter identification:

$$\sum_{t=1}^{T} \beta_{xt} = 0$$ (so intercepts capture average response levels)

Results from Traditional LC Cluster Model -- These 2 Segments are Not Very Useful



Overview of Presentation

- LC Cluster Models
- **LC Regression Models**
- Correlated Component Regression (CCR) to Select Predictors and Interactions

LC Regression Model

- A typical LC regression model with 2 predictors $Z=(Z_1, Z_2)$

$$P(Y \mid Z) = \sum_x P(X)P(Y \mid X, Z)$$

- For example, for Y continuous we have the LC linear regression model

$$Y = \alpha_x + b_{1x} Z_1 + b_{2x} Z_2 + \varepsilon_x$$

Model 1: LC Regression with Random Intercept and Discrete Random PRODUCT Effects

$$logit(Y_{im.t}) = \alpha_{im} + \beta_{xt}$$

$$\alpha_{im} = \alpha_m + \lambda F_i$$

Thus,

$$E(\alpha_{im}) = \alpha_m$$

$$V(\alpha_{im}) = \lambda^2$$

where:

logit($Y_{j.k}$) is the adjacent category logit associated with rating Y = m (vs. m-1) for product t

C-Factor F_i is the factor score for the ith respondent

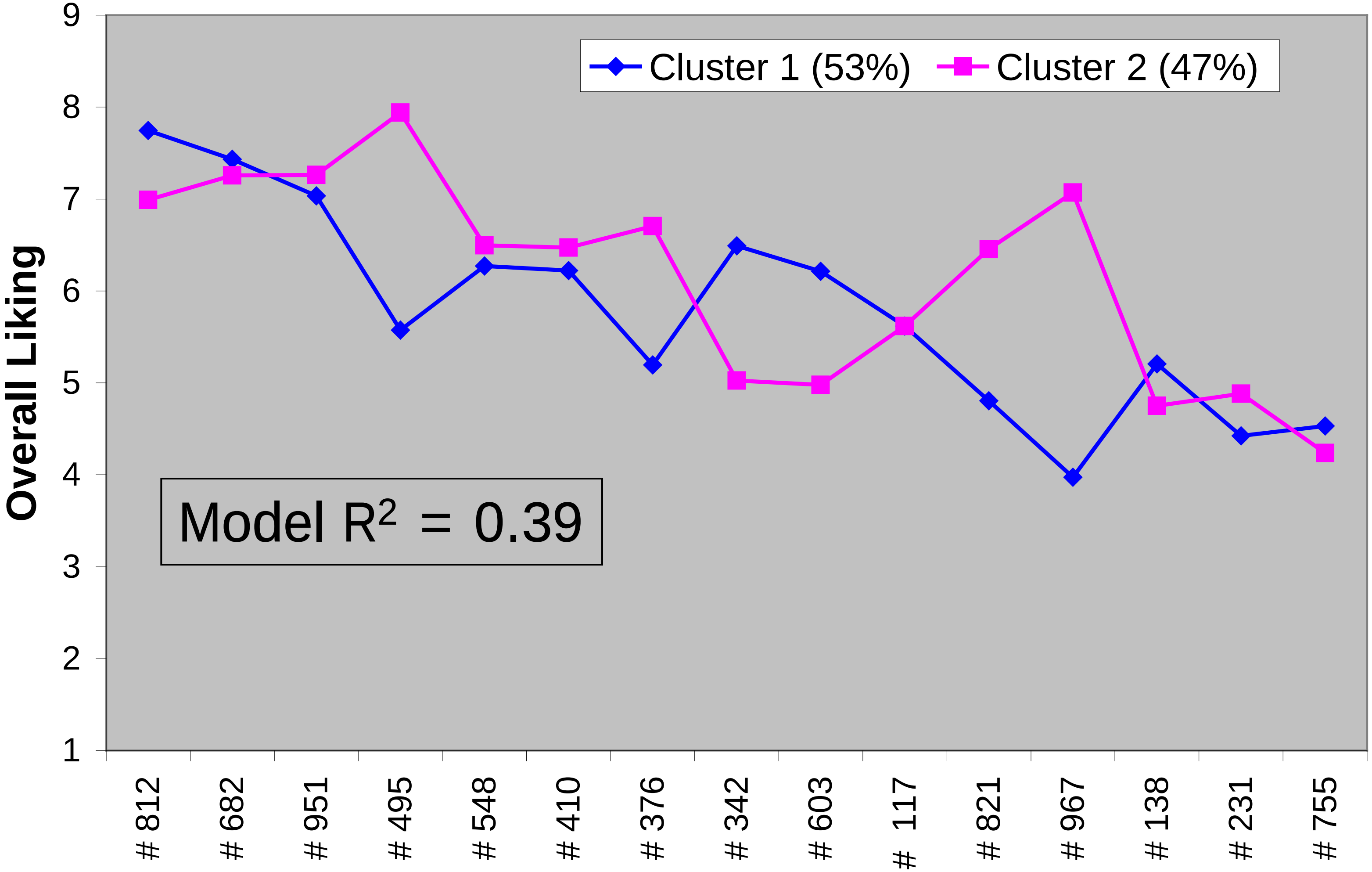
β_{xt} is the effect of the t^{th} product for class x

