An Introduction to Latent Class and Finite Mixture Modeling

JEROEN K. VERMUNT
Department of Methodology and Statistics,
Tilburg University, The Netherlands

Institüt für Soziology, Abteilung für empirische Sozialforschung

Johannes Kepler Universität Linz

January 14, 2005

1. Latent Class (LC) and Finite Mixture (FM) Models

- General definition: parameters of the model of interest differ across unobserved subgroups

- Most important applications:
  - (probabilistic) clustering
  - scaling / measurement error / discretized IRT
  - density estimation
  - unobserved heterogeneity / correlated observations / (nonparametric) random-effects modeling

- Latent GOLD implements these applications in three modules: LC cluster, LC factor, and LC regression
  - not as flexible as possible, but flexible enough
  - as easy (interface, graphs, output measures), save (random starts / Bayes constants, diagnostics), fast, and stable (EM +NR) as possible
  - work in progress: restrictions / choice / longitudinal / multilevel / complex sampling

- Traditional LC model (Lazarsfeld & Goodman)
  - LC cluster model with nominal observed variables
  - independence model with different parameters for each latent class (=local independence)

2. General Latent GOLD Model

- Simple probability structure from which the three modules are derived

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

  - Y is a set of dependent/response/endogenous variables
  - Z is a set of independent/predictor/exogenous variables
  - X is a set of nominal/ordinal latent variables

  - density *P(Y|Z)* is a weighted average of class-specific densities, which are from the exponentional family (multinomial, Poisson, normal)
  - in most cases: local independence between replications
  - MIMIC model

- For example, the traditional latent class model is:

$$P(y_1 y_2 y_3) = \sum_{x_1} P(x_1) P(y_1 \mid x_1) P(y_2 \mid x_1) P(y_3 \mid x_1)$$

  - note: all Y's are nominal, no Z's, a single X

- A GLM is specified for each Y variable

  - with nominal/ordinal variables these are logit models:

$$P(y_1 \mid x_1) = \frac{\exp(\alpha_{y_1} + \beta_{y_1 x_1})}{\sum_{y_1} \exp(\alpha_{y_1} + \beta_{y_1 x_1})}$$

3. The Traditional LC Model

- EXAMPLE General Social Survey 1982, white sample (see McCutcheon, 1987; Magidson & Vermunt, Soc. Meth. 2001)
  - evaluation of surveys by respondent (2 questions)
  - evaluation of respondent by interviewer (2 questions)
  - are there different types of survey respondents?

PURPOSE	ACCURACY	UNDERSTANDING	COOPERATION	Whites	Blacks
Good	Mostly true	Good	Interested	419	117
			Cooperative	35	14
			Impatient/Hostile	2	3
		Fair, poor	Interested	71	34
			Cooperative	25	19
			Impatient/Hostile	5	5
	Not true	Good	Interested	270	95
			Cooperative	25	10
			Impatient/Hostile	4	3
		Fair, poor	Interested	42	23
			Cooperative	16	14
			Impatient/Hostile	5	2
Depends	Mostly true	Good	Interested	23	7
			Cooperative	4	1
			Impatient/Hostile	1	0
		Fair, poor	Interested	6	3
			Cooperative	2	1
			Impatient/Hostile	0	0
	Not true	Good	Interested	43	19
			Cooperative	9	1
			Impatient/Hostile	2	2
		Fair, poor	Interested	9	2
			Cooperative	3	1
			Impatient/Hostile	2	1
Waste	Mostly true	Good	Interested	26	6
			Cooperative	3	0
			Impatient/Hostile	0	0
		Fair, poor	Interested	1	3
			Cooperative	2	1
			Impatient/Hostile	0	0
	Not true	Good	Interested	85	30
			Cooperative	23	9
			Impatient/Hostile	6	1
		Fair, poor	Interested	13	9
			Cooperative	12	7
			Impatient/Hostile	8	4

Profile Output

	Cluster 1 Ideal	Cluster 2 Believers	Cluster 3 Skeptics
LC Probabilities	0.6169	0.2038	0.1793
Conditional Probabilities			
PURPOSE			
Good	0.8905	0.9157	0.1592
Depends	0.0524	0.0706	0.2220
Waste	0.0570	0.0137	0.6189
ACCURACY			
Mostly True	0.6148	0.6527	0.0426
Not True	0.3852	0.3473	0.9574
UNDERSTAND			
Good	0.9957	0.3241	0.7532
Fair, poor	0.0043	0.6759	0.2468
COOPERATE			
Interested	0.9452	0.6879	0.6432
Cooperative	0.0547	0.2583	0.2559
Impatient/ Hostile	0.0001	0.0538	0.1009

