Introduction to Latent Class Analysis

Francesco Bartolucci

Department of Economics

University of Perugia (IT)

https://sites.google.com/site/bartstatistics/

francesco.bartolucci@unipg.it

Presentation at RISIS 2019 - Application of Latent Class Modelling to Research Policy and Higher Education Studies

September 9th, 2019

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

Outline

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

Heterogeneity

- Suppose we observe a *set of response variables* for n statistical units, with $\boldsymbol{Y}_i$ denoting the corresponding random vector for the i-th unit

- A statistical model allows us to account for the *heterogeneity* among the statistical units, which may be of two types:
  - *observed*: it may be explained on the basis of the observed covariates collected in vectors $\boldsymbol{X}_i$
  - *unobserved*: it cannot be explained on the basis of the observed covariates (it depends on factors that are not observed/observable)

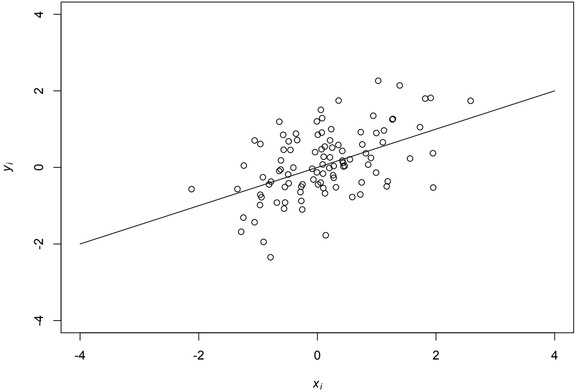
- *Standard statistical/econometric models* (in particular when one response variable is observed) only account for the observed heterogeneity

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

- This is the case of the *linear regression model* (for a single response variable):

$$Y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

  - $\boldsymbol{x}_i$: observed vector of covariates
  - $\boldsymbol{\beta}$: vector of regression coefficients
  - ε_i: error term



RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

- When $\boldsymbol{Y}_i$ is *a vector of more response variables* for each statistical unit, it is possible to account for the unobserved heterogeneity; typical situations:
  - *different response variables* are considered at the same time (e.g., different performance indicators)
  - *repeated observations of the same response variable* at different time occasions (longitudinal/panel data)
  - *(1st level) units are grouped in clusters*, which are 2nd level units (multilevel data)
- Almost all statistical/econometric models that account for unobserved heterogeneity may be cast in the class of *Latent Variable Models (LVMs)*, among which the *Latent Class (LC) model* is very important
- LVMs may also be used to *account for measurement errors* or *summarizing different measurements*
- *Main references*: Skrondal & Rabe-Hesketh (2004), Bartholomew *et al.* (2011)

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

Latent Variable Models

- *Latent variables* ($\boldsymbol{U}_i$) are unobservable variables supposed to exist and to affect $\boldsymbol{Y}_i$; these may be correlated with $\boldsymbol{X}_i$

- An LVM formulates *assumptions* on:
  - the conditional distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{U}_i$ and $\boldsymbol{X}_i$, $f(\boldsymbol{y}_i|\boldsymbol{u}_i, \boldsymbol{x}_i)$ (*measurement model*)
  - the conditional distribution of $\boldsymbol{U}_i$ given $\boldsymbol{X}_i$, $f(\boldsymbol{u}_i|\boldsymbol{x}_i)$ (*structural model*)

- A common assumption of LVMs is that of *local independence (LI)*, according to which the response variables are conditionally independent given the latent variables and the covariates:

$$f(\boldsymbol{y}_i|\boldsymbol{u}_i, \boldsymbol{x}_i) = \prod_{j=1}^{J} f(y_{ij}|\boldsymbol{u}_i, \boldsymbol{x}_i)$$

  - J: number of response variables
  - y_{ij}: single element of $\boldsymbol{y}_i$

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

- By marginalizing out the latent variables we obtain the *manifest distribution*:

$$f(\boldsymbol{y}_i|\boldsymbol{x}_i) = \int f(\boldsymbol{y}_i|\boldsymbol{u}, \boldsymbol{x}_i) f(\boldsymbol{u}|\boldsymbol{x}_i) d\boldsymbol{u}$$

- By the Bayes theorem we obtain the *posterior distribution*

$$f(\boldsymbol{u}_i|\boldsymbol{x}_i, \boldsymbol{y}_i) = \frac{f(\boldsymbol{y}_i|\boldsymbol{u}_i, \boldsymbol{x}_i) f(\boldsymbol{u}_i|\boldsymbol{x}_i)}{f(\boldsymbol{y}_i|\boldsymbol{x}_i)}$$

  that is used for predicting the latent variables on the basis of the manifest variables

- *Estimation* is typically based on the maximum likelihood approach relying on numerical algorithms, among which the Newton-Raphson and particularly the Expectation-Maximization (EM) are very popular