$F_i \sim N(0,1)$ or $\alpha_{im} \sim N(\alpha_m, \lambda^2)$ m = 2,3,…,M

and effect coding is used for parameter identification:

$$\sum_{t=1}^{T} \beta_{xt} = 0$$

Model 1: LC Regression with Random Intercept and Discrete Random PRODUCT Effects



Model 1: LC Regression with Random Intercept and Discrete Random PRODUCT Effects

- Correlation of random intercept with average liking is 0.997 (was 0.87 for D-Factor #1)
- Inclusion of random intercept is conceptually similar to mean-centering each respondents' liking ratings
  - LC Cluster model of the mean-centered data produces similar results
- Advantages of LC Regression over mean-centering
  - maintains ordinal metric
  - can be used with partial profile (incomplete block) designs

Including Sensory Attributes as Predictors

- **Products**: 15 crackers
- **Consumers**: n=157 (category users)
  - evaluated all products over three days
  - 9-point liking scale (dislike extremely➔like extremely)
  - completely randomized block design balanced for the effects of day, serving position, and carry-over
- **Sensory attribute evaluations**: trained sensory panel (n=8)
  - 18 flavor attributes, 20 texture attributes, 14 appearance rated on 15-point intensity scales (low➔high)
  - reduced (via PCA) to four appearance, four flavor, and four texture factors

LC Regression Models

Restructure the data for LC regression:

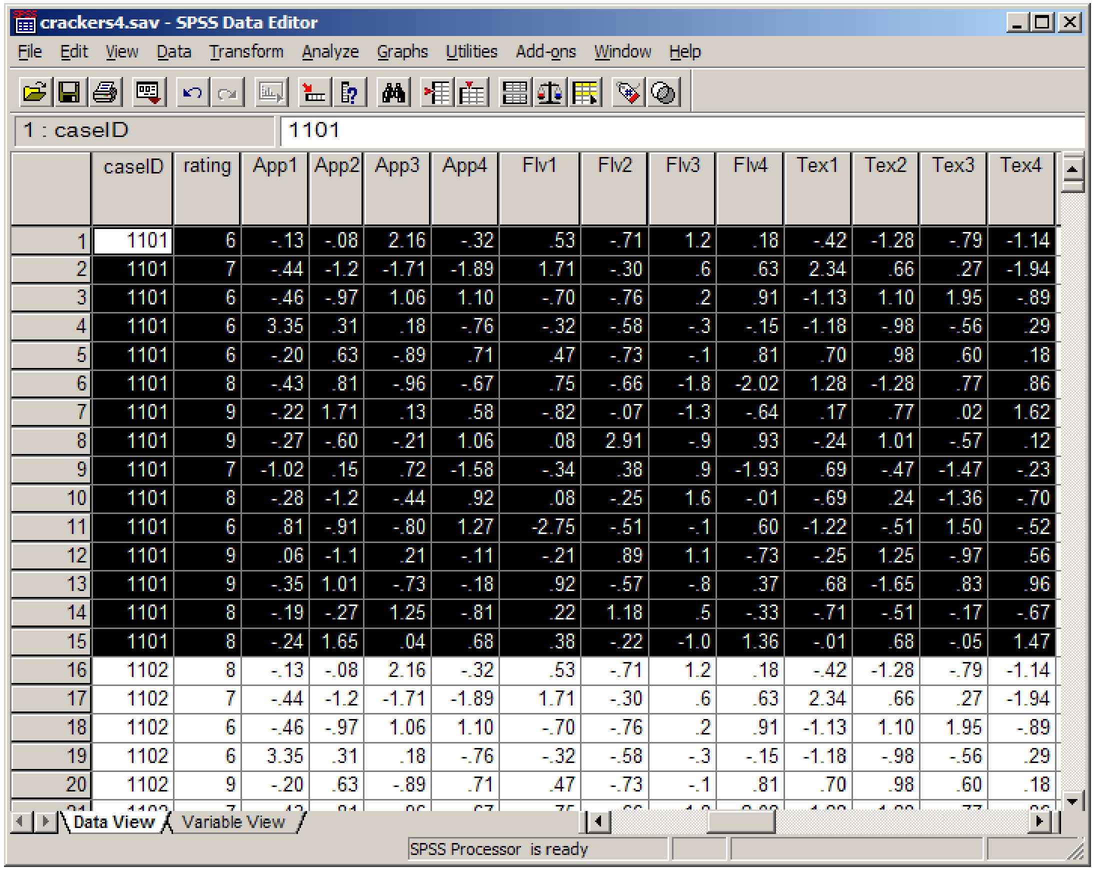
- Dependent variable = overall liking of product 1,2,...,15
  - T = 15 records (replications) per case

