Some Extensions of the Basic Latent Class Model

Jeroen K. Vermunt
Department of Methodology and Statistics, Tilburg University
www.jeroenvermunt.nl

Introduction

- So far, we discussed simple LC models for dichotomous and nominal indicators.
- We also looked at one important extension; that is, the inclusion of covariates.
- Some other extensions of the basic model are:
    1. Local dependencies
    2. Multiple latent variables (discrete factor models)
    3. Ordinal indicators
    4. Continuous indicators (latent profile analysis)
- In this video, I will discuss these extensions.

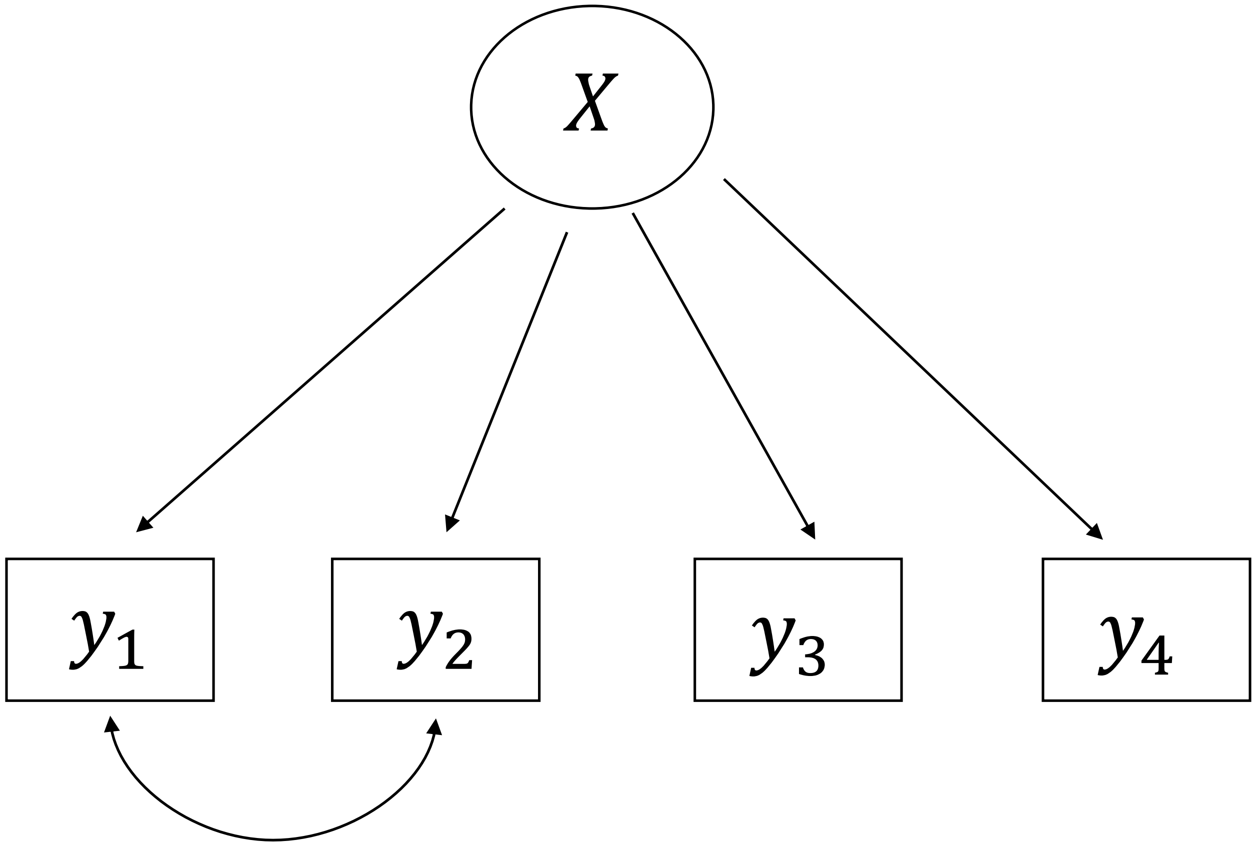
Local dependencies

- What is the reason a specified LC model may not fit the data well?
  - Because the local independence assumption is violated