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

- LVMs may be *classified* according to:
  - *type of response variables* (discrete, continuous, categorical, etc.)
  - *type of latent variables* (discrete or continuous)
  - *presence or absence of covariates* (that may be included in different ways)
- LVMs based on discrete latent variables are of particular interest as they permit:
  - to naturally *group units in homogeneous latent clusters*, also named latent groups or latent classes (model-based clustering)
  - to account for the *unobserved heterogeneity* in a nonparametric way (it is necessary to specify a parametric distribution for the latent variables)
- The *LC model* is one of the most important LVMs based on discrete latent variables and may be seen as a *Finite Mixture (FM) model* for categorical response variables

RISIS RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

Expectation-Maximization algorithm

- This is a general approach for maximum likelihood estimation in the presence of *missing data* (Dempster *et al.*, 1977)
- In our context, missing data correspond to the latent variables; then:
  - *incomplete (observable) data*: covariates and response variables $(\boldsymbol{X}, \boldsymbol{Y})$
  - *complete (unobservable) data*: incomplete data + latent variables $(\boldsymbol{U}, \boldsymbol{X}, \boldsymbol{Y})$
- The corresponding *log-likelihood functions* are:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\boldsymbol{y}_i|\boldsymbol{x}_i)$$

$$\ell^*(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log[f(\boldsymbol{y}_i|\boldsymbol{U}_i, \boldsymbol{x}_i) f(\boldsymbol{U}_i|\boldsymbol{x}_i)]$$

  - $\boldsymbol{\theta}$: overall vector of model parameters

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

- The EM algorithm maximizes $\ell(\boldsymbol{\theta})$ by alternating *two steps* until convergence (h=iteration number):
  - *E-step*: compute the expect value of $\ell^*(\boldsymbol{\theta})$ given the current parameter value $\boldsymbol{\theta}^{(h-1)}$ and the observed data, obtaining

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(h-1)}) = E[\ell^*(\boldsymbol{\theta})|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\theta}^{(h-1)}]$$

  - *M-step*: maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(h-1)})$ with respect to $\boldsymbol{\theta}$ obtaining $\boldsymbol{\theta}^{(h)}$

- *Convergence is* checked on the basis of the difference

$$\ell(\boldsymbol{\theta}^{(h)}) - \ell(\boldsymbol{\theta}^{(h-1)}) \quad \text{or} \quad \|\boldsymbol{\theta}^{(h)} - \boldsymbol{\theta}^{(h-1)}\|$$

- The algorithm is usually *easier to implement and much more stable* with respect to the Newton-Raphson algorithm, but it may be much slower

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

Finite Mixture Model

- *Main references*: Lindsay (1995), McLachlan & Peel (2000), Bouveyron *et al.* (2019)
- The underlying idea is that statistical units come from different groups, where the grouping is unobserved (*latent groups or clusters*)
- *Model assumptions*:
    - there exist unit-specific *discrete latent variables* U_i, $i = 1, \dots, n$, with the same finite distribution with k levels defining the groups
    - the groups have *prior probabilities (weights)* $\pi_u = p(U_i = u)$, $u = 1, \dots, k$
    - for each group we have a specific *conditional response distribution* $f(\boldsymbol{y}_i|u) = f(\boldsymbol{y}_i|U_i = u)$, $u = 1, \dots, k$
- An FM model may include or not *individual covariates*; these may directly affect the measurement model or the structural model

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

- An advantage of FM models with respect to LVMs based on continuous latent variables is that manifest and posterior distributions may be *explicitly computed*

- The *manifest distribution* is a weighted average of density or probability mass functions:

$$f(\boldsymbol{y}_i) = \sum_{u=1}^{k} f(\boldsymbol{y}_i|u)\pi_u$$

- By the Bayes theorem we obtain the *posterior distribution*

$$p(U_i = u|\boldsymbol{y}_i) = \frac{f(\boldsymbol{y}_i|u)\pi_u}{f(\boldsymbol{y}_i)}, \quad u = 1, \dots, k,$$

that is used to assign each unit to a specific latent group on the basis of the manifest variables

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

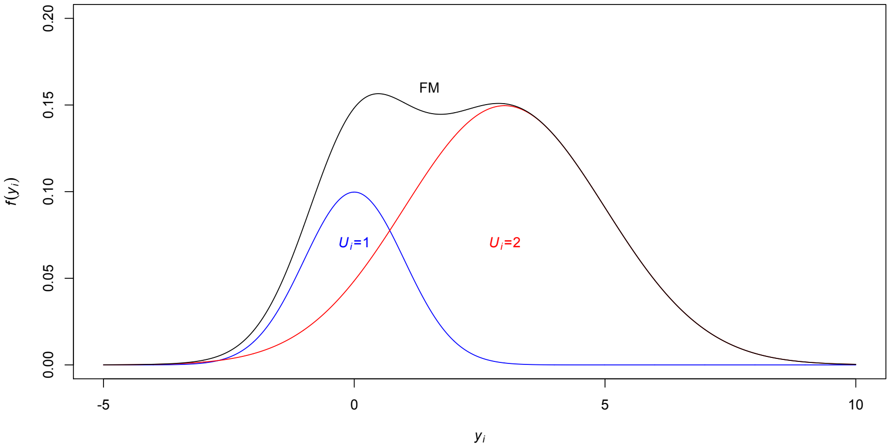
- *Example*: for a single response variable, $k = 2$ components exist with Normal distribution with specific means and variances and different weights:
  - $U_i = 1 \; \rightarrow \; Y_i \sim N(0, 1)$
  - $U_i = 2 \; \rightarrow \; Y_i \sim N(3, 4)$
  - $\pi_1 = 0.25$, $\pi_2 = 0.75$