- Predictor = nominal PRODUCT variable (Model 1)

  OR

  Predictors = 12 sensory attributes (Model 2)

LC Regression Data Layout



	caseID	rating	App1	App2	App3	App4	Flv1	Flv2	Flv3	Flv4	Tex1	Tex2	Tex3	Tex4
1	1101	6	-.13	-.08	2.16	-.32	.53	-.71	1.2	.18	-.42	-1.28	-.79	-1.14
2	1101	7	-.44	-1.2	-1.71	-1.89	1.71	-.30	.6	.63	2.34	.66	.27	-1.94
3	1101	6	-.46	-.97	1.06	1.10	-.70	-.76	.2	.91	-1.13	1.10	1.95	-.89
4	1101	6	3.35	.31	.18	-.76	-.32	-.58	-.3	-.15	-1.18	-.98	-.56	.29
5	1101	6	-.20	.63	-.89	.71	.47	-.73	-.1	.81	.70	.98	.60	.18
6	1101	8	-.43	.81	-.96	-.67	.75	-.66	-1.8	-2.02	1.28	-1.28	.77	.86
7	1101	9	-.22	1.71	.13	.58	-.82	-.07	-1.3	-.64	.17	.77	.02	1.62
8	1101	9	-.27	-.60	-.21	1.06	.08	2.91	-.9	.93	-.24	1.01	-.57	.12
9	1101	7	-1.02	.15	.72	-1.58	-.34	.38	.9	-1.93	.69	-.47	-1.47	-.23
10	1101	8	-.28	-1.2	-.44	.92	.08	-.25	1.6	-.01	-.69	.24	-1.36	-.70
11	1101	6	.81	-.91	-.80	1.27	-2.75	-.51	-.1	.60	-1.22	-.51	1.50	-.52
12	1101	9	.06	-1.1	.21	-.11	-.21	.89	1.1	-.73	-.25	1.25	-.97	.56
13	1101	9	-.35	1.01	-.73	-.18	.92	-.57	-.8	.37	.68	-1.65	.83	.96
14	1101	8	-.19	-.27	1.25	-.81	.22	1.18	.5	-.33	-.71	-.51	-.17	-.67
15	1101	8	-.24	1.65	.04	.68	.38	-.22	-1.0	1.36	-.01	.68	-.05	1.47
16	1102	8	-.13	-.08	2.16	-.32	.53	-.71	1.2	.18	-.42	-1.28	-.79	-1.14
17	1102	7	-.44	-1.2	-1.71	-1.89	1.71	-.30	.6	.63	2.34	.66	.27	-1.94
18	1102	6	-.46	-.97	1.06	1.10	-.70	-.76	.2	.91	-1.13	1.10	1.95	-.89
19	1102	6	3.35	.31	.18	-.76	-.32	-.58	-.3	-.15	-1.18	-.98	-.56	.29
20	1102	9	-.20	.63	-.89	.71	.47	-.73	-.1	.81	.70	.98	.60	.18

The data file is now restructured so that the dependent variable RATING can be predicted as a function of 1) PRODUCT or 2) the taste attributes.