Fit measures

Model	L^2	prop. red. L^2	df	p-value	BIC	AIC	class. errors
1-Cluster	257.26	0.00	29	0.00	51.60	199.26	0.00
2-Cluster	79.51	0.69	22	0.00	-76.51	35.51	0.08
3-Cluster	22.09	0.91	15	0.11	-84.29	-7.91	0.13
4-Cluster	6.61	0.97	8	0.58	-50.12	-9.39	0.20

4. Model Selection, Classification, Estimation, and Problems

4.1 Model selection

- Chi-squared statistics
  - overall tests: L-squared and X-squared
  - conditional tests: regularity conditions
  - sparseness

- Analysis of association
  - proportional reduction of L-squared

- Information criteria: BIC, AIC, CAIC
  - model comparison
  - parsimony versus fit

- Bivariate residuals
  - local lack of fit

- Quality of classification
  - see classification
  - fit, parsimony, and classification: AWE

4.2 Classification

- After estimating a LC model, we can classify individuals into latent classes

- Latent classification or posterior membership probabilities

$$P(x_1 \mid y_1 y_2 y_3) = \frac{P(x_1)P(y_1 \mid x_1)P(y_2 \mid x_1)P(y_3 \mid x_1)}{\sum_{x_1} P(x_1)P(y_1 \mid x_1)P(y_2 \mid x_1)P(y_3 \mid x_1)}$$

- Allocation methods
  - modal allocation
  - proportional allocation
  - random allocation (+ multiple)

- Quality of classification
  - proportion of classification errors
  - reduction of classification errors
  - other reduction of errors measures

EXAMPLE GSS'82 data (continued)

Classification output for 3-cluster model

PURPOSE	ACCURACY	UNDERSTA	COOPERAT	Modal	Cluster1	Cluster2	Cluster3
1	1	1	1	1	0.9197	0.0786	0.0017
1	1	1	2	1	0.6382	0.3537	0.0081
1	1	1	3	2	0.0155	0.9435	0.0410
1	1	2	1	2	0.0238	0.9729	0.0033
1	1	2	2	2	0.0037	0.9927	0.0036
1	1	2	3	2	0.0000	0.9932	0.0068
1	2	1	1	1	0.8780	0.0637	0.0583
1	2	1	2	1	0.5188	0.2442	0.2369
1	2	1	3	3	0.0068	0.3503	0.6429
1	2	2	1	2	0.0246	0.8528	0.1227
1	2	2	2	2	0.0038	0.8644	0.1318
1	2	2	3	2	0.0000	0.7761	0.2238
2	1	1	1	1	0.8653	0.0968	0.0380
2	1	1	2	1	0.4934	0.3579	0.1487
2	1	1	3	2	0.0070	0.5560	0.4370
2	1	2	1	2	0.0173	0.9257	0.0570

4.3 Estimation

- Maximum likelihood (ML)
  - standard estimation method in LC analysis
  - known favorable properties like unbiasedness, consistency, etc.
  - probability density for data: multinomial distribution (not restrictive at all)
  - easy to adapt to other types of Y variables (see mixture models and LC cluster)

- Kernel log-likelihood function:

$$\ln Lik = \sum_{y_1 y_2 y_3} n_{y_1 y_2 y_3} \ln P(y_1 y_2 y_3)$$
$$= \sum_{y_1 y_2 y_3} n_{y_1 y_2 y_3} \ln \sum_{x_1} P(x_1) P(y_1 \mid x_1) P(y_2 \mid x_1) P(y_3 \mid x_1)$$

- EM algorithm
  - general iterative method for ML estimation with missing data (here X_1 is missing)
  - E step: fill in missing data
  - M step: estimate parameters in the usual way
  - stable (converges irrespective of starting values)
  - fast per iteration, but may need many iterations

- E step

$$\hat{n}_{x_1 y_1 y_2 y_3} = n_{y_1 y_2 y_3} P(x_1 \mid y_1 y_2 y_3)$$

$$P(x_1 \mid y_1 y_2 y_3) = \frac{P(x_1)P(y_1 \mid x_1)P(y_2 \mid x_1)P(y_3 \mid x_1)}{\sum_{x_1} P(x_1)P(y_1 \mid x_1)P(y_2 \mid x_1)P(y_3 \mid x_1)}$$

- M step

$$\ln Lik^* = \sum_{x_1 y_1 y_2 y_3} \hat{n}_{x_1 y_1 y_2 y_3} \ln P(x_1)P(y_1 \mid x_1)P(y_2 \mid x_1)P(y_3 \mid x_1)$$

$P(x_1) = \hat{n}_{x_1} / N$, $P(y_1 \mid x_1) = \hat{n}_{x_1 y_1} / \hat{n}_{x_1}$, etc..

- M step becomes a bit more complicated when a conditional density is restricted by a GLM. But, standard complete data methods can be used.