- Through the general rule for LVMs we obtain the *manifest distribution* of Y_i:

$$f(y_i) = 0.25\,\phi(y_i; 0, 1) + 0.75\,\phi(y_i; 3, 4)$$

  - $\phi(y_i; \mu, \sigma^2)$: density function of the Normal distribution with mean μ and variance σ^2

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES



- Apart from model-based clustering, an FM model is a valid approach for *density estimation* given its flexibility (able to easily reproduce skewed and multimodal distributions)

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

- The *posterior distribution* of U_i may be found by the Bayes theorem:

$$p(U_i = 1|y_i) = \frac{0.25\,\phi(y_i;0,1)}{0.25\,\phi(y_i;0,1) + 0.75\,\phi(y_i;3,4)}$$

$$p(U_2 = 1|y_i) = \frac{0.75\,\phi(y_i;3,4)}{0.25\,\phi(y_i;0,1) + 0.75\,\phi(y_i;3,4)}$$

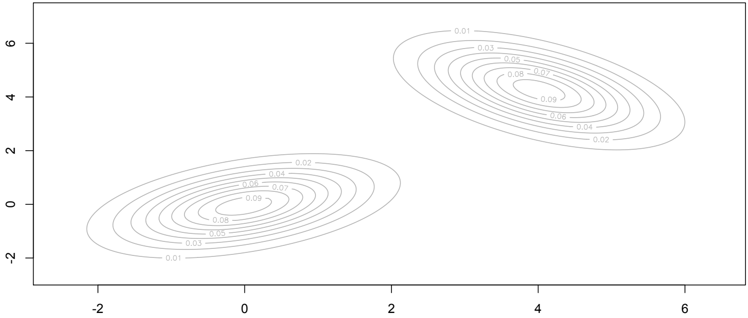
- Units are assigned to the latent groups on the basis of the *Maximum A-Posterior (MAP) rule*; example:

	U_i		Assigned
y_i	1	2	group
-3	0.400	**0.600**	2
0	**0.673**	0.327	1
2	0.093	**0.907**	2
4	0.000	**1.000**	2

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

- The term *finite mixture model* may be used for the general discrete LV approach, although it is typically used for continuous data

- For *continuous data*, $f(\boldsymbol{y}_i|u)$ are typically multivariate Normal distributions (*FM of Normal distributions*) with specific mean vectors $\boldsymbol{\mu}_u$ and variance-covariance matrices $\boldsymbol{\Sigma}_u$



- For *categorical data* we obtain the LC model that has peculiarities in terms of assumptions and its estimation

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

Estimation of FM models

- *Estimation* of an FM model is based on the maximum likelihood approach that is easily carried out through the Expectation-Maximization (EM) algorithm

- To introduce the EM algorithm it is convenient to substitute each latent variable U_i with the *(binary) indicator variables* $Z_{i1}, \dots, Z_{ik}$, where $U_i = u$ iff $Z_{iu} = 1$ with all other variables equal to 0

- *Example* with $k = 4$:

U_i	Z_{i1}	Z_{i2}	Z_{i3}	Z_{i4}
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

- In this way the *complete data log-likelihood* is expressed as

$$\ell^*(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{u=1}^{k}[Z_{iu}\log f(\boldsymbol{y}_i|u) + Z_{iu}\log \pi_u]$$

- The EM algorithm consists in alternating two steps:
  - *E-step*: compute the posterior expect value of each indicator variable Z_{iu} by the Bayes theorem:

$$\hat{z}_{iu} = E(Z_{iu}|\boldsymbol{y}_i) = p(Z_{iu} = 1|\boldsymbol{y}_i) = p(U_i = u|\boldsymbol{y}_i)$$

  - *M-step*: maximize function $\ell^*(\boldsymbol{\theta})$ with each indicator variables Z_{iu} substituted by $\hat{z}_{iu}$ obtained at the E-step

- At the M-step, an *explicit solution* exists for the class weights:

$$\pi_u = \frac{1}{n}\sum_{i=1}^{n}\hat{z}_{iu}, \quad u = 1,\dots,k$$

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

- For the *FM of Normal distributions*, an explicit solution exists for the other parameters ($u = 1, \dots, k$):

$$\boldsymbol{\mu}_u = \frac{1}{\sum_{i=1}^n \hat{z}_{iu}} \sum_{i=1}^n \hat{z}_{iu} \boldsymbol{y}_i$$

$$\boldsymbol{\Sigma}_u = \frac{1}{\sum_{i=1}^n \hat{z}_{iu}} \sum_{i=1}^n \hat{z}_{iu}(\boldsymbol{y}_i - \boldsymbol{\mu}_u)(\boldsymbol{y}_i - \boldsymbol{\mu}_u)' \quad \text{(homoskedastic)}$$

$$\boldsymbol{\Sigma} = \frac{1}{n} \sum_{i=1}^n \sum_{u=1}^k \hat{z}_{iu}(\boldsymbol{y}_i - \boldsymbol{\mu}_u)(\boldsymbol{y}_i - \boldsymbol{\mu}_u)' \quad \text{(heteroskedastic)}$$