Model 2: LC Regression with Random Intercept and Discrete Random Product Attribute Effects

$$logit(Y_{im.t}) = \alpha_{im} + \beta_{x1} Z_1 + \beta_{x2} Z_2 + \ldots + \beta_{xT} Z_Q$$

$$\alpha_{im} = \alpha_m + \lambda F_i$$

Thus,

$$E(\alpha_{im}) = \alpha_m$$

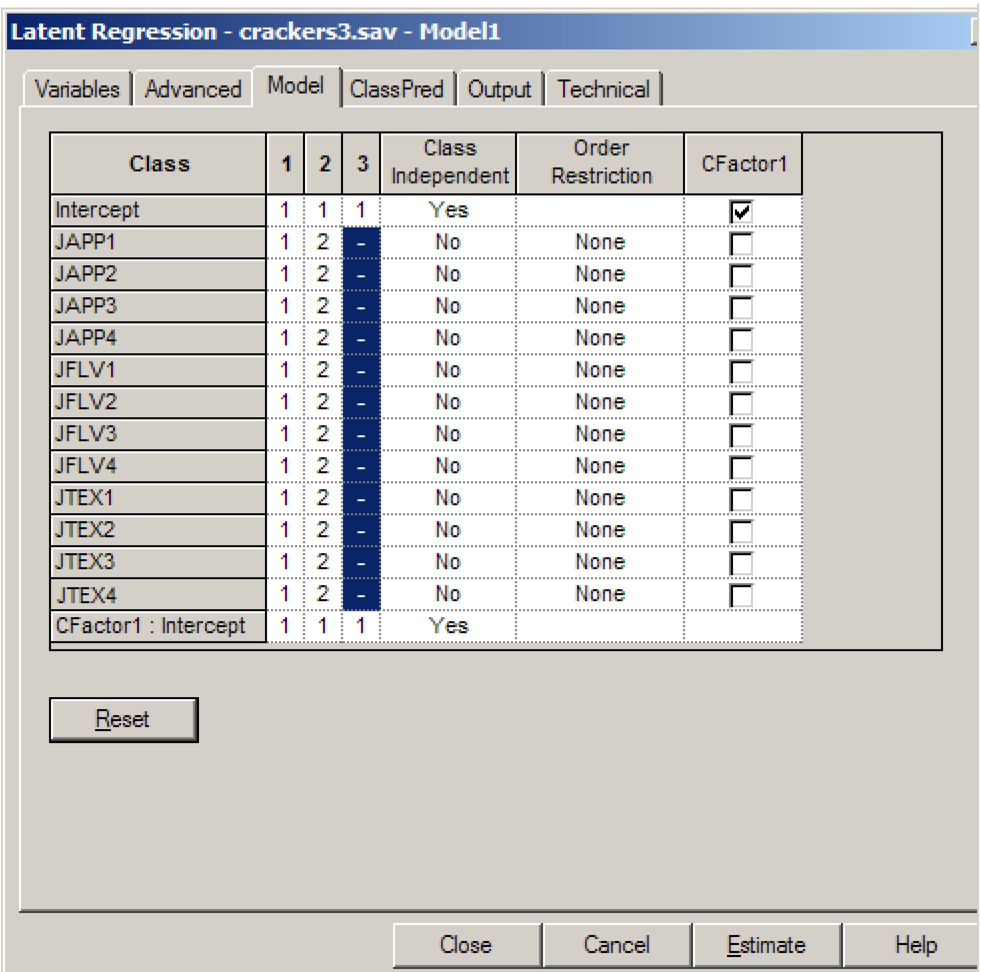
$$V(\alpha_{im}) = \lambda^2$$

where:

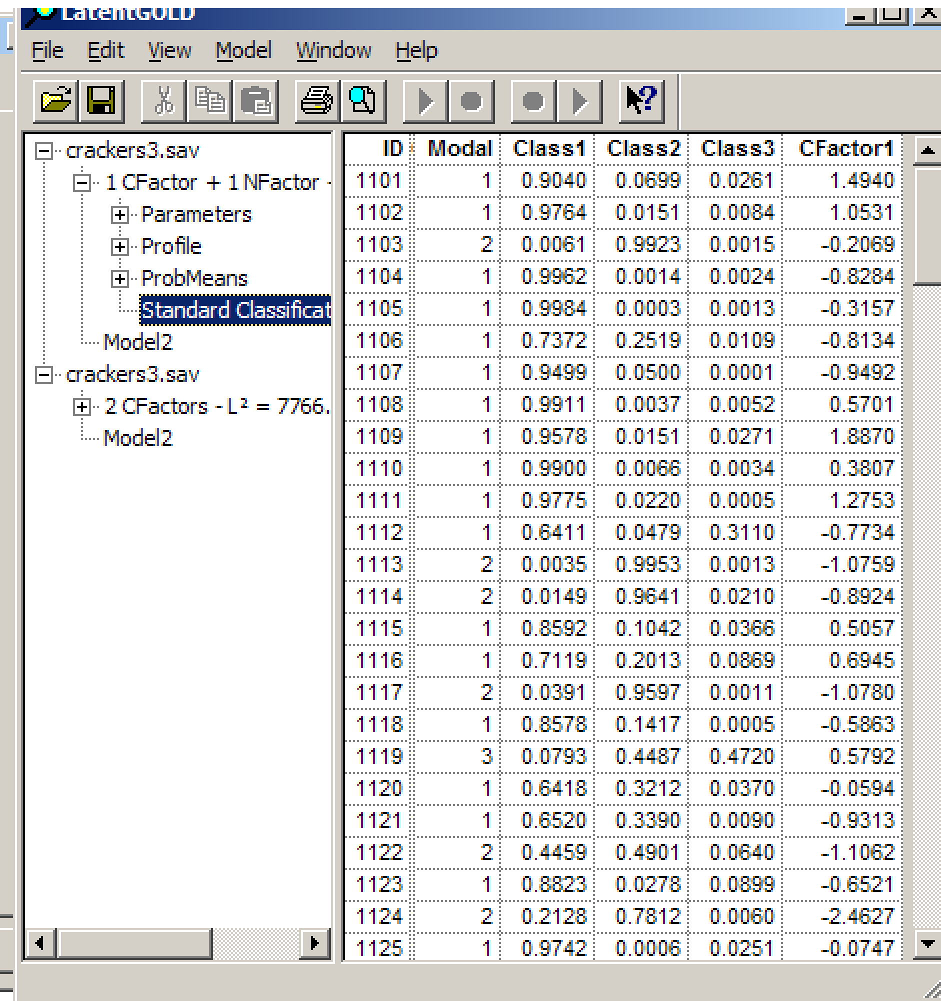
$logit(Y_{im.t})$ is the adjacent category logit for product t with attributes $Z_1, Z_2, \ldots, Z_Q$

β_{xq} is the effect of the qth attribute for class x

Setup and Classification Output for 3-class Random Intercept Model 2 where Attributes do Not Predict Liking for Class 3