- Newton-Raphson algorithm
  - standard optimization algorithm
  - first and second derivatives log-likelihood
  - unstable (needs good starting values)
  - fast when near to the maximum

- Latent Gold: hybrid algorithm
  - EM when far from the maximum
  - Newton-Raphson when close enough to maximum

4.4 Problems and possible solutions

- Non-identification
  - different sets of parameter estimates yield the same value of L-squared: estimates are not unique
  - detection: running the model with different sets of starting values, or formally from the rank of the information matrix
  - famous example: 3-cluster model for 4 dichotomous indicators

- Local maxima
  - there are different sets of 'ML' parameter estimates with different L^2 values
  - we want the solution with lowest L^2 (highest likelihood)
  - solution: multiple sets of starting values
  - example GSS'82: 3-cluster model with L^2 of 37.57 instead of 22.09

- Boundary solutions
  - estimated probability becomes zero, or log-linear parameters extremely large negative
  - solution: using priors to smooth the estimates
  - example GSS'82: 3-cluster model contains 2 zero probabilities

5. Extensions of the Traditional LC Model

- 4 most important extensions of traditional LC model
  - several latent variables (LC factor)
  - local dependencies
  - ordinal variables
  - covariates

5.1 Several latent variables (LC factor model)

- When items don't measure a single dimension, we may introduce a second latent variable
  - Similar to (confirmatory/exploratory) factor analysis

- For example, two latent variables:

$$P(y_1 y_2 y_3) = \sum_{x_1 x_2} P(x_1 x_2) P(y_1 \mid x_1 x_2) P(y_2 \mid x_1 x_2) P(y_3 \mid x_1 x_2)$$

with: $P(y_1 \mid x_1 x_2) = \dfrac{\exp(\alpha_{y_1} + \beta_{y_1 x_1} + \beta_{y_1 x_2})}{\sum_{y_1} \exp(\alpha_{y_1} + \beta_{y_1 x_1} + \beta_{y_1 x_2})}$

some β's may be set equal to zero

- This is done in the LC factor module, with the additional restriction that the latent variables are dichotomous/ordinal (see Magidson and Vermunt, Soc. Meth. 2001).

5.2 Local dependencies

- Why doesn't a LC model fit?
  - because local independence assumption is violated

- 3 possible solutions:
  1. increase the number of clusters or latent classes
  2. increase the number of factors or latent variables
  3. allow for local dependencies or direct relationships between certain items

- Option 3 is similar to correlated errors in structural equation models

- For example:

$$P(y_1 y_2 y_3) = \sum_x P(x_1) P(y_1 y_2 \mid x_1) P(y_3 \mid x_1)$$

with:

$$P(y_1 y_2 \mid x_1) = \frac{\exp(\alpha_{y_1} + \alpha_{y_2} + \alpha_{y_1 y_2} + \beta_{y_1 x_1} + \beta_{y_2 x_1})}{\sum_{y_1 y_2} \exp(\alpha_{y_1} + \alpha_{y_2} + \alpha_{y_1 y_2} + \beta_{y_1 x_1} + \beta_{y_2 x_1})}$$

- Interpretation: two items are more strongly related than can be explained by clusters/factors

- Exploratory versus theoretical justification

- Detection: bivariate residuals (kind of modification indices)

EXAMPLE: Political orientation (Hagenaars, 1993)

5 dichotomous variables

- system responsiveness (low/high)
- ideological level (nonideologies/ideologies)
- repression potential (high/low)
- protest approval (low/high)
- conventional participation (low/high)

Fit measures:

Model	L-squared	BIC	df	p-value
1-cluster	296.56	113	26	0.00
2-cluster	95.82	-45	20	0.00
3-cluster	24.80	-74	14	0.04
4-cluster	7.45	-49	8	0.49
2-factor	12.30	-86	14	0.58
2-cluster + 1 direct effect	19.10	-115	19	0.45

Bivariate residuals in 2-cluster model

Indicators	SYS_RESP	IDEO_LEV	REP_POT	PROT_APP
IDEO_LEV	0.6751			
REP_POT	3.5917	0.3085		
PROT_APP	0.1405	0.0739	**57.8794**	
CONV_PAR	2.5345	0.5543	3.2535	2.0772

5.3 Ordinal indicators

- So far: dichotomous or unordered polytomous indicators

- When indicators are ordinal, we can restrict their relationship with the latent variable(s)

- Several options
  1. different types of logits
     - adjacent-category logit
     - cumulative logit (or probit)
     - continuation-ratio logit
  2. inequality restrictions
  3. binomial count (k out of K)

- In Latent Gold, we work with adjacent category logit models with fixed scores for item categories, for example

$$\beta_{y_1 x_1} = \beta_{x_1} y_1$$

- Assumption: local odds-ratios are category independent

- Note: in the LC factor model, we assume that the latent variables are ordinal

EXAMPLE GSS'82 data (continued)

Fit measures

Model		L^2	BIC	df	p-value
Nominal	1-Cluster	257.26	51.60	29	0.00
	2-Cluster	79.51	-76.51	22	0.00
	3-Cluster	22.09	-84.29	15	0.11
Ordinal	1-Cluster	257.26	51.60	29	0.00
	2-Cluster	82.60	-87.60	24	0.00
	3-Cluster	30.67	-104.07	19	0.04

Profile output ordinal 3-cluster (constraint?)

	Cluster1	Cluster2	Cluster3
Cluster Size	0.6407	0.1925	0.1668
PURPOSE			
good	0.8792	0.2512	0.9167
depends	0.0756	0.1475	0.0579
waste	0.0451	0.6013	0.0254
ACCURACY			
mostly true	0.6233	0.0415	0.6753
not true	0.3767	0.9585	0.3247
UNDERSTA			
good	0.9949	0.7483	0.2027
fair/poor	0.0051	0.2517	0.7973
COOPERAT			
interested	0.9362	0.6365	0.6971
cooperative	0.0609	0.2754	0.2412
impatient/hostile	0.0029	0.0881	0.0617

5.4 Covariates

- Other terms: exogenous, explanatory, external, independent, concomitant, or predictor variables

- Probably most important extension of standard LC model

- Covariates to predict class membership: joint classification and prediction

- More elegant than separate classification and prediction

- Covariates may be nominal or continuous/numeric

- A model with two nominal covariates:

$$P(y_1 y_2 y_3 \mid z_1 z_2) = \sum_{x_1} P(x_1 \mid z_1 z_2) P(y_1 \mid x_1) P(y_2 \mid x_1) P(y_3 \mid x_1)$$

with a logit model for the latent variable:

$$P(x_1 \mid z_1 z_2) = \frac{\exp(\alpha_{x_1} + \beta_{x_1 z_1} + \beta_{x_1 z_2})}{\sum_{x_2} \exp(\alpha_{x_1} + \beta_{x_1 z_1} + \beta_{x_1 z_2})}$$

- Alternative: inactive covariates

EXAMPLE: Political orientation (Hagenaars, 1993)

Parameters output

Latent Variable(s) (gamma)			
	Cluster1	Cluster2	Cluster3
Intercept	0.5907	-0.0762	-0.5146
Covariates	Cluster1	Cluster2	Cluster3
SEX			
Men	0.2951	-0.2291	-0.0660
Women	-0.2951	0.2291	0.0660
EDUC			
Some College	0.9812	-0.5570	-0.4242
Less Than College	-0.9812	0.5570	0.4242
AGE			
16-34	-0.3910	-0.8987	1.2897
35-57	0.2776	0.0491	-0.3266
58-91	0.1134	0.8497	-0.9631

ProbMeans output

	Cluster1	Cluster2	Cluster3
Overall Probability	0.4167	0.3526	0.2307
Covariates			
SEX			
Men	0.5283	0.2507	0.2210
Women	0.3294	0.4325	0.2381
EDUC			
Some College	0.7660	0.0969	0.1371
Less Than College	0.1767	0.5285	0.2949
AGE			
16-34	0.3898	0.0979	0.5124
35-57	0.5213	0.3466	0.1320
58-91	0.3296	0.6153	0.0551

6. Simple Mixture Models

- Basic idea of mixture models
  - unobserved heterogeneity
  - parameters of the specified model are different for subgroups
  - unknown to which subgroup someone belongs
  - subgroups are latent classes

- Applicable to:
  1. univariate distributions: mean + variance
     - density estimation
     - clustering
  2. multivariate distributions: means + variances/covariances
     - multivariate density estimation
     - clustering
  3. regression models: coefficients
     - clustering using a regression model
     - unobserved heterogeneity
     - dependence between observations
  4. more complex models: coefficients

- Note: standard LC model is also a mixture model

- Estimation, testing, and classification issues same as in models discussed so far