- In general, different starting values must be used in order to face the problem of the *multimodality of the log-likelihood function*

- The EM algorithm is *implemented* in several softwares such as `R` (package `mclust`) and `Stata`

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

- In order to *select the number of components* (k) two criteria can be adopted when there is no precise idea based on substantial reasons:

$$\begin{aligned}\text{Akaike Information Criterion (AIC)} &= -2\ell(\hat{\boldsymbol{\theta}}_k) + 2 \times \#\text{par.}\\ \text{Bayesian Information Criterion (BIC)} &= -2\ell(\hat{\boldsymbol{\theta}}_k) + \log(n) \times \#\text{par.}\end{aligned}$$

- These criteria rely on *penalized versions* of the log-likelihood function (Akaike, 1973; Schwarz,1978): the selected model is that with the minimum value of AIC (or BIC), corresponding to the best compromise between goodness-of-fit and model parsimony

- Certain authors prefer to report AIC (or BIC) with *reversed sign* (search for the maximum)

- BIC often selects a *more parsimonious model* with respect to AIC

- Many *other selection criteria* are available and these may be used in general (not only to select k)

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

Example 1: simulated data

- Consider an FM model for $n = 100$ *bivariate responses* with $k = 2$ components:
  - $U_i = 1 \rightarrow \boldsymbol{Y}_i \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$
  - $U_i = 2 \rightarrow \boldsymbol{Y}_i \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$
  - $\pi_1 = \pi_2 = 0.5$

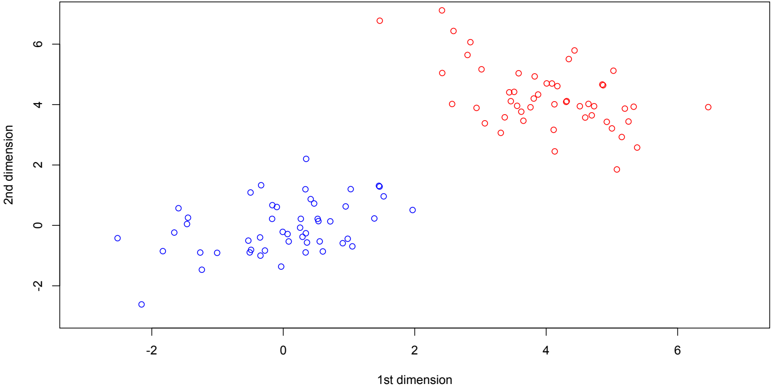
$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1.0 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 4 \\ 4 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1.0 & -0.5 \\ -0.5 & 1 \end{pmatrix}$$

- This is a *heterockedatic FM model* because the two variance-covariance matrices are different

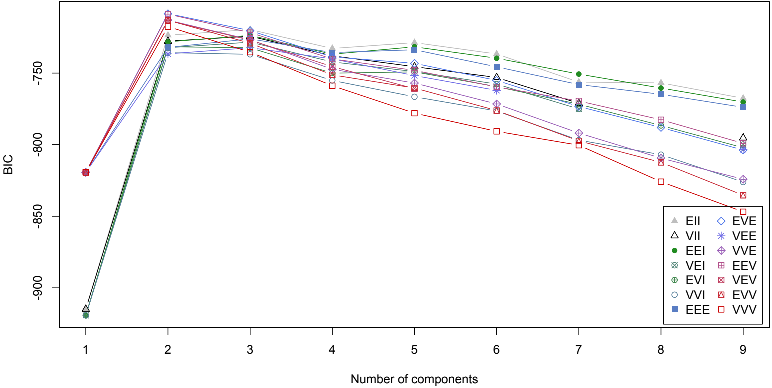
RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

- *Data representation* (blue = 1st component, red = 2nd component):



- These data are analyzed by package `mclust` *of* R

- *Model selection* based on BIC (with sign reversed) in terms of k and structure of the variance-covariance matrices (EII, VII,...):



- *Two groups* are indeed selected, with different variance-covariance matrices

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

- *Estimated parameters*:

$$\hat{\boldsymbol{\mu}}_1 = \begin{pmatrix} -0.015 \\ -0.058 \end{pmatrix}, \quad \hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 1.006 & 0.403 \\ 0.403 & 0.835 \end{pmatrix}$$

$$\hat{\boldsymbol{\mu}}_2 = \begin{pmatrix} 4.012 \\ 4.251 \end{pmatrix}, \quad \hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 0.869 & -0.519 \\ -0.519 & 1.089 \end{pmatrix}$$