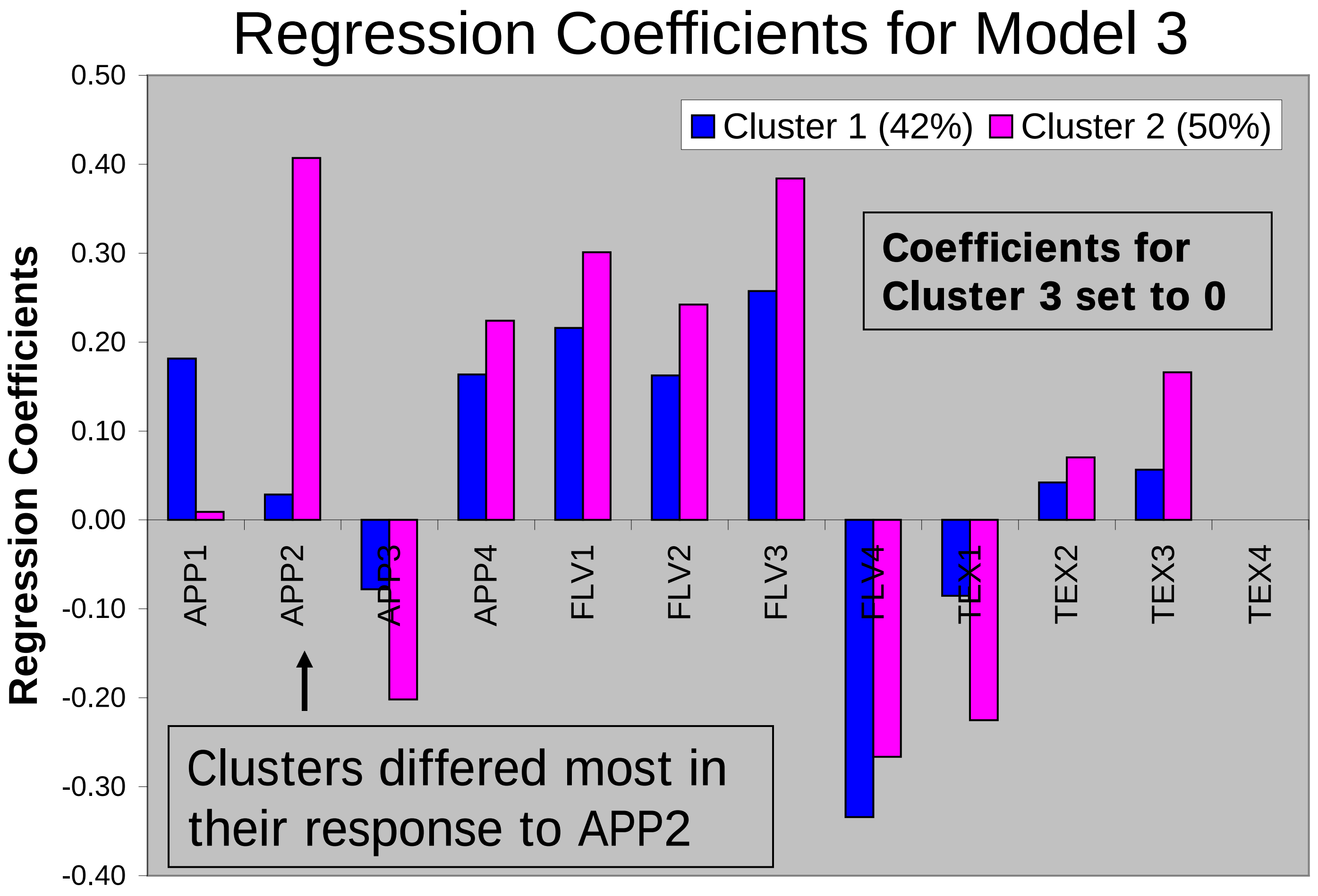


Class	1	2	3	Class Independent	Order Restriction	CFactor1
Intercept	1	1	1	Yes		☑
JAPP1	1	2	-	No	None	☐
JAPP2	1	2	-	No	None	☐
JAPP3	1	2	-	No	None	☐
JAPP4	1	2	-	No	None	☐
JFLV1	1	2	-	No	None	☐
JFLV2	1	2	-	No	None	☐
JFLV3	1	2	-	No	None	☐
JFLV4	1	2	-	No	None	☐
JTEX1	1	2	-	No	None	☐
JTEX2	1	2	-	No	None	☐
JTEX3	1	2	-	No	None	☐
JTEX4	1	2	-	No	None	☐
CFactor1 : Intercept	1	1	1	Yes		