Mixtures of univariate distributions

- A single Y , which mean and variance differs across subgroups

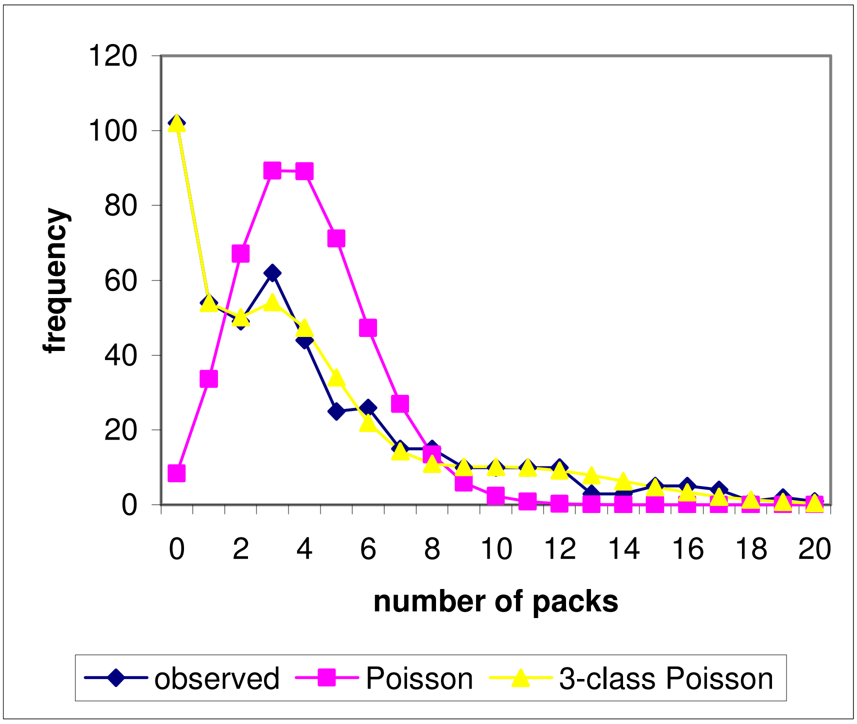
- In Latent Gold, Y may be nominal, ordinal, count, or continuous

- Formula: $P(y_1) = \sum_{x_1} P(x_1)P(y_1 \mid x_1)$

- The exact form of $P(y_1 \mid x_1)$ depends of the type of Y variable

EXAMPLE: Mixture of Poisson distributions (weekly consumption of hard-candy)

Observed and estimated distributions



Fit measures

Model	LL	BIC	Npar
1-Class	-1545.00	3096.12	1
2-Class	-1188.83	2396.03	3
3-Class	-1132.04	2294.70	5
4-Class	-1130.08	2303.01	7
5-Class	-1130.07	2315.25	9

7. LC or Model-Based Clustering

- LC analysis is becoming a popular clustering tool

- Is application of mixtures of multivariate distributions

- Clustering via LC or mixture models has many advantages over traditional types of cluster analysis:
  - large samples
  - mixed mode data
  - no linear scale transformation necessary
  - favorable properties of statistical model: model options, statistical tests

- Basic formula:

$$P(y_1 y_2 y_3) = \sum_{x_1} P(x_1) P(y_1 \mid x_1) P(y_2 \mid x_1) P(y_3 \mid x_1)$$

- Same basic assumption as standard LC model, which is, in fact, the special case with all variables nominal/ordinal. Local independence assumption can be relaxed.

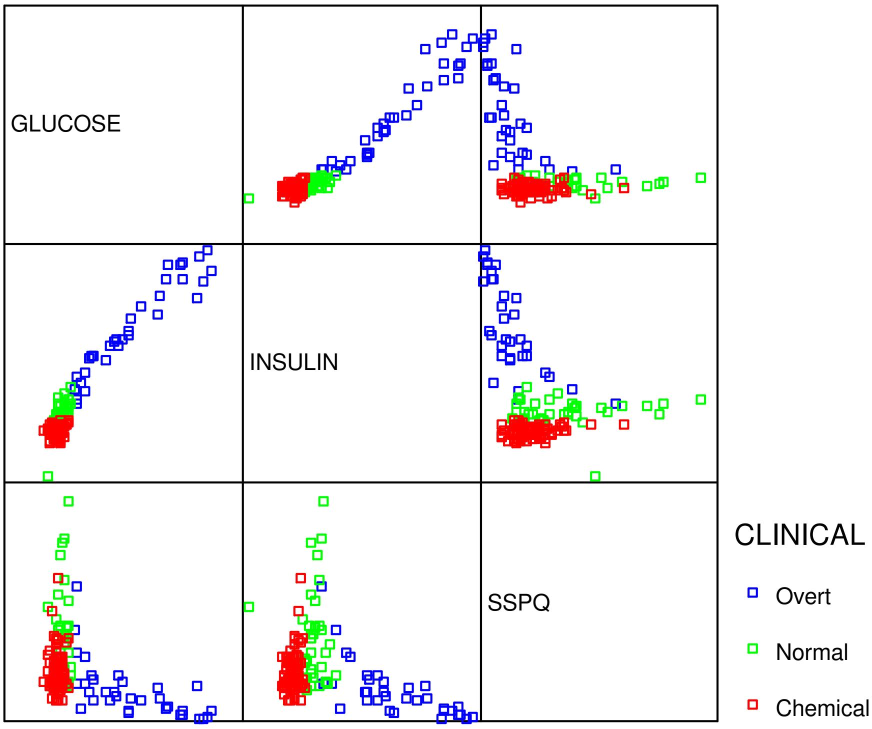
- Mixed mode data: choosing the appropriate probability density function for each indicator variable
  - nominal: multinomial
  - ordinal: restricted multinomial
  - counts: Poisson / binomial
  - continuous: (multivariate) normal

- LC cluster versus K-means

EXAMPLE: Diabetes data

Three continuous measures to diagnose diabetes: Glucose, Insulin, and SSPG (steady-state plasma glucose)

Information on the clinical diagnosis: "normal", "chemical diabetes", and "overt diabetes"

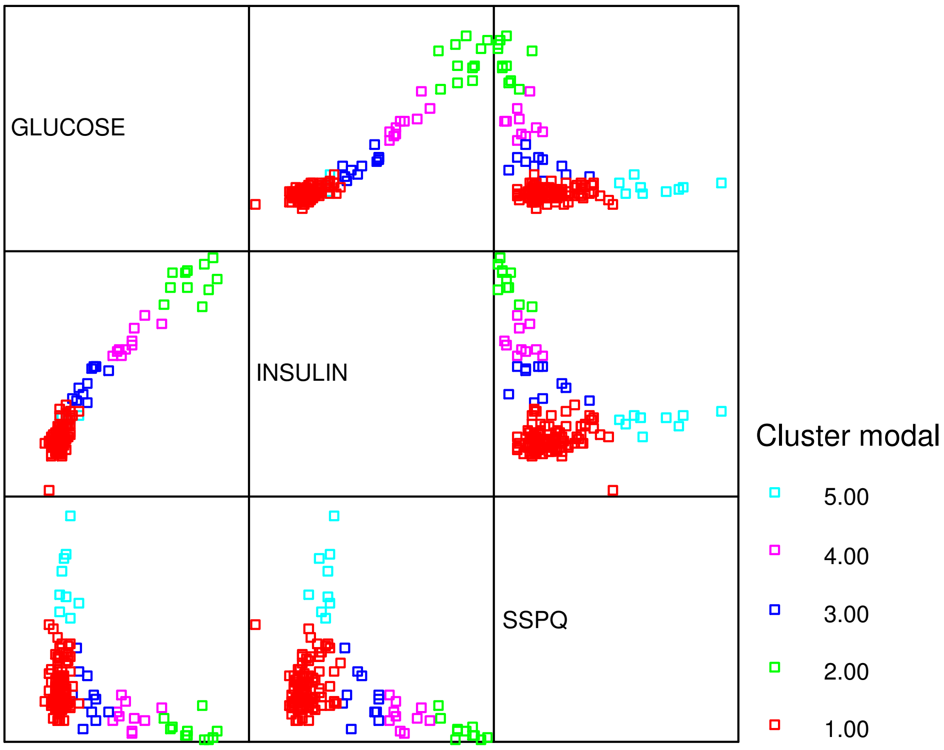


Note: cluster forms/shapes are very different from K-means clusters

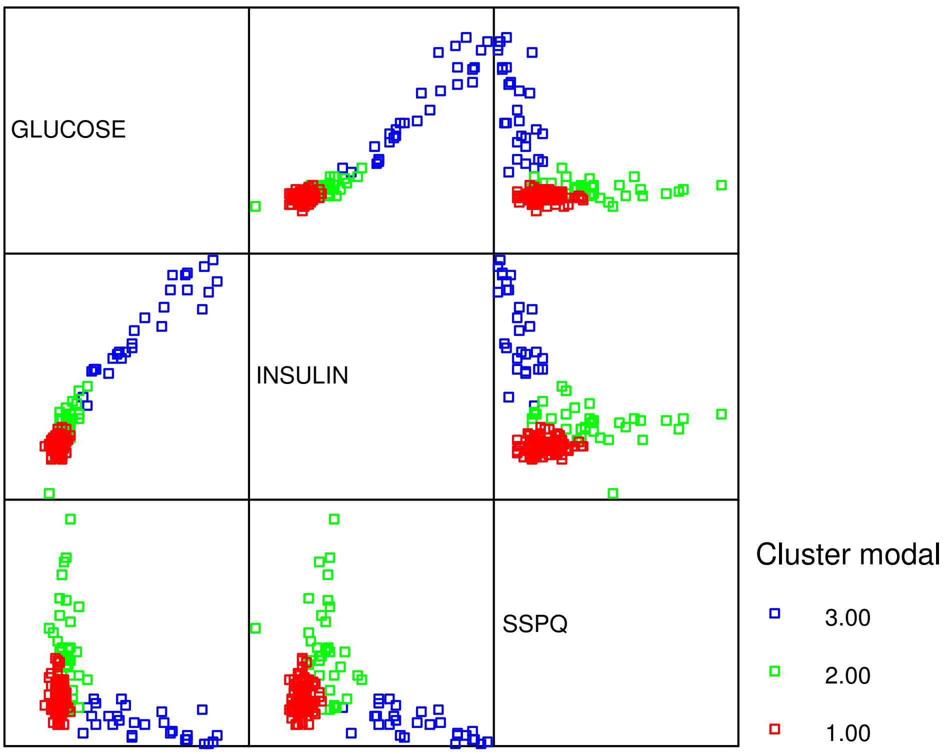
Fit measures

Model	LL	BIC	Npar
Equal and diagonal			
1-Cluster	-2750.13	5530.13	6
2-Cluster	-2559.88	5169.52	10
3-Cluster	-2464.78	4999.24	14
4-Cluster	-2424.46	4938.49	18
5-Cluster	-2392.56	4894.60	22
Unequal and diagonal			
1-Cluster	-2750.13	5530.13	6
2-Cluster	-2446.12	4956.94	13
3-Cluster	-2366.92	4833.38	20
4-Cluster	-2335.38	4805.13	27
5-Cluster	-2323.13	4815.47	34
Unequal and full			
1-Cluster	-2546.83	5138.46	9
2-Cluster	-2359.12	4812.80	19