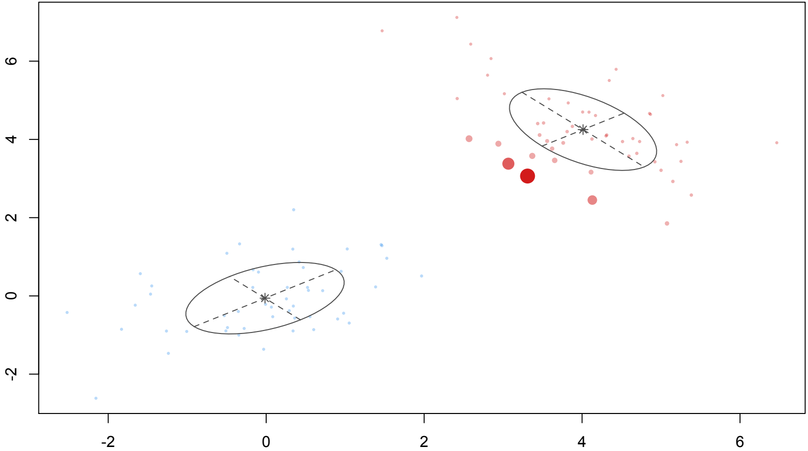
$$\hat{\pi}_1 = \hat{\pi}_2 = 0.5$$

- *Posterior probabilities and clustering*:

i	$\hat{z}_{i1}$	$\hat{z}_{i2}$	Assigned group	True group
1	1.0000	0.0000	1	1
⋮	⋮	⋮	⋮	⋮
70	0.0000	1.0000	2	2
⋮	⋮	⋮	⋮	⋮
74	0.0027	0.9973	2	2

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

- *Clustering representation* (blue = 1st component, red = 2nd component) with measure of uncertainty:



Example 2: real data

- Data about all world countries regarding certain *macro-economic and demographic indicators* for 2016 (source: World Bank):

i	Name	Code	GDP	Ages	Life expect	School enp	School ens
1	Afghanistan	AFG	1793.89	43.86	65.02	NA	51.75
2	Albania	ALB	11355.62	17.72	80.45	109.78	94.98
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
85	Italy	ITA	34655.26	13.61	84.90	100.36	102.83
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
170	Switzerland	CHE	57421.55	14.83	85.10	104.41	102.29
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
186	United States	USA	53399.36	19.03	81.20	101.36	98.77
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
236	Sub-Saharan Africa (IDA & IBRD)	TSS	3482.87	42.88	62.09	97.18	42.81

- Countries with at least one *missing value* are eliminated so that $n = 162$ countries are considered

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

- A model with $k = 3$ components and unequal variance-covariance matrices (VEV structure) are *selected by BIC*

- *Estimated means and weights*:

$$\hat{\boldsymbol{\mu}}_1 = \begin{pmatrix} 3428.67 \\ 39.38 \\ 64.18 \\ 104.13 \\ 49.33 \end{pmatrix}, \quad \hat{\boldsymbol{\mu}}_2 = \begin{pmatrix} 16661.98 \\ 24.01 \\ 77.34 \\ 102.93 \\ 93.48 \end{pmatrix}, \quad \hat{\boldsymbol{\mu}}_3 = \begin{pmatrix} 51140.18 \\ 16.67 \\ 83.23 \\ 102.22 \\ 113.31 \end{pmatrix}$$

$$\hat{\pi}_1 = 0.265, \quad \hat{\pi}_2 = 0.529, \quad \hat{\pi}_3 = 0.205$$

- The *estimated variances* have a size that tend to increase with the mean and, generally, with a negative correlation between variable Ages and the other variables

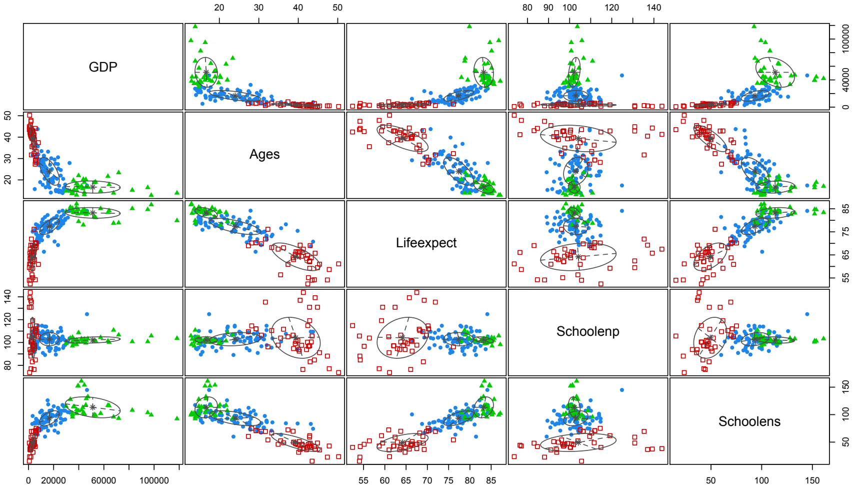
RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

- *Posterior probabilities and clustering*:

i	Code	$\hat{z}_{i1}$	$\hat{z}_{i2}$	$\hat{z}_{i3}$	Assigned group
2	ALB	0.006	0.988	0.007	2
⋮	⋮	⋮	⋮	⋮	⋮
85	ITA	0.000	0.199	0.801	3
⋮	⋮	⋮	⋮	⋮	⋮
170	CHE	0.000	0.000	1.000	3
⋮	⋮	⋮	⋮	⋮	⋮
186	USA	0.000	0.000	1.000	3
⋮	⋮	⋮	⋮	⋮	⋮
236	TSS	1.000	0.000	0.000	1