- 4 possible solutions:
  1. increase the number of latent classes
  2. increase the number of latent variables (as in factor analysis)
  3. allow for local dependencies or direct effects between certain indicators
  4. remove certain indicators

Local dependencies

- A LC model for 4 indicators, where y_1 and y_2 are dependent within classes:



Local dependencies

- Equation:

$$P(y_1,\ldots,y_4)=\sum_{c=1}^{C}P(X=c)\,P(y_1,y_2\mid X=c)\,P(y_3\mid X=c)\,P(y_4\mid X=c)$$

with:

$$P(y_1,y_2\mid X=c)=\frac{\exp(\alpha^{1}_{y_1}+\alpha^{2}_{y_2}+\alpha^{12}_{y_1y_2}+\beta^{1}_{y_1c}+\beta^{2}_{y_2c})}{\sum_{m_1=1}^{M_1}\sum_{m_2=1}^{M_2}\exp(\alpha^{1}_{m_1}+\alpha^{2}_{m_2}+\alpha^{12}_{m_1m_2}+\beta^{1}_{m_1c}+\beta^{2}_{m_2c})}$$

- Term $\alpha^{12}_{y_1y_2}$ captures the association between y_1 and y_2 within classes.

Local dependencies

- Interpretation: two items have a stronger association than can be explained by the fact that persons belong to different latent classes

- Exploratory (post hoc) versus theoretical justification

- Detection: via bivariate residuals

Example: GSS82.sav data set

- Bivariate residuals in 2-cluster model

Indicators	purpose	accuracy	understa	cooperat
purpose	.			
accuracy	0.1063	.		
understa	0.3849	1.7099	.	
cooperat	0.2162	0.4147	58.4408	.

- Local dependency (or direct effect) between cooperation and understanding can be included in Latent GOLD via “Residuals” tab