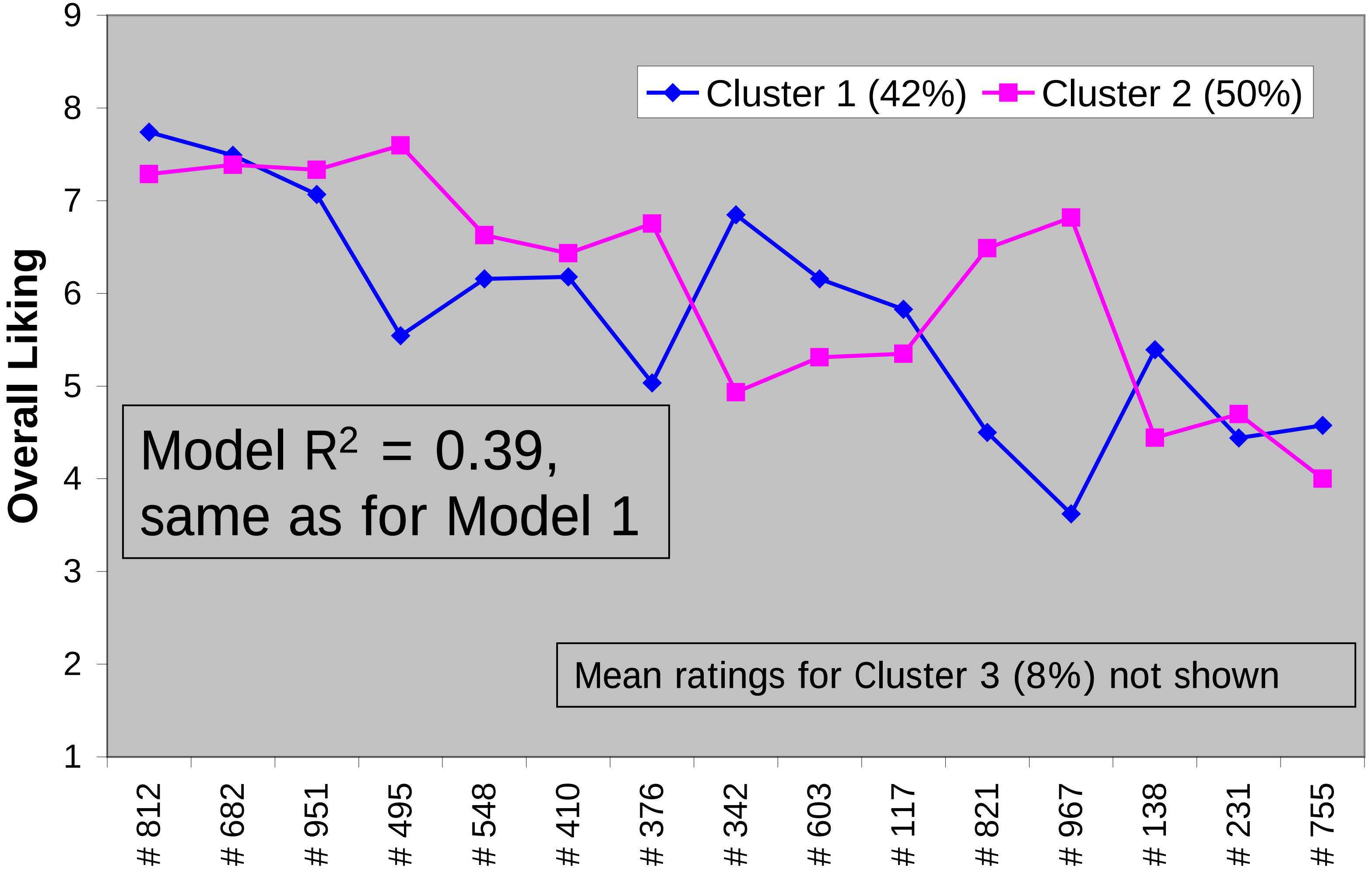


ID	Modal	Class1	Class2	Class3	CFactor1
1101	1	0.9040	0.0699	0.0261	1.4940
1102	1	0.9764	0.0151	0.0084	1.0531
1103	2	0.0061	0.9923	0.0015	-0.2069
1104	1	0.9962	0.0014	0.0024	-0.8284
1105	1	0.9984	0.0003	0.0013	-0.3157
1106	1	0.7372	0.2519	0.0109	-0.8134
1107	1	0.9499	0.0500	0.0001	-0.9492
1108	1	0.9911	0.0037	0.0052	0.5701
1109	1	0.9578	0.0151	0.0271	1.8870
1110	1	0.9900	0.0066	0.0034	0.3807
1111	1	0.9775	0.0220	0.0005	1.2753
1112	1	0.6411	0.0479	0.3110	-0.7734
1113	2	0.0035	0.9953	0.0013	-1.0759
1114	2	0.0149	0.9641	0.0210	-0.8924
1115	1	0.8592	0.1042	0.0366	0.5057
1116	1	0.7119	0.2013	0.0869	0.6945
1117	2	0.0391	0.9597	0.0011	-1.0780
1118	1	0.8578	0.1417	0.0005	-0.5863
1119	3	0.0793	0.4487	0.4720	0.5792
1120	1	0.6418	0.3212	0.0370	-0.0594
1121	1	0.6520	0.3390	0.0090	-0.9313
1122	2	0.4459	0.4901	0.0640	-1.1062
1123	1	0.8823	0.0278	0.0899	-0.6521
1124	2	0.2128	0.7812	0.0060	-2.4627
1125	1	0.9742	0.0006	0.0251	-0.0747

Parameter Estimates from LC Regression on Sensory Variables with Random Intercept



Results from LC Regression on Sensory Variables with Random Intercept



LC Regression Model 2 Results

- A 2-class model was preferred over a 3-class model according to BIC.
- BIC for a 3-class restricted model was slightly better than for a 2-class unrestricted model
  - The third class was restricted to have regression coefficients of 0 for all 12 predictors and represents individuals whose liking does not depend on the 12 sensory attributes
  - This group can be of substantive interest for follow-up or be excluded as outliers. Here the group was small (8%)