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

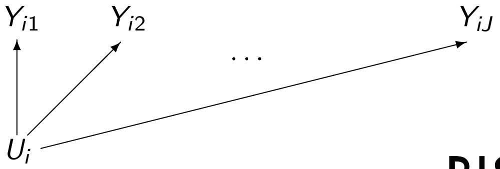
- *Clustering representation* (red = 1st component, blue = 2nd component, green = 3rd component):



RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

Latent Class Model

- *Main references*: Lazarfeld (1968), Goodman (1974)
- It follows an FM approach that:
  - is suitable for *categorical response variables* Y_{ij} with categories labeled from 0 to $c_j - 1$, $j = 1, \ldots, J$
  - *assumes LI* so that the latent variable is the only explanatory factor of the responses



- The *conditional probability of a response configuration* given the latent class is obtained by a single product:

$$p(\boldsymbol{y}_i|u) = \prod_{j=1}^{J} \eta_{j,y_{ij}|u}$$

  - $\eta_{j,y|u}$: probability that $Y_{ij} = y$ given $U_i = u$

- In the *binary case* it may be expressed using the Bernoulli probability mass function:

$$p(\boldsymbol{y}_i|u) = \prod_{j=1}^{J} \lambda_{j|u}^{y_{ij}} (1 - \lambda_{j|u})^{1-y_{ij}}$$

  - $\lambda_{j|u}$: probability that $Y_{ij} = 1$ given $U_i = u$ corresponding to $\eta_{j,1|u}$

- The *manifest distribution* becomes

$$p(\boldsymbol{y}_i) = \sum_{u=1}^{k} \left(\prod_{j=1}^{J} \eta_{j,y_{ij}|u} \right) \pi_u$$

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

- The *posterior distribution* becomes

$$p(U_i = u|\boldsymbol{y}_i) = \frac{\sum_{u=1}^{k}\left(\prod_{j=1}^{J}\eta_{j,y_{ij}|u}\right)\pi_u}{p(\boldsymbol{y}_i)}, \quad u = 1,\ldots,k$$

- The *number of free parameters* is in general

$$\#\mathrm{par} = \underbrace{k-1}_{\pi_u} + \underbrace{\prod_{j=1}^{J}(c_j - 1)}_{\eta_{j,y|u}}$$

that in the binary case becomes

$$\#\mathrm{par} = \underbrace{k-1}_{\pi_u} + \underbrace{kJ}_{\lambda_{j|u}}$$

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

Estimation

- Estimation is based on an *EM algorithm* having the same structure of that for FM models; it may be easily implemented using the binary indicator variable representation (Z_{iu} vs U_i)

- The *complete data log-likelihood* is expressed as

$$\ell^*(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{u=1}^{k}\left[Z_{iu}\sum_{j=1}^{J}\log(\eta_{j,y_{ij}|u}) + Z_{iu}\log(\pi_u)\right]$$

- At the M-step an *explicit solution* exists for the $\eta_{j,y|u}$ parameters:

$$\eta_{j,y|u} = \frac{1}{\sum_{i=1}^{n}\hat{z}_{iu}}\sum_{i=1}^{n}\hat{z}_{iu}I(y_{ij}=y)$$

that in the binary case becomes

$$\lambda_{j|u} = \frac{1}{\sum_{i=1}^{n}\hat{z}_{iu}}\sum_{i=1}^{n}\hat{z}_{iu}y_{ij}$$

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

- A crucial issue is still that of *multimodality of the log-likelihood function* that requires the use of different starting points

- Typically, apart from a deterministic initialization (depending on the observed data), several *random initializations* are tried with each probability π_u and $\eta_{j,y|u}$ drawn from a uniform distribution from 0 to 1 and then suitably renormalized

- A crucial point after estimation is that of *assigning individuals to the latent classes*, which is still based on the posterior probabilities $\hat{z}_{iu}$ using the MAP rule

- The same *model selection criteria* as for FM models (in particular AIC and BIC) are typically adopted for model selection

- The EM algorithm is *implemented* in several softwares such as `R` (package `MultiLCIRT`) and `Stata`

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

Example

- *Data* are collected on 216 subjects who responded to $J = 4$ items concerning the behavior in certain role conflict situations (Goodman, 1974)
- Each binary *response variable* is equal to 1 if the interviewed individual has a universalistic behavior and 0 if he/she has a particularistic behavior
- Data may be represented by a 2^4-dimensional *vector of frequencies* for all the response configurations:

Y_{i1}	Y_{i2}	Y_{i3}	Y_{i4}	Frequency
0	0	0	0	42
0	0	0	1	23
0	0	1	0	6
⋮	⋮	⋮	⋮	⋮
1	1	1	1	20

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

- *Selection of the number of classes*:

k	$\ell(\hat{\boldsymbol{\theta}})$	#par	AIC	BIC
1	-543.65	4	1095.30	1108.80
2	-504.47	9	1026.94	1057.31
3	**-503.30**	**14**	**1034.60**	**1081.86**
4	-503.11	19	1044.22	1108.35