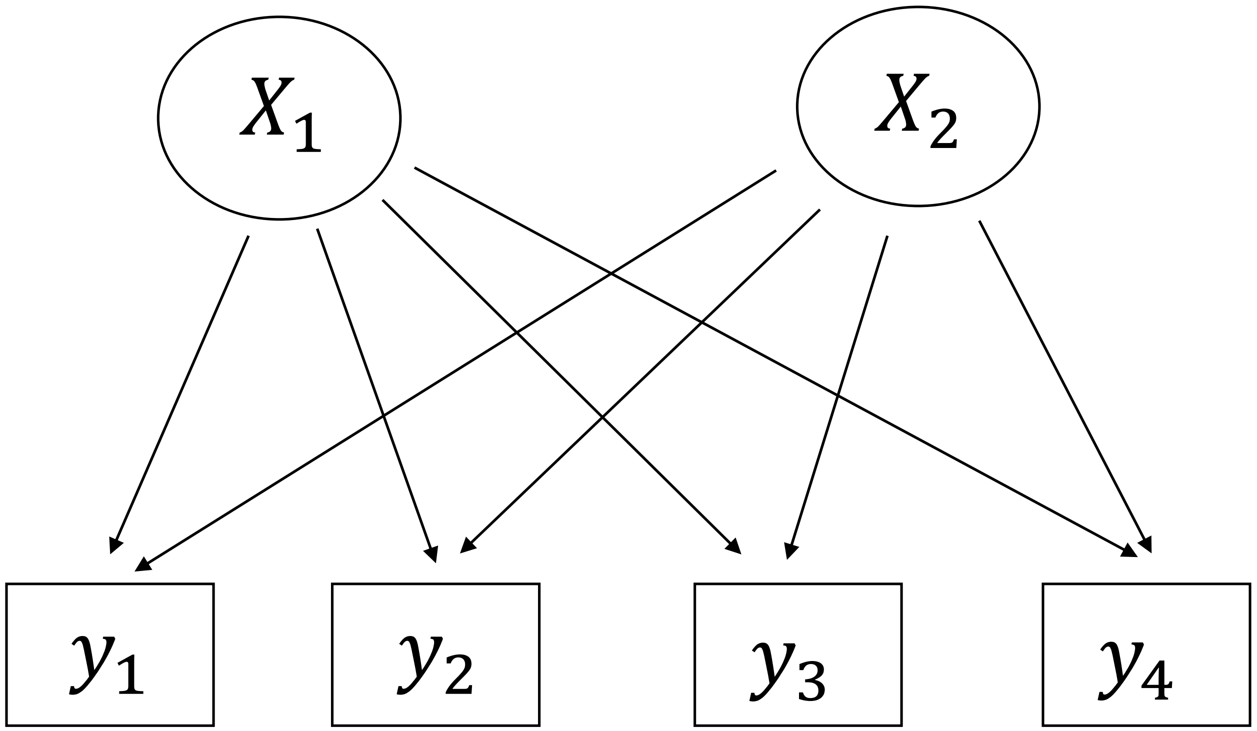
Example: GSS82.sav data set

- Fit measures of standard LC models and expanded models:

Model	BIC	L^2	df	p-value
1-cluster	8154.05	357.47	29	2.8e-58
2-cluster	7977.60	129.18	22	3.6e-17
3-cluster	7931.04	30.79	15	0.0095
4-cluster	7960.39	8.31	8	0.40
2-cluster + 1 direct effect	7877.11	13.88	20	0.84
2-dfactor “exploratory”	7912.23	11.98	15	0.68
2-dfactor “confirmatory”	7886.90	23.67	20	0.26

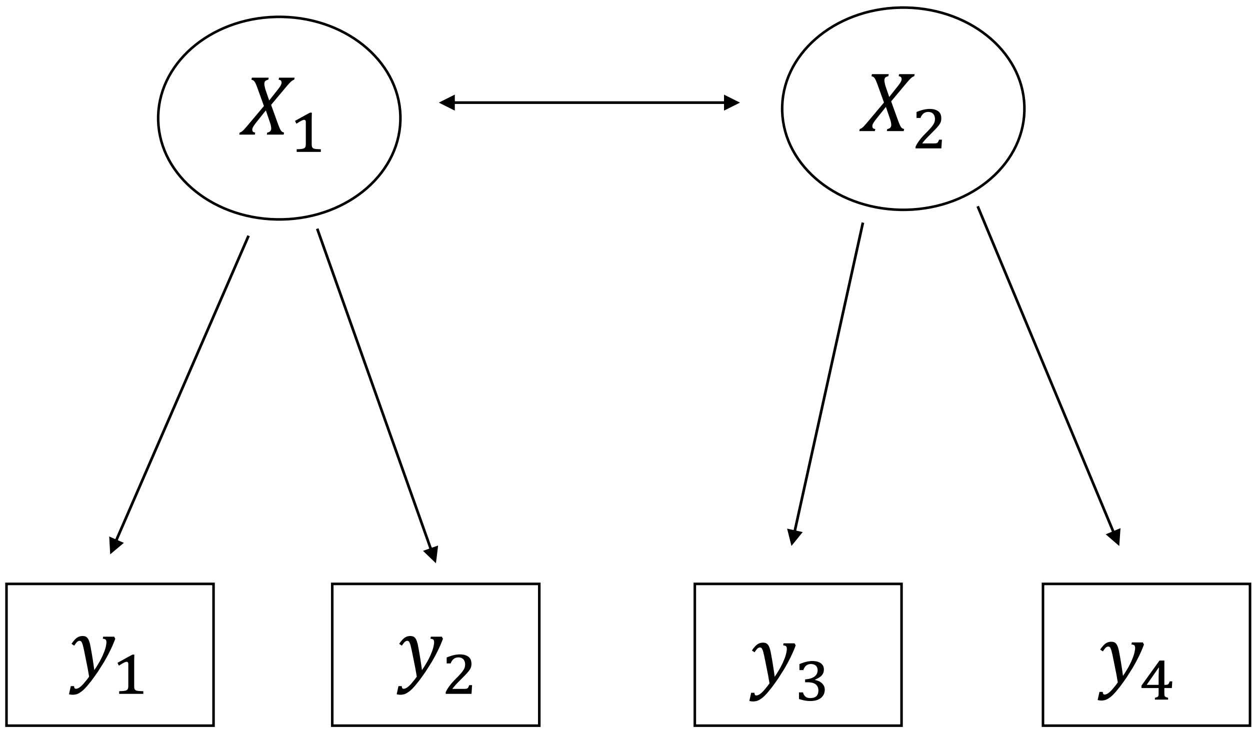
Multiple latent variables

- Exploratory 2-DFactor model:



Multiple latent variables

- Confirmatory 2-DFactor model:



Multiple latent variables

- Equation:

$$P(y_1,\ldots,y_J) = \sum_{c_1=1}^{C_1} \sum_{c_2=1}^{C_2} P(X_1 = c_1, X_2 = c_2) \prod_{j=1}^{J} P(y_j \mid X_1 = c_1, X_2 = c_2)$$

with:

$$P(y_j \mid X_1 = c_1, X_2 = c_2) = \frac{\exp(\alpha_{y_j}^{j} + \beta_{y_j}^{j1} c_1 + \beta_{y_j}^{j2} c_2)}{\sum_{m=1}^{M_j} \exp(\alpha_m^{j} + \beta_m^{j1} c_1 + \beta_m^{j2} c_2)}$$

- Confirmatory: some beta’s are set equal to zero

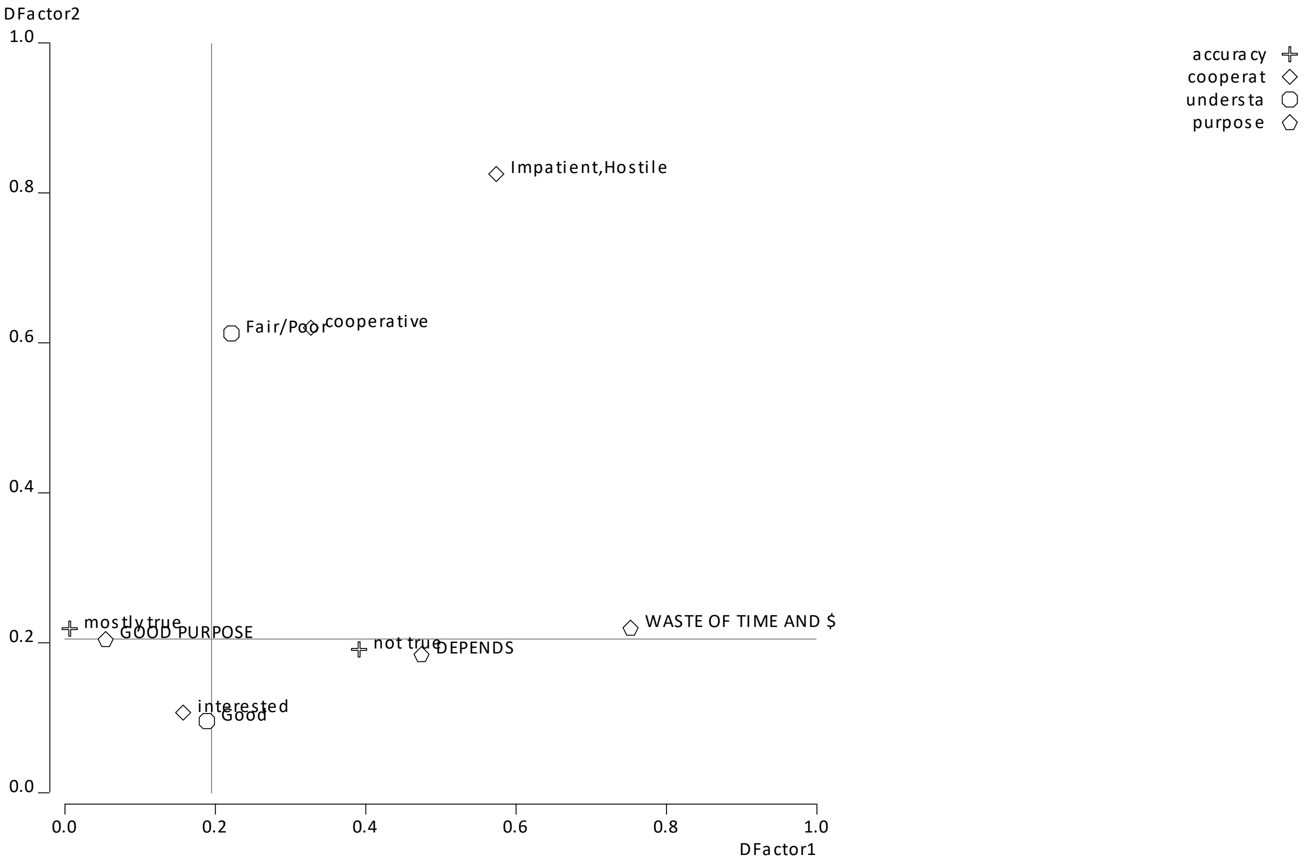
Example: GSS82.sav data set

- Fit measures of standard LC models and expanded models:

Model	BIC	L^2	df	p-value
1-cluster	8154.05	357.47	29	2.8e-58
2-cluster	7977.60	129.18	22	3.6e-17
3-cluster	7931.04	30.79	15	0.0095
4-cluster	7960.39	8.31	8	0.40
2-cluster + 1 direct effect	7877.11	13.88	20	0.84
2-dfactor "exploratory"	7912.23	11.98	15	0.68
2-dfactor "confirmatory"	7886.90	23.67	20	0.26

Example: GSS82.sav data set

- ProbMeans biplot of “exploratory” 2dfactor model



Ordinal indicators

- When an indicator has ordered categories, one may use this in the specification of its relationship with the latent variable(s).
- Latent GOLD uses an adjacent-category logit model which treats the indicator as numeric, but *discrete* (so no distributional assumptions).
- The beta parameters are restricted as follows: $\beta^{j}_{y_j c} = \beta^{j}_{c} y_j$.
  - This implies local odds-ratios are category independent.
- In Profile, you get class-specific means in addition to conditional probabilities for all categories.

Example: GSS82.sav data set

- Fit measures for nominal and ordinal models

Model		BIC	L^2	df	p-value
Nominal	1-Cluster	8154.05	357.47	29	2.8e-58
	2-Cluster	7977.60	129.18	22	3.6e-17
	3-Cluster	7931.04	30.79	15	0.0095
Ordinal	1-Cluster	8154.05	357.47	29	2.8e-58
	2-Cluster	7967.24	133.63	24	3.4e-17
	3-Cluster	7910.83	40.20	19	0.0031

Example: GSS82.sav data set

- Profile output of 3-class ordinal model

	Cluster1	Cluster2	Cluster3
Cluster Size	0.5827	0.2323	0.1850
Indicators			
accuracy			
mostly true	0.5999	0.6697	0.0217
not true	0.4001	0.3303	0.9783
Mean	1.4001	1.3303	1.9783
cooperat			
interested	0.9490	0.6638	0.6227
cooperative	0.0493	0.2651	0.2879
Impatient,Hostile	0.0017	0.0711	0.0894
Mean	1.0528	1.4073	1.4667
understa			
Good	0.9955	0.3141	0.7245
Fair/Poor	0.0045	0.6859	0.2755
Mean	1.0045	1.6859	1.2755
purpose			
GOOD PURPOSE	0.8737	0.9031	0.2436
DEPENDS	0.0766	0.0637	0.1423
WASTE OF TIME AND $	0.0496	0.0332	0.6141
Mean	1.1759	1.1301	2.3706

Continuous indicators: latent profile analysis

- Basic formula:

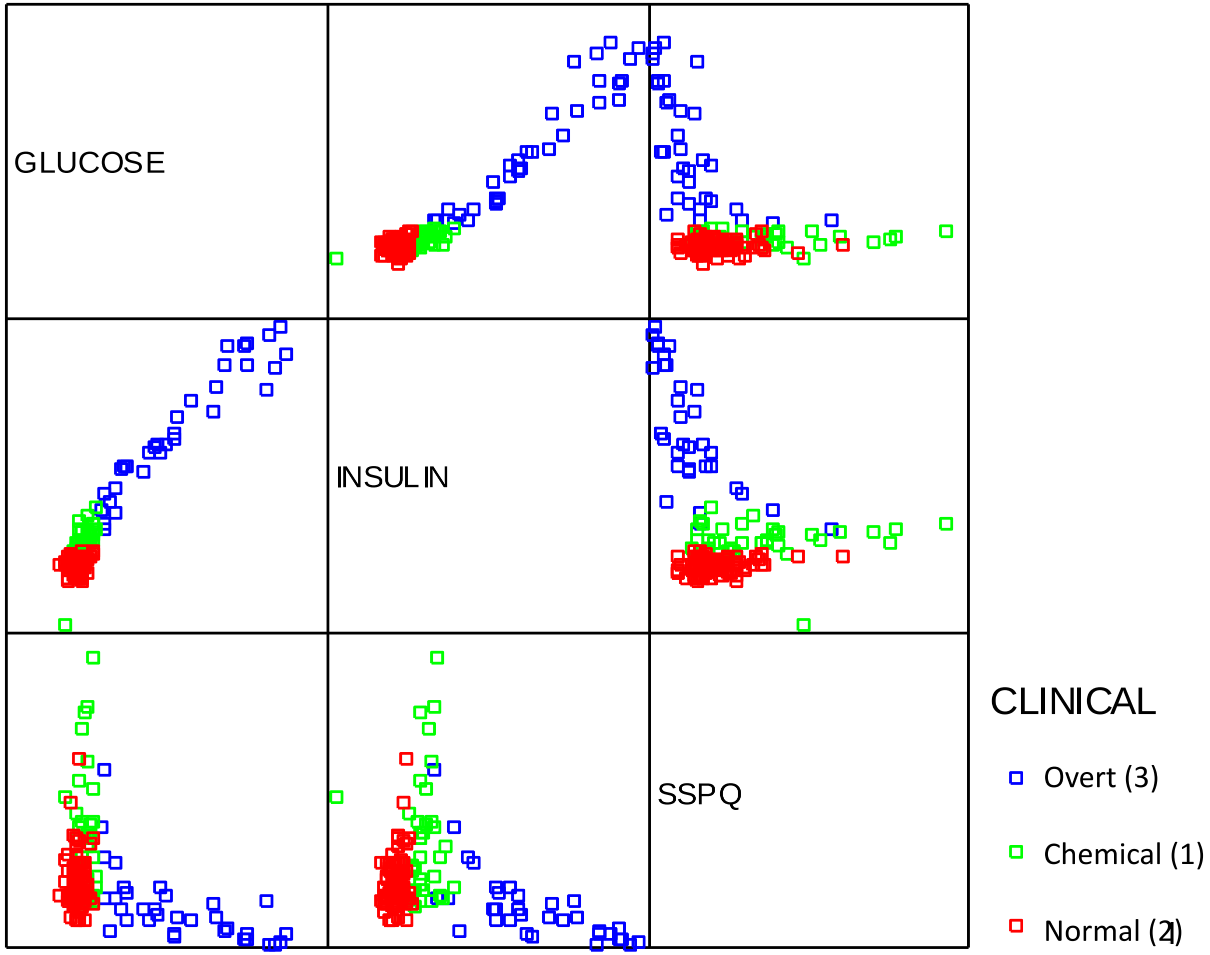
$$f(y_1,...,y_J) = \sum_{c=1}^{C} P(X=c) f(y_1,...,y_J \mid X=c) = \sum_{c=1}^{C} P(X=c) \prod_{j=1}^{J} f(y_j \mid X=c)$$

- Same basic assumptions as standard LC model. However, local independence (here, zero covariances) assumption must often be relaxed.
- $f(y_j \mid X=c)$ is a normal distribution, with a mean and variance that differs across latent classes. Note that normality within classes is a strong additional assumption compared to the LC models we have seen so far.

Example: diabetes.dat data set

- Data set consists of:
  - 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”
- Question:
  - Can we recover the clinical diagnosis with a latent class model using the three available measures?

Example: diabetes.dat data set



Example: diabetes.dat data set

- Fit measures for local independence models and models with one free covariance

Model	LL	BIC	Npar
Standard			
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
With y_1-y_2 covariance			
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

Example: diabetes.dat data set

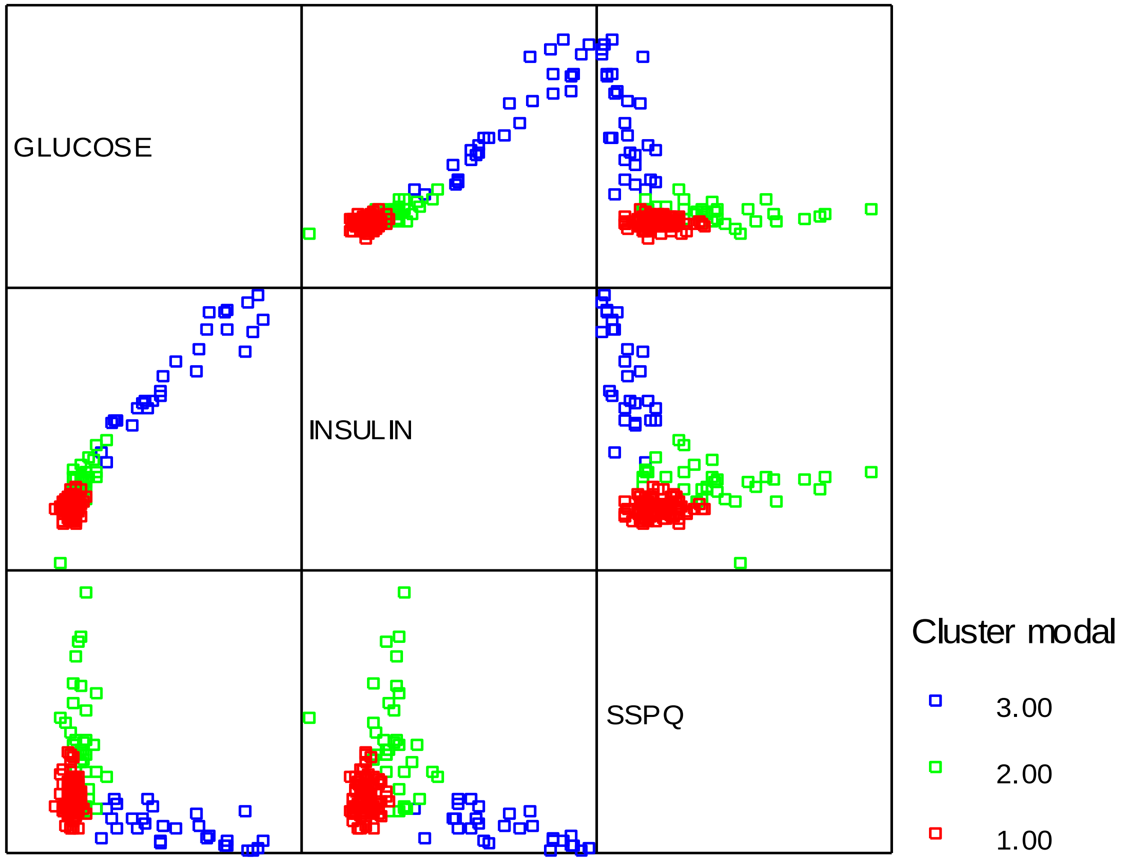
- Profile output of 3-cluster model with free y_1-y_2 covariance:

	Cluster1	Cluster2	Cluster3	Overall
Cluster Size	0.5391	0.2696	0.1913	
Indicators				
Glucose				
Mean	91.2315	104.0049	234.7598	122.1324
Insulin				
Mean	359.2211	495.0568	1121.0893	541.5922
SSPG				
Mean	163.1271	309.4323	76.9772	186.0971
Covariates				
true<l>				
1	0.1230	0.6638	0.0163	0.2484
2	0.8770	0.1859	0.0005	0.5230
3	0.0000	0.1503	0.9832	0.2286

- Other output you may look at are error variances, covariances, and correlations

Example: diabetes.dat data set

- Plot of modal classifications based on final 3-class model



Other extensions

- LC regression and LC growth models
  - LCs that differ in the parameters of a regression/growth model
- Latent Markov models
  - Transition between latent states over time
- Multilevel LC models
  - Latent classes at both the individual and the group level
- Models combining discrete and continuous latent variables
  - Mixture factor analysis
  - Mixture item response theory (IRT) models
  - LC regression/growth models with random effects