3-Cluster	-2308.64	4761.61	29
4-Cluster	-2298.59	4791.28	39
5-Cluster	-2284.97	4813.79	49
Unequal and y_1-y_2 free			
1-Cluster	-2560.40	5155.64	7
2-Cluster	-2380.27	4835.19	15
3-Cluster	-2320.57	4755.61	23
4-Cluster	-2303.14	4760.56	31
5-Cluster	-2295.05	4784.19	39

K-Means type model



Final 3-cluster model



EXAMPLE: cancer data (mixed mode)

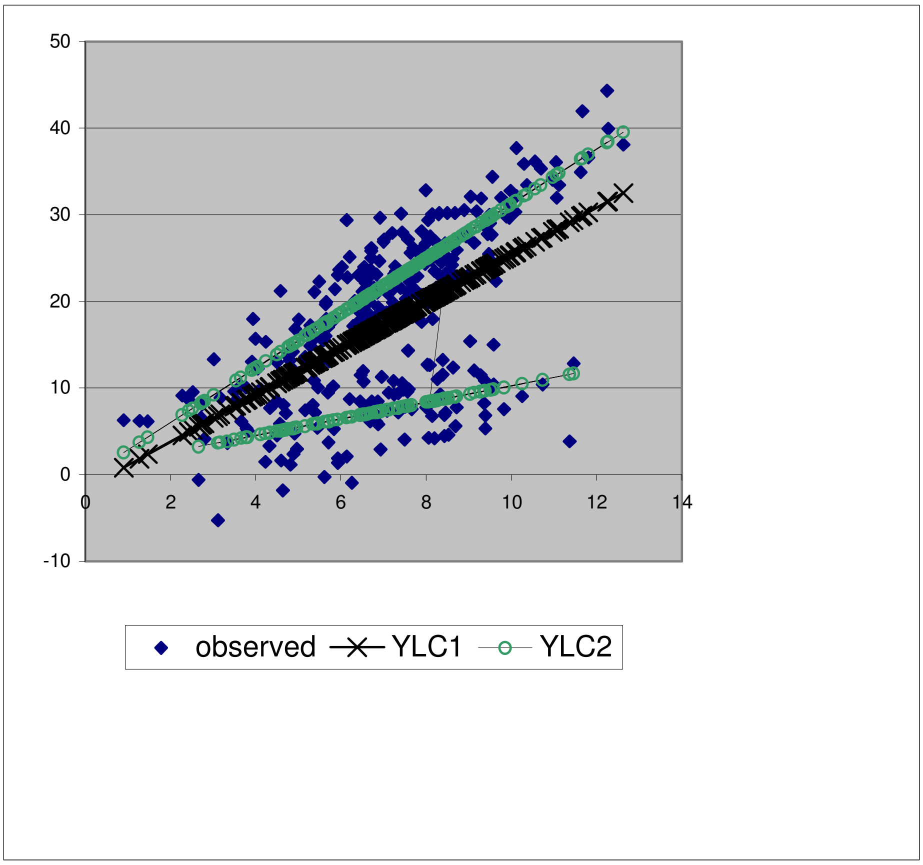
mixed mode data: 8 continuous and 4 nominal variables

class-dependent variances

Model	LL	BIC	Npar	class. errors
1 class	-11798	23762	27	0.000
2 class	-11386	23112	55	0.036
3 class	-11289	23089	83	0.100
4 class	-11202	23088	111	0.111

8. LC or Mixture Regression Models

- Single Y and one or more Z's
- Formula: $P(y_1 \mid Z) = \sum_{x_1} P(x_1) P(y_1 \mid x_1, Z)$
- The parameters of the model in which Y is regressed on Z differ across latent classes



- One way of seeing the model: cluster analysis based on a regression model

- Another perspective: two-level nonlinear mixed model

    - non-parametric specification random-effects
    - replications / repeated measures
    - level-1 predictors and level-2 predictors (covariates)
    - fixed (class-independent) and random (class-dependent) coefficients

    - very fast and stable compared to standard nonlinear mixed models

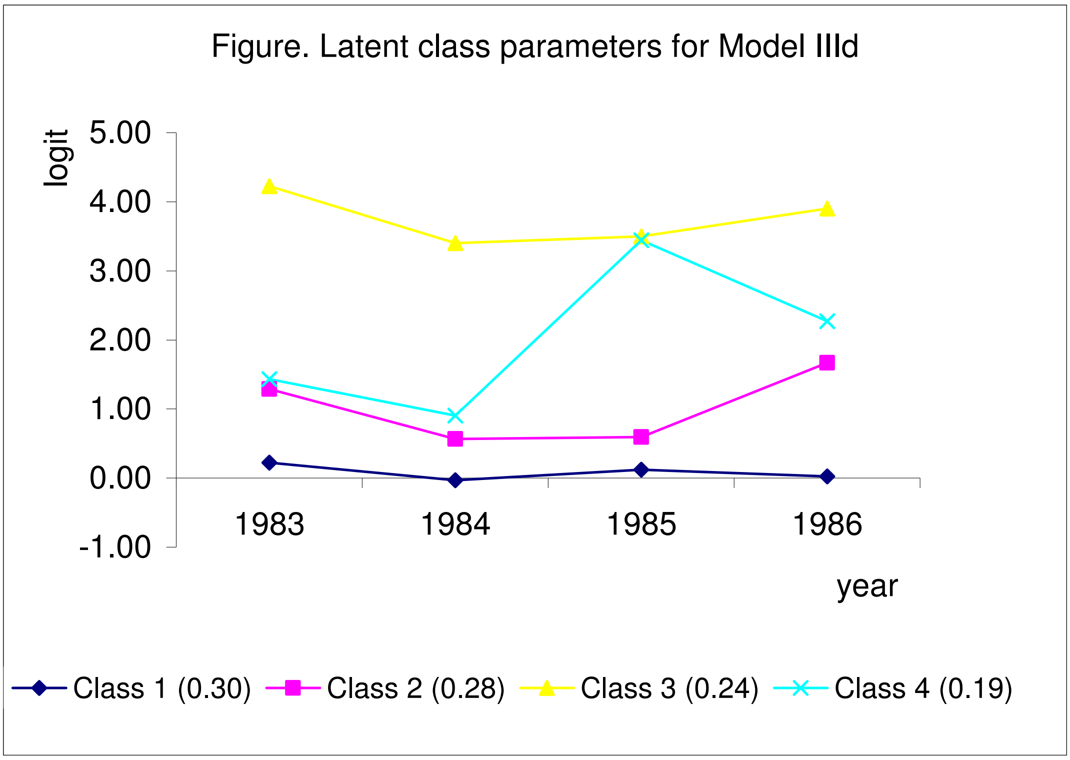
EXAMPLE: mixture binomial regression for longitudinal data

- Attitude towards abortion measured at 4 time points
- Class-independent predictor religion
- Non-parametric random-coefficient model
- See Vermunt & Van Dijk, 2001, Multilevel Modelling Newsletter

Fit measures

	LL	BIC	Npar
1-Class	-2188.38	4415.80	7
2-Class	-1745.42	3557.75	12
3-Class	-1682.71	3460.21	17
4-Class	-1656.72	3436.10	22
5-Class	-1645.20	3440.96	27

Time effects for 4 latent classes



9. LC Models for Choice Data

EXAMPLE: choice based conjoint experiment

- special case of LC regression
  - conditional logit model
  - choice-specific predictors

- 3 data files: responses, alternatives, choice sets

- responses on 8 choice tasks (3 alternatives + none)

- 400 cases

- effect of product attributes on choice
  - fashion (traditional/modern)
  - quality (low/high)
  - price (2-6)
  - none dummy

- purpose: find market segments for market simulations

10. Other Types of LC and FM Models

- Latent class scaling models
- Mixture variants of (regression) models for paired comparison, ranking, and other types of choice data
- Mixture variants of models for dependent/longitudinal data, such as Markov models and Rasch models
- Latent Markov models
- Loglinear models with latent variables
- Mixture variants of structural equation models

11. Some References on LC and FM models

- Hagenaars and McCutcheon (2002, eds.), "Applied Latent Class Analysis", Cambridge University Press.

- Agresti (2002) "Categorical Data Analysis", second edition, chapter 13. Wiley.

- McLachlan and Peel (2000). "Finite Mixture models". Wiley.

- Vermunt (1997). "Log-linear Models for Event Histories", chapters 3 and 5. Sage Publications.

- Papers on website www.latentclass.com.