LC Regression Model 2 Results

- Model 2 incorporates sensory information that provides direction for product development:
  - overall, respondents agree that they prefer crackers that are high in Flav1-3, low in Flav4, low in Tex1 and high in Tex2-3
  - segments differ primarily in their reaction to the appearance attributes: Cluster 1 prefers products high in APP2 and low in APP3. Cluster 2 was not highly influenced by these two characteristics, but preferred crackers high in APP1.
- Model 2 also provides information about the size the third cluster of respondents who are not affected by the sensory variables

Summary of Results

- The traditional LC Cluster model confounded different taste preferences with response level effects
  - Cluster 1 rated almost all products higher than Cluster 2

- LC Regression with a random intercept provided clear evidence of segment differences in consumers' liking ratings
  - While some products appealed to everybody, some products appealed much more to one segment than the other.
  - LC Regression Model 2 produced a 3-segment solution which showed how the segments were affected by the sensory attributes.

Conclusion and Follow-up Issue of Variable Selection with Small Samples

- Separate food products may be developed for each segment based on their different sensory preferences for crackers.
- However, there may be hundreds of sensory attributes, and for a given number of attributes there may be a large number of 2-way interactions (i.e., the effect of texture may vary depending upon appearance or flavor). Beyond 15–1 = 14 predictors, traditional techniques can not improve prediction (high-dimensional data)

Overview of Presentation

- **LC Cluster Models**
- LC Regression Models
- **Correlated Component Regression (CCR)** to Select Predictors and Interactions

Variable Selection: Small Samples and Many Predictors

Current approaches for analyzing *high dimensional data*:

1. Penalty Approaches – tends to omit predictors that are highly correlated with other predictors in model
2. PLS Regression – requirement that components be orthogonal yields extra components
3. Correlated Component Regression (CCR) – Similar to PLS Regression but **fewer, more interpretable components** than PLS
   - Comparisons of these methods with Sparse Data: **Performance favors CCR over the other approaches**

Results from Simulated Data -- Comparison of Several Variable Selection Methods:

Correlated Component Regression (CCR), Elastic Net (L1 + L2 regularization, Zou and Hastie, 2005), Lasso (L1 regularization), and sparse PLS regression (sgpls, Chun and Keles, 2009)

Design: Data simulated according to assumptions of Linear Discriminant Analysis

G_1 = 28 predictors (including 15 weak predictors) plus G_2 = 28 irrelevant predictors
2 Groups: $N_1 = N_2 = 25$; 100 simulated samples

Method M select G*(M) < 56 predictors for final model; Each method tuned using same sized validation file. Final models from each method evaluated based on large independent 'test' file.

Results favor CCR over the other approaches (Magidson and Yuan, 2010)

Lowest misclassification error rate:
  CCR (17.4%), sparse PLS (19.1%), Elastic net (20.2%), lasso (20.8%)
Fewest irrelevant variables:
  CCR (3.4), lasso (6.2), Elastic net (11.5), sparse PLS (13.1)
Most sparse solution (average # predictors in model):
  CCR (14.5), lasso (17.3), Elastic net (28.3), sparse PLS (32.3)

COREXpress^{TM}
Correlated Component Regression (CCR)

COREXpress^{TM} Beta Program

To apply for a beta version of COREXpress^{TM} contact:

Will Barker
Sales & Marketing
375 Concord Ave., Suite 007
Belmont, MA 02478
+1 (617) 489-4490
will@statisticalinnovations.com

Acknowledgment

I wish to thank The Kellogg Company for providing the data for this case study, and for allowing it to be distributed along with a tutorial for Latent GOLD.

References

Magidson, J., D. Thomas, J.K. Vermunt (*2009)* A New Model for the Fusion of Maxdiff Scaling and Ratings Data", *Sawtooth Software Conference Proceedings,* 83-103.

Magidson, J., J.K. Vermunt (*2002a) "Latent Class Modeling as a Probabilisitic Extension of K-Means Clustering",Quirk's Marketing Research Review*, March 2002, 20 & 77-80.

Magidson, J., J.K. Vermunt (*2002b) "Latent Class Models for Clustering: A Comparison with K-Means", Canadian Journal of Marketing Research*, 20, 36-43.

Vermunt, J. and Magidson, J. (2003) Latent Class Analysis. *Encyclopedia of Social Science Research Methods*, Sage Publications.

Vermunt, J. and Magidson, J. (2000) Latent GOLD Technical Guide, Belmont MA.: Statistical Innovations.