- Both AIC and BIC select *$k = 3$ latent classes*

- *Parameter estimates*:

	$\hat{\lambda}_{j\|u}$				
Class (u)	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$\hat{\pi}_u$
1	0.003	0.023	0.006	0.101	0.200
2	0.164	0.519	0.563	0.807	0.601
3	0.548	0.922	0.736	0.928	0.199

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

- *Posterior probability* for each possible response configuration:

y_{i1}	y_{i2}	y_{i3}	y_{i4}	$\hat{z}_{i1}$	$\hat{z}_{i2}$	$\hat{z}_{i3}$	Assigned class
0	0	0	0	**0.894**	0.105	0.001	1
0	0	0	1	0.183	**0.801**	0.016	2
0	0	1	0	0.040	**0.946**	0.013	2
0	0	1	1	0.001	**0.957**	0.042	2
0	1	0	0	0.150	**0.793**	0.057	2
0	1	0	1	0.004	**0.815**	0.180	2
0	1	1	0	0.001	**0.865**	0.134	2
0	1	1	1	0.000	**0.676**	0.324	2
1	0	0	0	0.105	**0.860**	0.035	2
1	0	0	1	0.003	**0.887**	0.110	2
1	0	1	0	0.001	**0.919**	0.080	2
1	0	1	1	0.000	**0.787**	0.213	2
1	1	0	0	0.002	**0.693**	0.305	2
1	1	0	1	0.000	0.423	**0.577**	3
1	1	1	0	0.000	**0.512**	0.488	2
1	1	1	1	0.000	0.253	**0.747**	3

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

Inclusion of covariates

- Two possible *choices to include individual covariates* collected in $\boldsymbol{x}_i$
- The first is in the *measurement model* so that we have random intercepts; for instance, for the LC model for binary variables we could assume:

$$\lambda_{ij|u} = p(Y_{ij} = 1 | U_i = u, \boldsymbol{x}_i),$$

$$\log \frac{\lambda_{ij|u}}{1 - \lambda_{ij|u}} = \alpha_u + \boldsymbol{x}_i' \boldsymbol{\beta}, \quad i = 1, \ldots, n, \; j = 1, \ldots, J, \; u = 1, \ldots, k$$

  - α_u: random intercepts
  - $\boldsymbol{\beta}$: vector of logistic regression parameters

- The latent variables are used to account for the *unobserved heterogeneity* in addition to the observed heterogeneity; the model may be seen as a "discrete version" of the random-effects logit model

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

- The second is in the *structural model* governing the distribution of the latent variables (via a multinomial logit parametrization):

$$\pi_{iu} = p(U_i = u | \boldsymbol{x}_i),$$
$$\log \frac{\pi_{iu}}{\pi_{i1}} = \boldsymbol{x}'_{i1}\gamma_u, \quad u = 2, \ldots, k$$

  - γ_u: vectors of regression coefficients specific for each latent class

- The *main interest is in the latent variable* that is measured through the observable response variables (e.g., health status) and on how this latent variable depends on the covariates

- Both extensions lead to FM/LC models that may be *estimated* by an EM algorithm having a structure similar to that of the corresponding models without covariates; particular care is necessary to obtain the *standard errors* for the regression coefficients

- Usual criteria may be used for *model selection* in terms of number of components (k) and other assumptions

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

Extension to longitudinal data

- The extension of FM/LC models to the analysis of longitudinal data is known as *Latent Markov (ML) model*

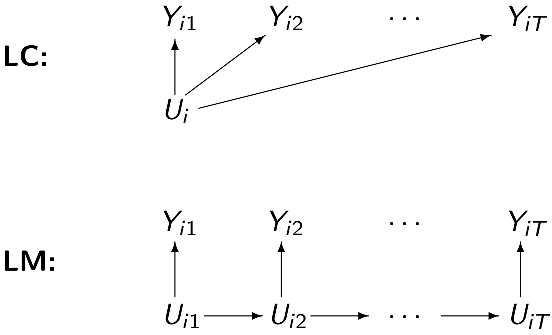
- *Main references*: Wiggins (1973), Bartolucci *et al.* (2013), Zucchini *et al.* (2016)

- For individual sequences of response variables at T time occasions, $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iT})'$, $i = 1, \ldots, n$, the *basic version of the LM model* assumes that:
  - (*LI*) the response variables in $\boldsymbol{Y}_i$ are conditionally independent given a latent process $\boldsymbol{U}_i = (U_{i1}, \ldots, U_{iT})'$
  - every latent process $\boldsymbol{U}_i$ follows a *first-order Markov chain* with state space $\{1, \ldots, k\}$, initial probabilities π_u, and transition probabilities $\pi_{v|u}$, $u, v = 1, \ldots, k$

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

Possible interpretation

- The LM model may be seen as a *generalization of the LC model* in which subjects are allowed to move between latent classes



RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

Model parameters

- Each latent state u ($u = 1, \dots, k$) corresponds to a *class of subjects* (or latent state) in the population, and is characterized by:
  - *initial probability*:
  $$\pi_u = p(U_{i1} = u)$$
  - *transition probabilities* (which may also be time-specific in the non-homogenous case):
  $$\pi_{v|u} = p(U_{it} = v|U_{i,t-1} = u), \quad t = 2, \dots, T, \; v = 1, \dots, k$$
  - *distribution of the response variables* (with categorical responses with c categories):
  $$\psi_{y|u} = p(Y_{it} = y|U_{it} = u), \quad t = 1, \dots, T, \; y = 0, \dots, c-1$$
- The transition probabilities are collected in the *transition matrix* $\mathbf{\Pi}$ of size $k \times k$

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

- LI implies that the *conditional distribution* of $\boldsymbol{Y}_i$ given $\boldsymbol{U}_i$ is:

$$p(\boldsymbol{y}_i|\boldsymbol{u}_i) = p(\boldsymbol{Y}_i = \boldsymbol{y}_i|\boldsymbol{U}_i = \boldsymbol{u}_i) = \prod_{t=1}^{T} \psi_{y_{it}|u_{it}}$$

- *Distribution* of $\boldsymbol{U}_i$: $$p(\boldsymbol{u}_i) = p(\boldsymbol{U}_i = \boldsymbol{u}_i) = \pi_{u_{i1}} \prod_{t>1} \pi_{u_{it}|u_{i,t-1}}$$

- *Manifest distribution* of $\boldsymbol{Y}_i$: $$p(\boldsymbol{y}_i) = p(\boldsymbol{Y}_i = \boldsymbol{y}_i) = \sum_{\boldsymbol{u}} p(\boldsymbol{y}_i|\boldsymbol{u})p(\boldsymbol{u})$$

- This may be *efficiently computed* through suitable recursions known in the hidden Markov literature (Baum *et al.*, 1970, Welch 2003)

RISIS RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

- The same tools available for FM/LC models may be used for *model estimation and selection*, although the EM algorithm requires particular care in the implementation based on certain recursions (Baum *et al.*, 1970, Welch 2003); package `LMest` in `R` may be used for applications

- After estimation, an important analysis is that of the *prediction of the latent states* for every unit and time occasion (dynamic clustering) that requires particular recursions (Viterbi, 1967; Juang & Rabiner, 1991)

- The LM model has been extended in several directions:
  - *multivariate longitudinal data* when more response variables are available at each time occasion
  - *inclusion of covariates* in the measurement or structural model with different interpretations and types of analysis
  - *multilevel longitudinal data* when units are clustered and so have a hierarchical structure

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

References

- Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In: B. N. Petrov and F. Csaki, eds., *Second International Symposium on Information Theory*, 267–281. Akademiai Kiado, Budapest.
- Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. John Wiley & Sons, Chichester, UK.
- Bartolucci, F., Bacci, S., and Gnaldi, M. (2015). *Statistical Analysis of Questionnaires: A Unified Approach Based on R and Stata*. Chapman and Hall/CRC, Boca Raton, FL.
- Bartolucci, F., Farcomeni, A., and Pennoni, F. (2013). *Latent Markov Models for Longitudinal Data*. Chapman and Hall/CRC press, Boca Raton, FL.
- Bartolucci, F., Lupparelli, M., and Montanari, G. E. (2009). Latent Markov model for longitudinal binary data: An application to the performance evaluation of nursing homes. *The Annals of Applied Statistics*, **3**:611–636.
- Bartolucci, F., Pandolfi, S., and Pennoni, F. (2017). LMest: An R package for latent Markov models for longitudinal categorical data. *Journal of Statistical Software*, **81**(4).
- Bartolucci, F., Pennoni, F., and Vittadini, G. (2011). Assessment of school performance through a multilevel latent Markov Rasch model. *Journal of Educational and Behavioral Statistics*, **36**:491–522.

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

- Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, **41**:164–171.
- Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). *Model-Based Clustering and Classification for Data Science: With Applications in R*. Cambridge University Press, Cambridge, UK.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**:1–38.
- Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, **61**:215–231.
- Juang, B. H. and Rabiner, L. R. (1991). Hidden Markov models for speech recognition. *Technometrics*, **33**:251–272.
- Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent Structure Analysis*. Houghton Mifflin, Boston.
- Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry and Applications*. In NSF-CBMS regional conference series in probability and statistics. Institute of Mathematical Statistics and the American Statistical Association.
- McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

- Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman and Hall/CRC, Boca Raton, FL.
- Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the IV Berkeley Symposium on Mathematical Statistics and Probability*, **4**:321–333.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**:461–464.
- Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). `mclust 5`: clustering, classification and density estimation using Gaussian finite mixture models, *The R Journal*, **8**:205-233.
- Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, **13**:260–269.
- Welch, L. R. (2003). Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter*, **53**:1–13.
- Wiggins, J. S. (1973). *Personality and Prediction: Principles of Personality Assessment*. Addison-Wesley Pub. Co.
- Zucchini, W., MacDonald, I. L., and Langrock, R. (2016). *Hidden Markov Models for Time Series: an Introduction Using R*. Chapman and Hall/CRC.

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES