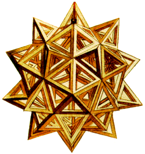
An URV Likelihood Approach to Cluster Ordinal Data

Roberto Rocci, University of Rome Tor Vergata.

joint work with

Monia Ranalli, Roma Tre University.



Milano, September 14, 2017

Today's themes

- **Introduction**
  How to cluster ordinal data?

- **Clustering ordinal data**
  Is pairwise likelihood a workable solution?
  (StCo 2014)

- **Simultaneous clustering and reduction**
  How to identify latent factors explaining the clustering structure?
  (e.g. factors explaining the between variability)
  By-products: noise variables identification, parsimonious modeling.
  (Psychometrika, 2017)

- **Clustering mixed-type data**
  Is composite likelihood a workable solution?
  (CSDA 2017)

Introduction

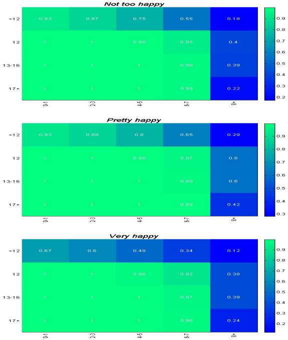
How to cluster ordinal data?

An illustrative example...

Three-way cross-classification of U.S. sample (N=1517) by happiness, years of schooling and number of siblings

Year of School Completed	*Number of Siblings*				
	0-1	2-3	4-5	6-7	8+
			Not too Happy		
< 12	15	34	36	22	61
12	31	60	46	25	26
13-16	35	45	30	13	8
17+	18	14	3	3	4
			Pretty Happy		
< 12	17	53	70	67	79
12	60	96	45	40	31
13-16	63	74	39	24	7
17+	15	15	9	2	1
			Very Happy		
< 12	7	20	23	16	36
12	5	12	11	12	7
13-16	5	10	4	4	3
17+	1	2	9	0	1

...our output: we classify the cells! (response profiles)



Finite mixture of Gaussians

A frequently used clustering model is the finite mixture of Gaussians.

$$f\left(\mathbf{x};\boldsymbol{\theta}\right)=\sum_{g=1}^{G}p_g\phi_P\left(\mathbf{x};\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g\right)$$

where:

- $\phi_P\left(\mathbf{x};\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g\right)$: P-variate Gaussian density with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$;
- $p_1, p_2, ..., p_G$: set of positive weights that sum to 1.

Use/Interpretation

- Each Gaussian density (component) is interpreted as a cluster (sub-population);
- p_g is the probability that an observation comes from the g-th sub-population.

After the estimation of model parameters (usually by ML), observations are assigned to clusters by maximizing the posterior probability

$$p(h|\mathbf{x}_r) = \frac{p_h \phi_P\left(\mathbf{x}_r; \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h\right)}{\sum_{g=1}^{G} p_g \phi_P\left(\mathbf{x}_r; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right)} \propto p_h \phi_P\left(\mathbf{x}_r; \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h\right)$$

i.e., the scaled density.

Problems with rank data

- It works on continuous data;
- Category scores are arbitrary;
- Is the Gaussian assumption true?

Conclusion. We should develop a new model appropriate for the ordinal nature of the data (or their measurements).

Theoretical Framework

Let

- $x_1, x_2, \ldots, x_P$ be the observed ordinal variables;
- $c_i = 1, \ldots, C_i$ be the associated categories for $i = 1, \ldots, P$;
- $\pi_r(\boldsymbol{\theta}) = Pr(x_1 = c_1, x_2 = c_2, \ldots, x_P = c_P; \boldsymbol{\theta})$ be the probability of the response pattern $\mathbf{x}_r$. It is a function of $\boldsymbol{\theta}$ s.t. $\pi_r(\boldsymbol{\theta}) \geq 0$ and $\sum_r \pi_r(\boldsymbol{\theta}) = 1$.

For a random i.i.d. sample of size N the log-likelihood is

$$\ell(\boldsymbol{\theta}; \mathbf{x}) = \sum_{r=1}^{R} n_r \log \pi_r(\boldsymbol{\theta}) \ ,$$

where n_r is the sample frequency of response pattern $\mathbf{x}_r$ and $\sum_r n_r = N$.

$$\Downarrow$$

Different models corresponds to different reparameterizations of π_r.

IRT & URV approaches (Factor Analysis)

Item Response Theory (IRT) - *Bock & Moustaki, 2007*

- $x_1, x_2, \ldots, x_P$: observed ordinal variables;
- $y_1, y_2, \ldots, y_Q$: latent variables $(Q < P)$;
- local independence assumption: the observed variables given the latent ones are independent.

The probability of a response pattern $\mathbf{x}_r$ is

$$\pi_r(\boldsymbol{\theta}) = Pr(x_1 = c_1, x_2 = c_2, \ldots, x_P = c_P; \boldsymbol{\theta}) = \int_{\mathcal{Y}} \prod_{i=1}^{P} p(c_i \mid \mathbf{y}) g(\mathbf{y}) d\mathbf{y}.$$

Underlying Response Variable (URV) - *Lee et al., 1990; Muthén, 1984*

- $x_1, x_2, \ldots, x_P$: observed ordinal variables;
- $y_1, y_2, \ldots, y_P$: latent continuous variables (usually Gaussians);
- the latent relationship between $\mathbf{x}$ and $\mathbf{y}$ is explained by a *threshold model*,
$$x_i = c_i \Leftrightarrow \gamma^{(i)}_{c_i - 1} \le y_i < \gamma^{(i)}_{c_i}.$$

The probability of a response pattern $\mathbf{x}_r$ is

$$\pi_r(\boldsymbol{\theta}) = Pr(x_1 = c_1, \ldots, x_P = c_P; \boldsymbol{\theta}) = \int_{\gamma^{(1)}_{c_1-1}}^{\gamma^{(1)}_{c_1}} \cdots \int_{\gamma^{(P)}_{c_P-1}}^{\gamma^{(P)}_{c_P}} \phi_P(\mathbf{y}; \mathbf{0}, \mathbf{R}) d\mathbf{y}.$$

A picture on the existing literature

Extensions of IRT and URV approaches to cluster analysis are obtained by assuming $\mathbf{y}$ as having a particular clustering structure.

- **IRT** → LCA: $Q = 1$ and y nominal discrete (*Goodman, 1974*); $\mathbf{y}$ finite mixture of Gaussians (*Cagnone & Viroli, 2012; McParland et al., 2012*);
- **URV** → ordinal variables are generated by thresholding a multivariate homoscedastic (*Everitt, 1988*) or heteroscedastic (*Lubke et al., 2008*) normal mixture density.

In both cases the computation of the likelihood is highly demanding because it requires the computation of multidimensional integrals.

Beside these approaches, we find: Mixture of latent variables for mixed data (*Browne & McNicholas, 2012*); Optimization clustering techniques (*Huang, 1998; Chaturvedi et al., 2001*); Probabilistic (*Giordan & Diana, 2011*);

Clustering ordinal data

Is pairwise likelihood a workable solution?

Our proposal

- **Aim**:
  Capturing both **cluster structure** and **dependence** within the groups. ⇒ Latent Gaussian Mixture following URV approach
- **Strengths**:
  - we do not assume the local independence;
  - it is possible to include an arbitrary number of ordinal variables;
  - it is possible to estimate both class conditional means and covariance matrices of the latent variables.
- **Main problems**:
  - How can we estimate this model efficiently?
    ⇒ Composite likelihood (pairwise) approach.
  - What are the conditions to identify the model?
    ⇒ Link with the log-linear models.
  - How can we classify the objects?
    ⇒ Different possible solutions.

Model assumptions

- $\mathbf{y} \sim f(\mathbf{y}) = \sum_{g=1}^{G} p_g \phi_P\left(\mathbf{y}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right)$;
- ordinal variables $\mathbf{x}$ are generated by thresholding $\mathbf{y}$.

It follows that the probability of response pattern $\mathbf{x}_r$ in cluster g is given by

$$\begin{aligned}\pi_r\left(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \gamma\right) &= Pr(x_1 = c_1, x_2 = c_2, \ldots, x_P = c_P | g)\\ &= \int_{\gamma^{(1)}_{c_1-1}}^{\gamma^{(1)}_{c_1}} \cdots \int_{\gamma^{(P)}_{c_P-1}}^{\gamma^{(P)}_{c_P}} \phi_P(\mathbf{y}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) d\mathbf{y},\end{aligned}$$

while the unconditional probability of response pattern $\mathbf{x}_r$ is given by

$$\pi_r(\boldsymbol{\theta}) = \sum_{g=1}^{G} p_g \pi_r\left(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \gamma\right).$$

For a random i.i.d. sample of size N the log-likelihood is

$$\ell(\boldsymbol{\theta}; \mathbf{x}) = \sum_{r=1}^{R} n_r \log \left(\sum_{g=1}^{G} p_g \pi_r \left(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \gamma \right) \right).$$

Existing proposals: Everitt (1988) and Lubke et al. (2008) use a MLE...

...This approach is **computationally demanding** and is not feasible for more than **few categorical variables**.

Estimation Problem - Looking at some possibilities...

- **Full Information Maximum Likelihood**: this approach is not feasible for more than few variables (Everitt and Merette, 1990).
- **Limited Information Maximum Likelihood**
    - **Three Stage Estimation** (Muthén, 1984) ⇒ it cannot be adopted. The thresholds cannot be estimated through the univariate marginal distributions, since they are not identified.
    - **Underlying Bivariate Normal** (Jöreskog and Moustaki, 2001) ⇒ maximizes the sum of all univariate and bivariate marginal loglikelihoods. Redundant, information contained in the univariate marginals is already in the bivariate ones.
    - **Composite maximum likelihood** (Lindsay, 1988; Varin et al., 2011): We suggest the use of the **Pairwise Likelihood Approach**.

Our choice/suggestion: Pairwise likelihood approach

The **pairwise log-likelihood** is the sum of all bivariate log-likelihoods

$$
\begin{aligned}
p\ell(\boldsymbol{\theta};\mathbf{x}) &= \sum_{i=1}^{P-1}\sum_{j=i+1}^{P}\ell(\boldsymbol{\theta};(x_i,x_j)) \\
&= \sum_{i=1}^{P-1}\sum_{j=i+1}^{P}\sum_{c_i=1}^{C_i}\sum_{c_j=1}^{C_j} n_{c_i c_j}^{(ij)} \log\left(\sum_{g=1}^{G} p_g \pi_{c_i c_j|g}^{(ij)}\right) \quad ,
\end{aligned}
$$

where $n_{c_i c_j}^{(ij)}$ is the observed frequency of a response in category c_i and c_j for variables x_i and x_j respectively, while $\pi_{c_i c_j|g}^{(ij)}$ is

$$
\pi_{c_i c_j|g}^{(ij)} = \int_{\gamma_{c_i-1}^{(i)}}^{\gamma_{c_i}^{(i)}}\int_{\gamma_{c_j-1}^{(j)}}^{\gamma_{c_j}^{(j)}} \phi\left(x_i,x_j;\mu_{i|g},\mu_{j|g},\sigma_{ii|g},\sigma_{jj|g},\rho_{ij|g}\right) dx_i dx_j
$$

In general, pairwise maximum likelihood estimators are less efficient than FML, even if in many cases the loss in efficiency is very small or almost null (Lindsay, 1988; Varin *et al.*, 2011), but much more efficient in terms of computational complexity (Ranalli et al. 2017). It has been proven that they are still consistent and asymptotically normal.

Parameter estimates are computed through a **pairwise EM algorithm**, whose **complete pairwise log-likelihood** is

$$p\ell_c(\boldsymbol{\theta};\mathbf{x}) = \sum_{i=1}^{P-1}\sum_{j=i+1}^{P}\sum_{c_i=1}^{C_i}\sum_{c_j=1}^{C_j}\sum_{g=1}^{G} n_{c_i c_j}^{(ij)} z_{c_i c_j|g}^{(ij)} \left[\log\left(\pi_{c_i c_j|g}^{(ij)}\right) + \log\left(p_g\right)\right].$$

Identifiability conditions

Investigating the identifiability conditions can be carried out...

- ...*empirically*: if the observed information matrix is full rank, then the model is **locally identified**;
- ...*heuristically*: if the same maximized likelihood is obtained with different parameter estimates starting the EM algorithm from different values, then the model is not identified;
- ...*theoretically*: the conditions needed to identify a model depend on its structure.

Necessary condition: given a $C_1 \times C_2 \times \ldots \times C_P$ contingency table, the number of model parameters has to be less than $C = \prod_{i=1}^{P} C_i - 1$.

- the necessary condition is

$$\prod_{i=1}^{P} C_i - 1 \geq \underbrace{G-1}_{p_g} + \underbrace{P(G-1)}_{\mu_2,\ldots,\mu_G} + \underbrace{P(P-1)/2}_{\mathbf{R}_1} + \underbrace{(G-1)P(P+1)/2}_{\Sigma_2,\ldots,\Sigma_G} + \underbrace{\sum_{i=1}^{P} C_i - P}_{thresholds}.$$

- the threshold parameters γ do not change over the components $\Rightarrow$ mean and variance of y_i, fixed to 0 and 1, respectively, only in one component;
- when $P = 1$ the model is not identified;
- when $P = 2$, in some cases the model could be identified (e.g. when $G = 2$ and $c_1 = c_2 = 4$), but it is not identified in others (e.g. when $G = 2$ and $c_1 = c_2 = 3$).

...**but** the pairwise loglikelihood works only with the bivariate marginals.

Example
Let us consider 3 binary variables

FML → 8 cells → 7 parameters

PML → 3 bivariate marginal distributions → 4 cells + 4 cells + 4 cells → 9 parameters?

Note that the 3 bivariate marginals can be reproduced by using a log-linear model with only two factor interactions (6 parameters).

The number of parameters needed to **saturate** the **pairwise likelihood** is equal to the number of parameters involved in a **hierarchical log linear model** with **two-factor interaction terms**

Necessary condition: the maximum number of **estimable parameters** has to be less than or equal to

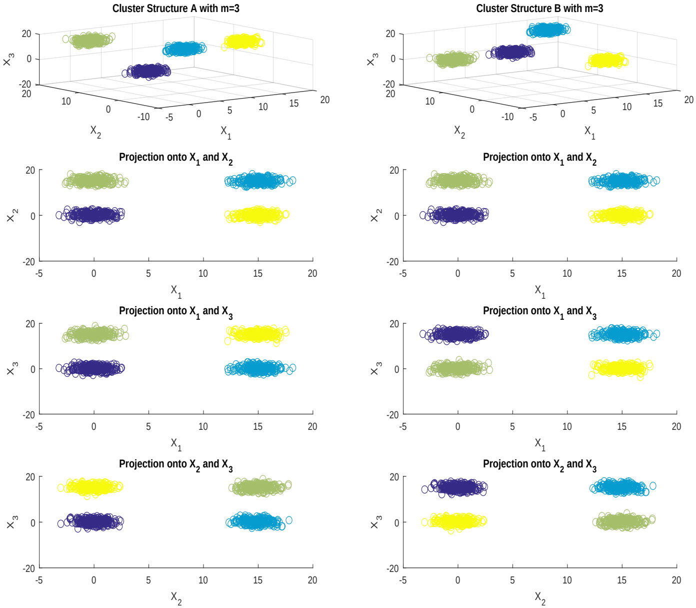
$$\sum_{i=1}^{P}(C_i - 1) + \sum_{i=1}^{P-1}\sum_{j=i+1}^{P}(C_i - 1)(C_j - 1).$$

Important: we are assuming that the clustering structure can be identified by the bivariate marginals.

How can we check for local identifiability? ...there are some *intuitions*

- look at the rank of the **Godambe Information** matrix;
- look at the rank of the **Jacobian matrix**, i.e. the derivatives of the log-linear parameters w.r.t. the latent mixture parameters (*Forcina*, 2008).

Is the clustering structure identified by the bivariate marginals?



Cluster Structure A with m=3

Cluster Structure B with m=3

Projection onto X_1 and X_2

Projection onto X_1 and X_2

Projection onto X_1 and X_3

Projection onto X_1 and X_3

Projection onto X_2 and X_3

Projection onto X_2 and X_3

Continuous case: finite mixtures of Gaussians

Proposition 1. Let $h(\mathbf{x};\boldsymbol{\theta}) = \sum_{g=1}^{G} p_g \phi(\mathbf{x};\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ be an element of $\mathcal{H}_{\mathcal{G}^{\mathcal{P}}}$ and $h_{ij}(x_i, x_j; \boldsymbol{\theta}^{(ij)})$ be its bivariate marginal density with respect to x_i and x_j, if for any $g \neq h$ $(g, h = 1, \ldots, G)$ we have

$$(\mu_{i,g}, \mu_{j,g}, \sigma_{ii,g}, \sigma_{ij,g}, \sigma_{jj,g}) \neq (\mu_{i,h}, \mu_{j,h}, \sigma_{ii,h}, \sigma_{ij,h}, \sigma_{jj,h}) \qquad (1)$$

then $h_{ij}(x_i, x_j; \boldsymbol{\theta}^{(ij)})$ belongs to $\mathcal{H}_{\mathcal{G}^2}$ and it has G components.

marginal (i, j)**:**
if the parameters are distinct
then the number of clusters is identified

Proposition 2. Let $h(\mathbf{x};\boldsymbol{\theta}) = \sum_{g=1}^{G} p_g \phi(\mathbf{x};\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ and $h(\mathbf{x};\tilde{\boldsymbol{\theta}}) = \sum_{g=1}^{\tilde{G}} p_g \phi(\mathbf{x};\tilde{\boldsymbol{\mu}}_g, \tilde{\boldsymbol{\Sigma}}_g)$ be two members of $\mathcal{H}_{\mathcal{G}^{\mathcal{P}}}$. If they are such that

$$p_1 > p_2 > \cdots > p_G, \tilde{p}_1 > \tilde{p}_2 > \cdots > \tilde{p}_G$$

and for every $i \neq j = 1, 2, \ldots, P$

$-h_{ij}(x_i, x_j; \boldsymbol{\theta}^{(ij)}) = h_{ij}(x_i, x_j; \tilde{\boldsymbol{\theta}}^{(ij)})$

$-(\mu_{i,g}, \mu_{j,g}, \sigma_{ii,g}, \sigma_{ij,g}, \sigma_{jj,g}) \neq (\mu_{i,h}, \mu_{j,h}, \sigma_{ii,h}, \sigma_{ij,h}, \sigma_{jj,h}), g \neq h = 1, \ldots, G$

$-(\tilde{\mu}_{i,g}, \tilde{\mu}_{j,g}, \tilde{\sigma}_{ii,g}, \tilde{\sigma}_{ij,g}, \tilde{\sigma}_{jj,g}) \neq (\tilde{\mu}_{i,h}, \tilde{\mu}_{j,h}, \tilde{\sigma}_{ii,h}, \tilde{\sigma}_{ij,h}, \tilde{\sigma}_{jj,h}), g \neq h = 1, \ldots, \tilde{G},$

then $h(\mathbf{x};\boldsymbol{\theta}) = h(\mathbf{x};\tilde{\boldsymbol{\theta}})$.

Ordinal case?

Further cautions...

How do we classify?

Response patterns are assigned to the components according to

$$\text{argmax}_h \; p(h \mid \mathbf{x}_r) = \text{argmax}_h \; p_h \pi_r \left(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h, \gamma\right),$$

BUT, we do not estimate $\pi_{r|g} = \pi_r \left(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \gamma\right)$ directly...

Three possible alternatives:

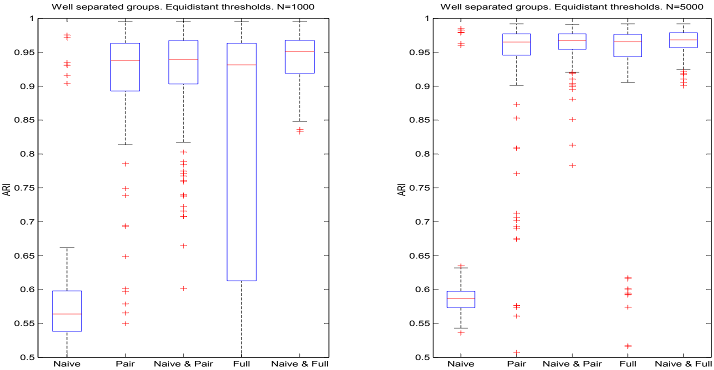
- **FMAP** compute $\pi_{r|g}$. There are multidimensional integrals involved but they have to be evaluated only once;
- **IMAP** approximate $\pi_{r|g}$ with a distribution having the same (estimated) bivariate marginals (IPF algorithm). We considered the case of a log-linear model having only the two-factor interactions different from zero;
- **CMAP** substitute the likelihood of the response on component h with its pairwise version.

Simulation Study

- **AIM**: evaluating the effect of some experimental factors on classification performance.
- The **pairwise likelihood** approach is compared to the **full maximum likelihood** and the **maximum likelihood for continuous data**.
- Simulation design: *250 samples* in eight different scenarios considering *three different experimental factors*:
  - *sample size* - $N = 1000, 5000$;
  - *thresholds* - equidistant or non;
  - *separation* between clusters - well or non well separated.
- To evaluate the classification performance of the methods two different measures of classification recovery have been considered
  - Crisp: **Adjusted Rand Index** (ARI), 1 best - 0 or < 0 worst
  - Fuzzy: **LOSS** measure, 0 best - 1 worst

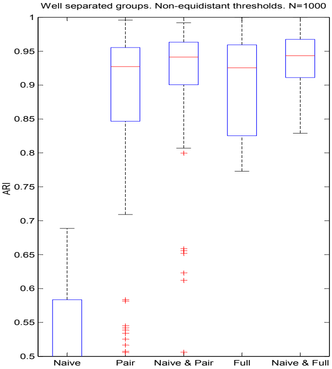
Results: ARI

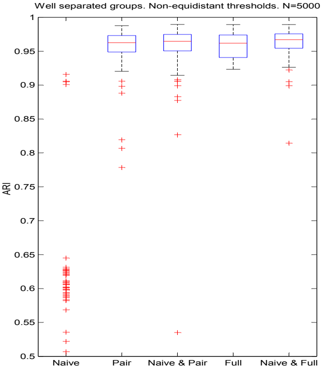
Equidistant thresholds - N=1000,5000 - *Well* separated groups.



Results: ARI

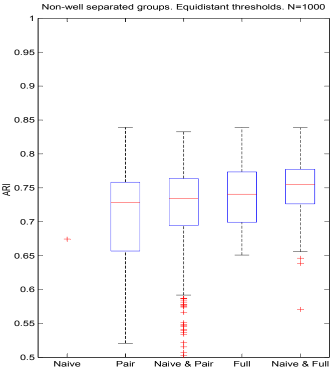
Non-Equidistant thresholds - N=1000,5000 - *Well* separated groups.



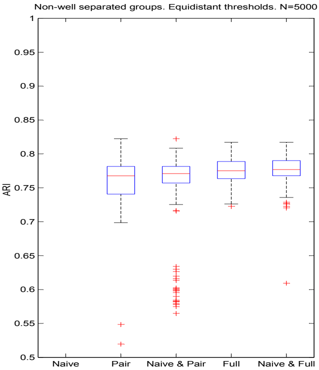


Results: ARI

Equidistant thresholds - N=1000,5000 - *Non-Well* separated groups.



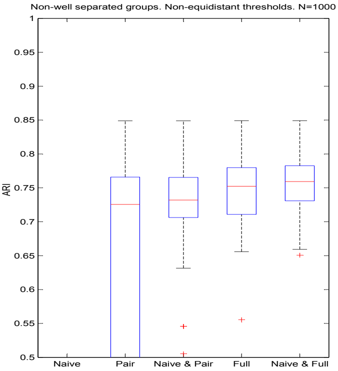
Non-well separated groups. Equidistant thresholds. N=1000

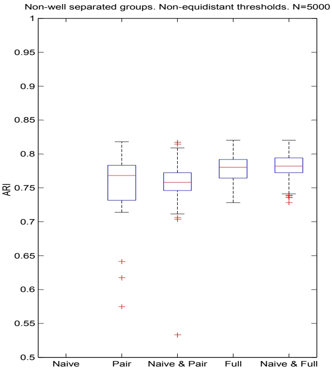


Non-well separated groups. Equidistant thresholds. N=5000

Results: ARI

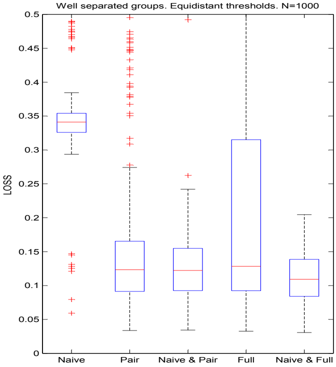
Non-Equidistant thresholds - N=1000,5000 - *Non-Well* separated groups.

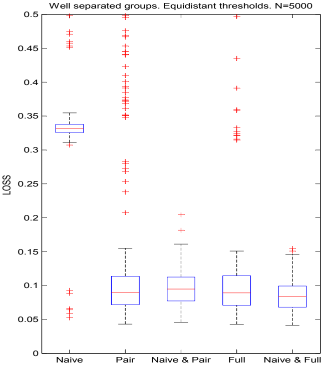




Results: LOSS

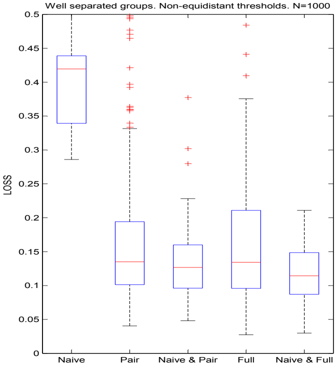
Equidistant thresholds - N=1000,5000 - *Well* separated groups.

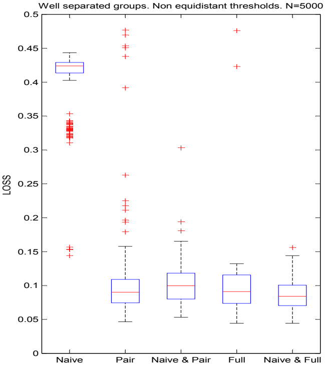




Results: LOSS

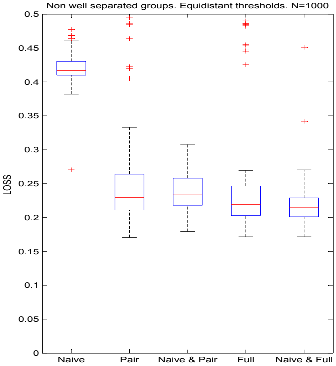
Non-Equidistant thresholds - N=1000,5000 - *Well* separated groups.

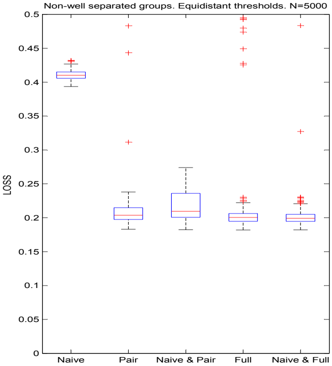




Results: LOSS

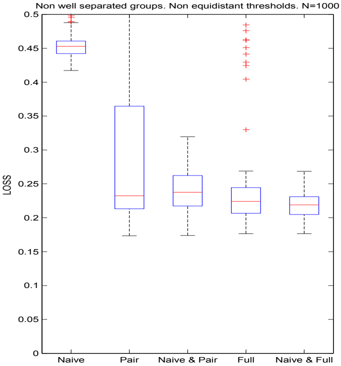
Equidistant thresholds - N=1000,5000 - *Non-Well* separated groups.

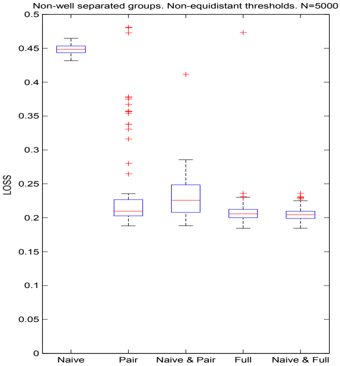




Results: LOSS

Non-Equidistant thresholds - N=1000,5000 - *Non-Well* separated groups.





Model selection: information criteria

How can we choose the number of components?

- Gao & Song (2010): C-BIC $= -2p\ell(\hat{\boldsymbol{\theta}}; \boldsymbol{x}) + \log N \mathrm{tr}\left(\hat{\boldsymbol{H}}^{-1}\hat{\boldsymbol{V}}\right)$;
- Varin & Vidoni (2005): C-AIC $= -2p\ell(\hat{\boldsymbol{\theta}}; \boldsymbol{x}) + 2\mathrm{tr}\left(\hat{\boldsymbol{H}}^{-1}\hat{\boldsymbol{V}}\right)$.

Where

- $\mathbf{H} = E(-\nabla^2 p\ell)$ is the sensitivity matrix;
- $\mathbf{V} = V(\nabla p\ell)$ is the variability matrix.

In FML the two matrices are equal.

Challenge: the estimation of **H** and **V**

General Social Survey

It is a three-way cross-classification table of **1,517** people on three ordinal variables: **happiness** (3 categories), **completed years of schooling** (4 categories), and **number of siblings** (5 categories), analysed by Goodman (1984) and re-analysed recently by Giordan et al. (2011).

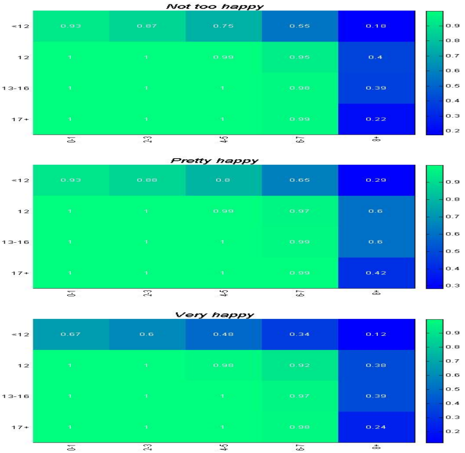
Year of School Completed	*Number of Siblings* 0-1	2-3	4-5	6-7	8+
	Not too Happy				
< 12	15	34	36	22	61
12	31	60	46	25	26
13-16	35	45	30	13	8
17+	18	14	3	3	4
	Pretty Happy				
< 12	17	53	70	67	79
12	60	96	45	40	31
13-16	63	74	39	24	7
17+	15	15	9	2	1
	Very Happy				
< 12	7	20	23	16	36
12	5	12	11	12	7
13-16	5	10	4	4	3
17+	1	2	9	0	1

Model choice - C-AIC and C-BIC

	G=1	**G=2**	**G=3**
C-AIC	22772	22730	22754
C-BIC	22937	22896	22972

General Social Survey - Output Analysis

- there is a **clear classification** between the two groups **as the years of school completed and number of siblings increase**;
- the variable **happiness** has **not** a **discriminative power**.

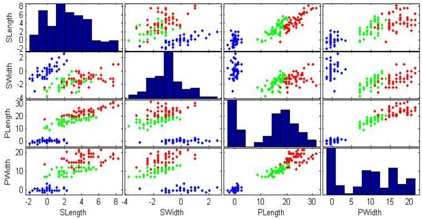


An advanced simulation study

Through a **model-based** method using a **pairwise likelihood** approach.

A clustering benchmark: when the latent mixture is observed...

Fisher's Iris data



Idea

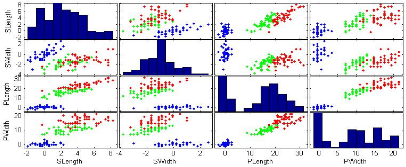
Ordinal variables are generated by *thresholding* a *latent mixture*

- is the cluster structure recovered?
- how good is the proposal compared to the existing models?

The robustness of clustering problem. Fisher's Iris data

- **150 four** dimensional observations (sepal length & width; petal length & width) of **three** different species of Iris: *Iris setosa*, *Iris versicolour* and *Iris virginica*;
- to re-analyze these data under our proposal, the variables have been **categorized**;
- the data have been **normalized** by the mean and the standard deviation of the first group (*Iris setosa*);
- the **threshold parameters** have been chosen such that the cluster structure has not been completely destroyed (ARI is maximized);
- combination of **four** and **three** categories & the **thresholds** are **equidistant**.

Response patterns				n_r	**Response patterns**				n_r
1	1	1	1	1	3	1	3	2	3
1	1	2	2	4	3	1	3	3	1
1	1	3	2	1	3	1	4	3	1
1	2	1	1	16	3	2	3	2	16
1	3	1	1	19	3	2	3	3	10
2	1	2	2	3	3	2	4	2	1
2	1	3	2	4	3	2	4	3	10
2	1	3	3	1	3	3	3	2	1
2	2	2	2	4	3	3	3	3	1
2	2	3	2	14	3	3	4	3	4
2	2	3	3	7	4	2	3	2	1
2	3	1	1	8	4	2	4	2	1
2	3	3	2	1	4	2	4	3	8
2	4	1	1	6	4	3	4	3	3



Fisher's Iris data - Comparative Analysis

Some observations...

- pairwise EM initialized from the *true empirical values* → ARI=**0.9222**;
- pairwise EM initialized *randomly* considering 1000 different starting points → ARI=**0.8005**;
- the **overlap** between the second and third group reflects on a larger classification uncertainty;
- **robustness** of clustering problem: even if the data were categorized, the cluster structure has been recovered satisfactorily.

ARI - Clustering performances

Ordinal variables as metric		**Ordinal variables as ordinal**	
k-means	0.6615	MCA & *k*-means (3 fact.)	0.5676
HomFMG(D)	0.6634	MCA & *k*-means (2 fact.)	0.7874
HomFMG(F)	0.5153	LFMG (TV)	**0.9222**
HetFMG(F)	0.4128	LFMG (RV)	**0.8005**

Confusion matrix - Fisher's Iris data

	G=1	**G=2**	**G=3**
G=1	50	0	0
G=2	0	46	4
G=3	0	4	46

Simultaneous clustering and reduction

How to identify latent factors explaining the clustering structure?

(e.g. factors explaining the between variability)

By-products: noise variables identification, parsimonious modeling.

Simultaneous reduction and clustering for ordinal data

Three-way cross-classification of U.S. sample

	Number of Siblings				
Year of School Completed	0-1	2-3	4-5	6-7	8+
	Not too Happy				
< 12	15	34	36	22	61
12	31	60	46	25	26
13-16	35	45	30	13	8
17+	18	14	3	3	4
	Pretty Happy				
< 12	17	53	70	67	79
12	60	96	45	40	31
13-16	63	74	39	24	7
17+	15	15	9	2	1
	Very Happy				
< 12	7	20	23	16	36
12	5	12	11	12	7
13-16	5	10	4	4	3
17+	1	2	9	0	1

...these are our main questions

What are the latent factor that explain the clustering structure?

and

How are they related with the observed variables?

The aim is to propose a model for **simultaneous clustering** and **dimensionality reduction** of the ordered categorical data using a **composite likelihood** approach.

Recalling model assumptions

- $x_1, x_2, \ldots, x_P$: observed ordinal variables;
- $y_1, y_2, \ldots, y_P$: latent continuous variables;
- the latent relationship between $\mathbf{x}$ and $\mathbf{y}$ explained by a *threshold model*,

$$x_i = c_i \Leftrightarrow \gamma^{(i)}_{c_i - 1} \leq y_i < \gamma^{(i)}_{c_i}.$$

The probability of a *response pattern* $\mathbf{x}_r$ is given by

$$Pr(x_1 = c_1, \ldots, x_P = c_P; \boldsymbol{\theta}) = \sum_{g=1}^{G} p_g \int_{\gamma^{(1)}_{c_1 - 1}}^{\gamma^{(1)}_{c_1}} \cdots \int_{\gamma^{(P)}_{c_P - 1}}^{\gamma^{(P)}_{c_P}} \phi(\mathbf{y}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) d\mathbf{y}$$

where p_g is the probability of belonging to group g subject to $p_g > 0$ and $\sum_{g=1}^{G} p_g = 1$.

Key points of the proposal

- observed ordinal variables are a *discretization* of underlying **first-order latent** continuous variables $\mathbf{y}$.
- first order latent variables are linear combinations of **second-order latent** variables $\tilde{\mathbf{y}}$.

$$\mathbf{x} \leftarrow \mathbf{y} = \mathbf{A}\tilde{\mathbf{y}} \leftarrow \tilde{\mathbf{y}}$$

To detect **informative/noise dimensions** second-order latent variables are divided into two groups:

- Q *informative/discriminant* factors distributed as a *finite mixture of Gaussians*;
- $\bar{Q} = P - Q$ *non informative/noise* factors distributed as a *Gaussian*.

All relevant information about the clustering structure is captured by the set of informative/discriminant latent variables.

Model parameters

Since $\mathbf{y} = \mathbf{A}\tilde{\mathbf{y}}$, then

$$\boldsymbol{\mu}_g = E(\mathbf{y}|g) = \mathbf{A}E(\tilde{\mathbf{y}}|g) = \mathbf{A}\begin{bmatrix} \eta_{g,1} \\ \vdots \\ \eta_{g,Q} \\ \eta_{0,Q+1} \\ \vdots \\ \eta_{0,P} \end{bmatrix} = \mathbf{A}\begin{bmatrix} \boldsymbol{\eta}_g \\ \boldsymbol{\eta}_0 \end{bmatrix}$$

and

$$\boldsymbol{\Sigma}_g = V(\mathbf{y}|g) = \mathbf{A}V(\tilde{\mathbf{y}}|g)\mathbf{A}' = \mathbf{A}\begin{bmatrix} \boldsymbol{\Omega}_g & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}_0 \end{bmatrix}\mathbf{A}'.$$

Parsimonious Modeling: some parameters are set to zero or equal to others.

Parameters estimation

The parameters are estimated by maximizing the pairwise log-likelihood (if the FML is infeasible).

Estimates computation

The maximization of the pairwise log-likelihood is done by using a *pairwise EM algorithm*.

Model selection

The number of components and/or discriminative dimensions is chosen by minimizing the C-BIC.

Identifiability conditions

- given a $C_1 \times C_2 \times \ldots \times C_p$ contingency table, the **necessary condition** for the identifiability is that the number of model parameters can be **at most**

$$\sum_{i=1}^{P}(C_i - 1) + \sum_{i=1}^{P-1}\sum_{j=i+1}^{P}(C_i - 1)(C_j - 1);$$

- the threshold parameters γ do not change over the components;
- the first two thresholds are fixed to 0 and 1, respectively;
- some constraints on $\mathbf{\Omega}_0$, $\mathbf{\Omega}_1$ and $\mathbf{A}$ to solve the rotational indeterminacy.

Rotational Indeterminacy

Writing $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2]$, we note that we have the sum of two factor analysis (FA) models

$$\mathbf{y} = \mathbf{A}\tilde{\mathbf{y}} = \mathbf{A}_1\tilde{\mathbf{y}}^Q + \mathbf{A}_2\tilde{\mathbf{y}}^{\bar{Q}}$$

They have the same rotational freedom of the FA model, i.e.

$$\begin{aligned} \mathbf{y} &= \mathbf{A}_1\mathbf{T}_1\mathbf{T}_1^{-1}\tilde{\mathbf{y}}^Q + \mathbf{A}_2\mathbf{T}_2\mathbf{T}_2^{-1}\tilde{\mathbf{y}}^{\bar{Q}} \\ &= \mathbf{A}_1^*\tilde{\mathbf{y}}^{*Q} + \mathbf{A}_2^*\tilde{\mathbf{y}}^{*\bar{Q}} \end{aligned}$$

In order to make the parameters identified, and then estimable, we put some constraints: $\mathbf{\Omega}_0 = \mathbf{I}$, $\mathbf{\Omega}_1 = \mathbf{I}$. Such constraints still allow a rotational freedom by orthonormal matrices. This can be eliminated by requiring a "lower" triangular form for the two loading matrices. In general, $\mathbf{A}_1$ and $\mathbf{A}_2$ have a lower triangular matrix in the first Q and $(P - Q)$ rows, respectively. Of course, after the estimation the parameter matrices can be rotated to enhance the interpretation.

Noise variables

- One important step is to identify the observed variables that could be considered as noise.
- Intuitively this information is included in the correlation matrix between the first and second order latent variables.

$$\mathbf{A} V\left(\tilde{\mathbf{y}}\right)$$

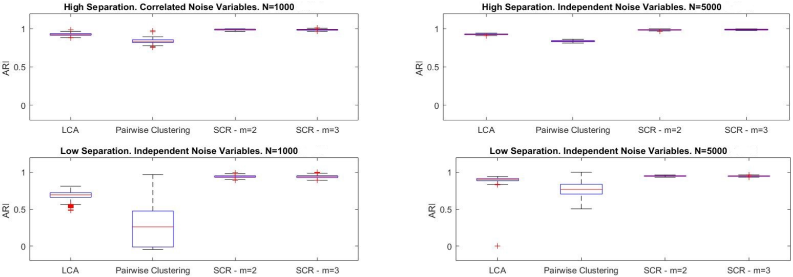
- $V(\tilde{\mathbf{y}})$ accounts for both the between and within variance of the mixture.
- The variables x corresponding to first order latent factors y that are well correlated with the noise factors $\tilde{\mathbf{y}}^{\bar{Q}}$ are identified as noise.
- It is also important to evaluate for each first order factor the percentage of variance explained by the informative second order factors.

Simulation Study

- **AIM**: evaluating the effect of some experimental factors on classification performance;
- **simultaneous clustering & dimensionality reduction approach** with $m = 2$ and $m = 3$ compared with the **naive clustering approach** (i.e. $Q = P$), LCA estimated with $m = P$;
- simulation design: *250 samples* in *eight* different scenarios considering *two different experimental factors*:
  - *sample size* - $N = 1000, 5000$;
  - *separation* between clusters - well or non well separated;
  - *# of components and variables*: G=2, P=5, Q=2; G=3, P=8, Q=3;
- we assess the assumptions of local independence of LCA (correlated vs non-correlated noise variables);
- we assess the performances of model selection;
- ARI is used to evaluate the performance in terms of goodness of recovery of the true clustering structure.

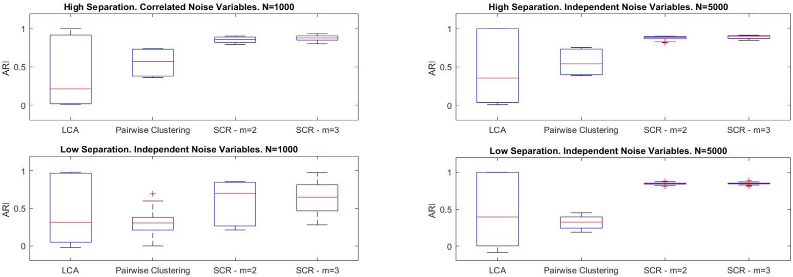
First Scenario - P=5 and G=2

Box-plots of ARI for the posterior probabilities. Data generated from a two-component latent mixture; 5 ordinal variables with 5 categories; 3 of them are noise variables. N=1000,5000. High/Low separation. Independent noise variables.



Second Scenario - P=8 and G=3

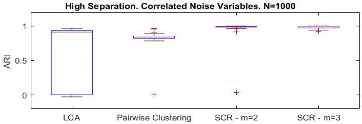
Box-plots of ARI for the posterior probabilities. Data generated from a three-component latent mixture; 8 ordinal variables with 5 categories; 5 of them are noise variables. N=1000,5000. High/Low separation. Independent noise variables.

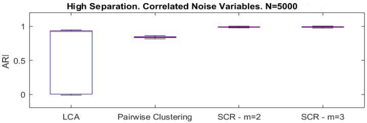


Third Scenario - Correlated Noise Variables

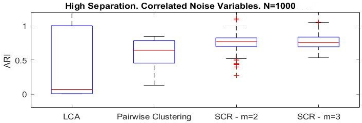
Box-plots of ARI for the posterior probabilities. 250 samples generated from $G = 2$ and $G = 3$, with correlated noise variables, and N=1000,5000. $G = 2$: 5 ordinal variables with 5 categories, 3 of them are noise variables. $G = 3$: 8 ordinal variables with 5 categories, 5 of them are noise variables.

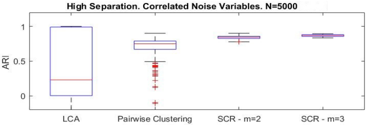
G=2


High Separation. Correlated Noise Variables. N=1000


High Separation. Correlated Noise Variables. N=5000

G=3


High Separation. Correlated Noise Variables. N=1000


High Separation. Correlated Noise Variables. N=5000

Fourth Scenario - Model Selection

ARI for the best model chosen through C-BIC compared to the ARI of the true model.
$G = 2$, 5 ordinal variables with 5 categories. High degree of separation and independent noise variables. $N = 1000$. 50 samples have been generated with $Q = 1, 2, 3, 4$.
For each of the 200 samples 5 different models have been fitted.

	Mean	**St.Dev**	**q=0.025**	**q=0.25**	**q=0.5**	**q=0.75**	**q=0.975**
C-BIC ARI	0.9674	0.0796	0.9431	0.9694	0.9959	1.0000	1.0000
True Fitted ARI	0.9797	0.0247	0.9488	0.9796	0.9918	1.0000	1.0000

General Social Survey

It is a three-way cross-classification table of *1,517* people on three ordinal variables: completed years of schooling (4 categories), number of siblings (5 categories), and happiness (3 categories).

Year of School Completed	*Number of Siblings* 0-1	2-3	4-5	6-7	8+
	Not too Happy				
<12	15	34	36	22	61
12	31	60	46	25	26
13-16	35	45	30	13	8
17+	18	14	3	3	4
	Pretty Happy				
<12	17	53	70	67	79
12	60	96	45	40	31
13-16	63	74	39	24	7
17+	15	15	9	2	1
	Very Happy				
<12	7	20	23	16	36
12	5	12	11	12	7
13-16	5	10	4	4	3
17+	1	2	9	0	1

General Social Survey

Model choice - C-BIC & BIC

	C-BIC - $m=2$			**BIC** - $m=3$		
	G=1	**G=2**	**G=3**	**G=1**	**G=2**	**G=3**
Q=1	24717	**22848**	22890	12950	**12610**	12770
Q=2	23151	22881	22891	13193	12634	12831
Q=3	22937	22896	22972	12294	12367	12463

Correlations between **y** (by rows) and $\tilde{\mathbf{y}}$ (by columns) variables

$$\left[\begin{array}{r|rr} 0.9987 & 0.0509 & 0.0000 \\ \hdashline -0.4951 & 0.8439 & -0.2065 \\ -0.1884 & 0.2074 & 0.9600 \end{array}\right]$$

General Social Survey - Output Analysis

- there is a **clear classification** between the two groups **as the education level increases**;
- the variable **happiness** has **not discriminative power**.

Table: Empirical Evidence on the presence of noise dimensions between years of schooling (X_1), number of siblings (X_2) and happiness (X_3) by pairs.

Bivariate marginals	Polychoric correlation	ϕ-Coefficient	Cramer's V	Goodman-Kruskal γ (s.e.) [C.I. 95%]
X_1 & X_2	-0.425	0.394	0.227	-0.425 (0.025) [-0.474,-0.377]
X_1 & X_3	-0.161	0.165	0.116	-0.169 (0.036) [-0.24, -0.099]
X_2 & X_3	0.073	0.13	0.092	0.07 (0.033) [0.006, 0.135]

ISSP - Output Analysis

Multi-way table taken from the International Social Survey Programme (ISSP) on environment in 1993.
The possible answers to each question are: (1) strongly agree, (2) somewhat agree, (3) neither agree nor disagree, (4) somewhat disagree, (5) strongly disagree.

Questions
X_1: We believe too often in science, and not enough in feelings and faith
X_2: Overall, modern science does more harm than good
X_3: Any change humans cause in nature, no matter how scientific, is likely to make things worse
X_4: Modern science will solve our environmental problems with little change to our way of life

Table: Model choice according to C-BIC and BIC for pairwise likelihood approach ($m = 2$) and full likelihood approach ($m = 4$), respectively.

	C-BIC - $m = 2$				**BIC** - $m = 4$			
	G=1	**G=2**	**G=3**	**G=4**	**G=1**	**G=2**	**G=3**	**G=4**
Q=1	37228	38003	34875	34887	35735	31070	31107	31130
Q=2	34850	39094	35223	34925	33815	31069	31076	31167
Q=3	37111	**33267**	34925	34968	37450	**31036**	31054	31161
Q=4	38957	38468	38092	40876	32569	31172	31246	31287

$$\begin{array}{cc} & \tilde{\mathbf{y}} \\ \mathbf{y} & \left[\begin{array}{ccc:c} 0.9258 & 0.5422 & 0.2135 & 0.4377 \\ 0.4350 & 0.9421 & 0.5070 & 0.4466 \\ 0.3094 & 0.4952 & 0.8756 & 0.5150 \\ \hdashline -0.2758 & -0.2592 & -0.4968 & 0.8082 \end{array}\right] \end{array}.$$

The noisy dimension is mainly given by the fourth question

Table: Empirical Evidence on the presence of a noise dimension between X_1, X_2, X_3 and X_4 by pairs.

Bivariate Marginals	Polychoric Corr.	ϕ-Coefficient	Cramer's V	Goodman-Kruskal γ (s.e.) [C.I. 95%]
X_1 & X_2	0.421	0.488	0.244	0.402 (0.0304) [0.335, 0.470]
X_1 & X_3	0.400	0.441	0.220	0.400 (0.035) [0.330, 0.467]
X_2 & X_3	0.488	0.545	0.273	0.474 (0.032) [0.411, 0.537]
X_1 & X_4	0.034	0.229	0.115	0.039 (0.039) [-0.037, 0.116]
X_2 & X_4	0.005	0.291	0.146	0.026 (0.040) [-0.052 0.105]
X_3 & X_4	0.072	0.393	0.197	-0.073 (0.041) [-0.154 0.007]

Looking at the posterior probabilities of the response patterns (although they are not included in the paper), it seems that the two groups cluster individuals based on the score assigned to the questions - high score on the first two questions (bad feeling towards science) and low score on the third one. As a consequence, the two groups can be interpreted as degree of belief in science (or faith, conversely) - strong vs. weak.

Clustering mixed-type data

Is composite likelihood a workable solution?

A real data example: Credit Scoring data

- **BAccount**: a factor with levels no - good running - bad running, quality of the credit clients bank account;
- **Months**: duration of loan in months;
- **Past**: a factor with levels bad payer - good payer if the client previosly have been a bad or good payer;
- **Use**: a factor with levels private – professional, the use to which the loan is made;
- **DM**: the size of loan in DM;
- **Gender**: a factor with levels M – F, sex of the client;
- **Status**: a factor with levels no single - single, status of the client.

There are **1000** loan applicants: **300** are **bad** and **700** are **good**.

We want to capture the cluster structure

Model assumptions

Continuous data $\cup$ **Ordinal data** $=?$

- $y_1, \dots, y_P$ continuous variables
- $x_1, \dots, x_Q (Q \leq P)$ ordinal variables
- $\mathbf{y} \sim f(\mathbf{y}) = \sum_{g=1}^{G} p_g \phi_P \left(\mathbf{y}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g \right)$;
- ordinal variables $\mathbf{x}$ are generated by thresholding $\mathbf{y}^Q$;

For a random i.i.d. sample of size N with mixed-type data the log-likelihood is

$$\ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{y}^{\bar{Q}}) = \sum_{n=1}^{N} \log \left[\sum_{g=1}^{G} p_g \phi_{\bar{Q}}(\mathbf{y}_n^{\bar{Q}}; \boldsymbol{\mu}_g^{\bar{Q}}, \boldsymbol{\Sigma}_g^{\bar{Q}}) \pi_n \left(\boldsymbol{\mu}_{n;g}^{Q|\bar{Q}}, \boldsymbol{\Sigma}_g^{Q|\bar{Q}}, \gamma \right) \right].$$

...Adopting a ML approach is **computationally demanding** and is not feasible for more than **few ordinal variables** (*Everitt & Merette, 1990*).

How can we estimate it efficiently?

We adopt a composite likelihood approach distinguishing three blocks of marginals

- sub-set of continuous variables
  marginal distribution of the continuous variables ∼ heteroscedastic Gaussian mixture
- sub-set of ordinal variables
  all bivariate marginal distributions of ordinal variables ∼ partial manifestation of the underlying heteroscedastic Gaussian mixture
- sub-set of mixed-type variables
  the marginal distributions given by all continuous variables and only one ordinal variable.

The **composite log-likelihood** is

$$
\begin{aligned}
c\ell(\boldsymbol{\theta}) = & \sum_{n=1}^{N} \log \left[\sum_{g=1}^{G} p_g \phi_{\bar{Q}}(\mathbf{y}_n^{\bar{Q}}; \boldsymbol{\mu}_g^{\bar{Q}}, \boldsymbol{\Sigma}_g^{\bar{Q}\bar{Q}}) \right] + \\
& + \sum_{i=1}^{Q-1} \sum_{j=i+1}^{Q} \sum_{n=1}^{N} \sum_{c_i=1}^{C_i} \sum_{c_j=1}^{C_j} \delta_{nc_ic_j}^{(ij)} \log \left[\sum_{g=1}^{G} p_g \pi_{c_ic_j}^{(ij)}(\boldsymbol{\mu}_g^{(ij)}, \boldsymbol{\Sigma}_g^{(ij)}, \gamma^{(ij)}) \right] + \\
& + \sum_{j=1}^{Q} \sum_{n=1}^{N} \sum_{c_j}^{C_j} \delta_{nc_j}^{(j)} \log \left[\sum_{g=1}^{G} p_g \pi_{c_j}^{(j|\bar{Q})}(\mu_{n;g}^{(j|\bar{Q})}, \sigma_g^{(j|\bar{Q})}, \gamma^j) \phi_{\bar{Q}}(\mathbf{y}_n^{\bar{Q}}; \boldsymbol{\mu}_g^{\bar{Q}}, \boldsymbol{\Sigma}_g^{\bar{Q}\bar{Q}}) \right]
\end{aligned}
$$

Estimates are computed by using an EM-like algorithm.

Classification & Model Selection

- **Classification**
  - **FMAP**
    The posteriors are computed by using the composite likelihood estimates.
  - **IMAP**
    Computationally infeasible.
  - **CMAP**
    Assign the observation to the component corresponding to the maximum composite fit.
- **Model Selection**

$$cCLC = -2c\ell(\hat{\psi}) + 2EN(\hat{\mathbf{p}}),$$

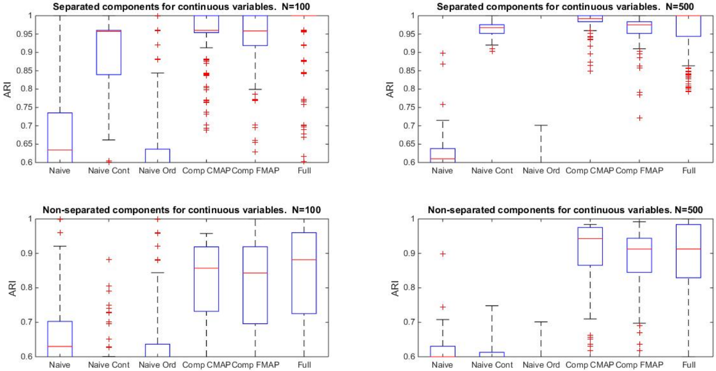
where *EN* works as a penalty term, it is the composite-entropy of the fuzzy classification obtained in the E-step of the EM-like algorithm.

Simulation study design

- 250 samples simulated from a latent mixture of Gaussians in 4 scenarios with 2 experimental factors: **sample size** ($N = 100, 500$) and **separation degree** between clusters.
- **competitors**:
  1. **Naive**. The ordinal nature is ignored and the variables are treated as they were metric;
  2. **Continuous Naive**. An heteroscedastic Gaussian mixture is fitted considering only continuous variables;
  3. **Ordinal Naive**. The ordinal nature is ignored and the variables are treated as they were metric;
  4. **Composite**. The proposed model is fitted through a composite likelihood. The observations are assigned to the components based on CMAP or FMAP;
  5. **Full**. The proposed model is fitted through full maximum likelihood.
- ARI is the index used to measure the goodness of models/algorithms;

Simulation study output

Figure: Box-plots of ARI for the posterior probabilities. Data generated from two-component mixture model partially observed; 3 ordinal variables with 5 categories and 3 continuous variables. N=100,500. Separated/non-separated means for continuous variables. 250 samples.



Credit Scoring Data

Has the cluster structure been captured?

Table: Model selection

	G=1	**G=2**	**G=3**
cCLC	24978	18864	22973

Table: Confusion matrix

	Default=1	Default=0
Cluster 1	220	10
Cluster 2	80	690

Credit risk profile

- mainly **single females**
- with a **lower credit-quality** bank account
- who were **bad payers** in the past
- and asked a **loan for private use**.

Some References

- Bock, D., Moustaki, I.: *Handbook of Statistics on Psychonometrics, chap. Item response theory in a general framework*, Elsevier, 2007.
- Cagnone S. and Viroli C., *A factor mixture analysis model for multivariate binary data*. Statistical Modelling, 12: 257-277, 2012.
- Everitt, B.S.: *A finite mixture model for the clustering of mixed-mode data*. Statistics Probability Letters, 6(5): 305-309, 1988.
- Everitt, B., Merette, C.: *The clustering of mixed-mode data: a comparison of possible approaches*, Journal of Applied Statistics **17**(3), 283–297, 1990.
- Gao, X., Song, P.X. *Composite likelihood EM algorithm with applications to multivariate hidden Markov model*, 2010.
- Goodman, L. A.: *Exploratory latent structure analysis using both identifiable and unidentifiable models*. Biometrika 61, 2, 215–231, 1974.
- Jöreskog, K.G., Moustaki, I.: *Factor analysis for ordinal variables: a comparison of three approaches*, Multivariate Behavioural Research 36, 347–387, 2001.
- Lee, S.Y., Poon, W.Y., Bentler, P.: *Full maximum likelihood analysis of structural equation models with polytomous variables*, Statistics & Probability Letters 9(1), 91–97, 1990.
- Lindsay, B.: *Composite likelihood methods*. Contemporary Mathematics, 80: 221-239, 1988.

Some References

- Lubke, G., Neale, M.: *Distinguishing between latent classes and continuous factors with categorical outcomes: Class invariance of parameters of factor mixture models*, Multivariate Behavioral Research 43(4), 592–620, 2008.
- McParland, D., Gormley, I., Clark, S., McCormick, T., Kabudula, C., Collinson, M.: *Clustering south african households based on their asset status using latent variable models*, Tech. rep., Department of Statistics, University of Washington, 2012.
- Muthén, B.: *A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators*, Psychometrika 49(1), 115–132, 1984.
- Ranalli,M., Rocci,R.:*Mixture models for ordinal data: a pairwise likelihood approach*, Statistics and Computing, DOI:10.1007/s11222-014-9543-4, 2014.
- Ranalli,M., Rocci,R.:*Mixture models for mixed-type data through a composite likelihood approach*, Computational Statistics & Data Analysis, Volume 110, 87–102, 2017.
- Ranalli, M., Lagona, F., Picone, M., Zambianchi, E.: *Segmentation of sea current fields by cylindrical hidden Markov models: a composite likelihood approach.* Journal of the Royal Statistical Society: Series C. DOI: 10.1111/rssc.12240, (2017).
- Ranalli,M., Rocci,R.:*A Model-Based Approach to Simultaneous Clustering and Dimensional Reduction of Ordinal Data*, Psychometrika, DOI: 10.1007/s11336-017-9578-5, 2017.
- Varin, C., Vidoni, P. *A note on composite likelihood inference and model selection.* Biometrika **92**(3), 519–528, 2005.
- Varin, C., Reid, N. and Firth, D. *An overview of composite likelihood methods*. Statistica Sinica, 21(1): 1-41, 2011.

Thank you!