Model-based variable clustering

Vincent Vandewalle, Thierry Mottet, Matthieu Marbac

Model-based variable clustering

Vincent VANDEWALLE[1,2], Thierry MOTTET[2], Matthieu MARCBAC[3]

[1] Université Lille 2, EA 2694
[2] Inria
[3] ENSAI

ERCIM 2017
Sunday 17th December 2017
London

Outline

Data

$\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ composed of n independent observations $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})$ defined on $\mathbb{R}^d$.

Goal

Cluster the data in G clusters.

What variables use in clustering?

- Well-posed problem in the supervised classification setting with objective criteria: error rate, AUC, ...
- Ill-posed problem in clustering since the class variable is not known by advance. Thus what are the most relevant variables with respect to this unknown variable?
- Pragmatic solution 1: Prior choice of the practitioner among available variables (according to some focus)
- Pragmatic soltuion 2: Posterior analysis of the correlation between the predicted cluster (based on all the variables) and each variable

The model based clustering solution

- Mixture models allow to perform clustering by modelling the distribution of the data as a mixture of G components each one corresponding to a cluster.
- Thus possibility to suppose that some variables do not depend (directly) on the cluster in the probabilistic model.

Some references

- Raftery & Dean (2006): some classifying, and some redundant variables. Redundant variables independent of the cluster given the classifying variables.
- Maugis & *al.* (2009): refinement of Raftery & Dean (2006) by specifying the role of each variable.

Advantages of these approaches

- Improve the accuracy of the clustering by decreasing the variance of the estimators
- Allow some specific interpretation of the classifying variables

Limitations of these approaches

- Combinatorial problem to select the best model with these refined approaches
- Search too hard to perform when the number of variables is large

Solution: use simpler models for a better search (Marbac & Serdki 2016,2017)

- Assumption of conditional independence of the classifying variables given the cluster
- Non-classifying variables are independent
- Optimisation of the integrated classification likelihood (ICL)
- Better results than previous approaches on large number of variables with moderated sample size
- The independence assumption allows to easily consider the heterogeneous data setting

Several clustering variables

- The variables in the data can convey several clustering view points with respect to different groups of variables
- Allow to find some clustering which could be hidden by other variables

Some references

- Galimberti & *al.* (2007): First proposition of a multiple Gaussian mixture model
- Galimberti & *al.* (2017): Refinement of the previous model with ideas similar to to Raftery & Dean (2006) et Maugis & *al.* (2009)

Remarks

- Smart modelling of the role of each variable
- Search hard to perform when the number of variables is large
- Specific to the Gaussian setting

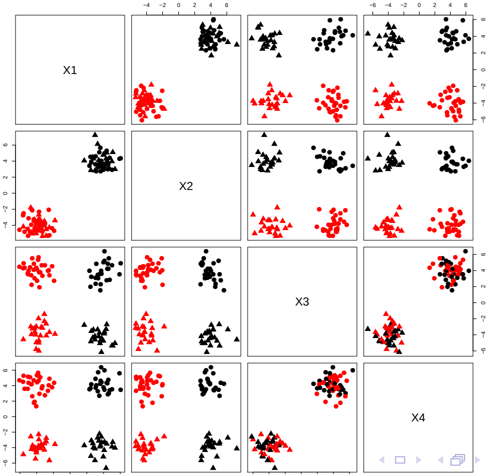
- B independent blocks.
- block b follows a Gaussian mixture with G_b components (for $b = 1, \ldots, B$), with the assumption of class conditional independence of the variables
- $\boldsymbol{\omega} = (\omega_j; j = 1, \ldots, d)$ the repartition of the variables in blocks; $\omega_j = b$ if variable j belongs to block b.
- The probability distribution function (pdf) of $\boldsymbol{x}_i$ is

$$p(\boldsymbol{x}_i|\boldsymbol{m}, \boldsymbol{\theta}) = \prod_{b=1}^{B} \sum_{g=1}^{G_b} \pi_{bg} \prod_{j \in \Omega_b} \phi(x_{ij}|\mu_{gj}, \sigma_{gj}^2),$$

  - $\boldsymbol{m} = (G_1, \ldots, G_B, \boldsymbol{\omega})$ defines the model
  - $\Omega_b = \{j : \omega_j = b\}$ the subset of variables belonging to block b
  - $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ model parameters
    - $\boldsymbol{\pi} = (\pi_{bg}; b = 1, \ldots, B; g = 1, \ldots, G_b)$ the proportions with $0 < \pi_{bg}$ and $\sum_{g=1}^{G_b} \pi_{bg} = 1$
    - $\boldsymbol{\mu} = (\mu_{gj}; g = 1, \ldots, G_{\omega_j}; j = 1, \ldots, d)$ the means
    - $\boldsymbol{\sigma} = (\sigma_{gj}; g = 1, \ldots, G_{\omega_j}; j = 1, \ldots, d)$ the standard deviations
  - $\phi(.|\mu_{gj}, \sigma_{gj}^2)$ the pdf of the Gaussian distribution with mean μ_{gj} and variance σ_{gj}^2.

Illustration :

- $n = 100$ from a MGMM with $B = 2$ blocks of two variables.
- Variable 1 and 2 belong to block 1 and variables 3 and 4 in block 2
- Each block follows a bi-component Gaussian mixture (*i.e.*, $G_b = 2$) with equal proportions (*i.e.*, $\pi_{bg} = 1/2$) and $\mu_{\mathbf{1}j} = 4$, $\mu_{\mathbf{2}j} = -4$ and $\sigma_{gj} = 1$.



Remarks

- Different partitions explained by subsets of variables.
- Generalizes approaches used for variable selection in model-based clustering (if $B = 2$ and $G_1 = 1$ then variables belonging to block 1 are not relevant for the clustering, while variables belonging to block 2 are relevant)
- MGMM permits **variable selection** and **multiple partitions** explained by **subsets of variables** (variables classification).
- Sparse model: number of parameters $\nu_{\boldsymbol{m}} = \sum_{b=1}^{B}(G_b - 1) + 2G_b\text{card}(\Omega_b)$
- Better model search expected than in the model of Galimberti & *al.* (2017)
- Natural extension to the heterogeneous data setting

Identifiability

Model identifiability is directly obtained from the identifiability of Gaussian mixture with local independence (Teicher, 1963, 1967).

Observed-data likelihood for sample x and model *m*

$$\ell(\boldsymbol{\theta}|\boldsymbol{m}, \mathbf{x}) = \sum_{b=1}^{B} \sum_{i=1}^{n} \ln \left(\sum_{g=1}^{G_b} \pi_{bg} \prod_{j \in \Omega_b} \phi(x_{ij}|\mu_{gj}, \sigma^2_{gj}) \right).$$

Observed-data likelihood for sample x and model *m*

- B independent mixtures
- $\mathbf{z} = (\boldsymbol{z}_{ib}; i = 1, \dots, n; b = 1, \dots, B)$ vectors of the component memberships
- $\boldsymbol{z}_{ib} = (z_{ib1}, \dots, z_{ibG_b})$ where $z_{ibg} = 1$ if observation i arose from component g for block b, and $z_{ibg} = 0$ otherwise

Completed-data likelihood for sample x and model *m*

$$\ell(\boldsymbol{\theta}|\boldsymbol{m}, \mathbf{x}, \mathbf{z}) = \sum_{b=1}^{B} f_{\boldsymbol{\pi}_b} + \sum_{j=1}^{d} f_j(\omega_j),$$

$f_{\boldsymbol{\pi}_b} = \sum_{i=1}^{n} \sum_{g=1}^{G_b} z_{ibg} \ln \pi_{bg}$ and $f_j(\omega_j) = \sum_{i=1}^{n} \sum_{g=1}^{G_{\omega_j}} z_{i\omega_j g} \ln \phi(x_{ij}|\mu_{\omega_j j}, \sigma^2_{\omega_j j})$.

EM algorithm

Starting from the initial value $\boldsymbol{\theta}^{[0]}$, iteration $[r]$ is composed of two steps:

E-step Computation of the fuzzy partitions $t_{ibg}^{[r]} := \mathbb{E}[Z_{ibg}|\boldsymbol{x}_i, \boldsymbol{m}, \boldsymbol{\theta}^{[r-1]}]$, hence

$$t_{ibg}^{[r]} = \frac{\pi_{bg}^{[r-1]} \prod_{j \in \Omega_b} \phi(x_{ij}|\mu_{gj}, \sigma_{gj}^2)}{\sum_{k=1}^{G_b} \pi_{bk}^{[r-1]} \prod_{j \in \Omega_b} \phi(x_{ij}|\mu_{kj}, \sigma_{kj}^2)},$$

M-step Maximization of the expected value of the complete-data log-likelihood over the parameters,

$$\pi_{bg}^{[r]} = \frac{n_{bg}^{[r]}}{n}, \mu_{gj}^{[r]} = \frac{1}{n_{\omega_j g}^{[r]}} \sum_{i=1}^{n} t_{i\omega_j g}^{[r]} x_{ij} \text{ and } \sigma_{gj}^{[r]} = \frac{1}{n_{\omega_j g}^{[r]}} \sum_{i=1}^{n} t_{i\omega_j g}^{[r]} (x_{ij} - \mu_{\omega_j j}^{[r]})^2.$$

Remarks

- Independence between the B blocks of variables permits to maximize the observed-data log-likelihood on each block separately.
- Possible modification to perform the block estimation and the parameter inference simultaneously.

Model collection $\mathcal{M}$

$$\mathcal{M} = \{\boldsymbol{m} : \omega_j \leq B_{\max} \text{ and } G_b \leq G_{\max}; j = 1, \dots, d; b = 1, \dots, B_{\max}\},$$

where $B_{\max}$ is the maximum number of blocks and $G_{\max}$ is the maximum number of components within block.

Model selection

Model selection often achieved by searching the model $\boldsymbol{m}^\star$ maximizing the BIC criterion which is defined by

$$\text{BIC}(\boldsymbol{m}) = \max_{\boldsymbol{\theta}_{\boldsymbol{m}}} \ell_{pen}(\boldsymbol{\theta}_{\boldsymbol{m}} | \boldsymbol{m}, \mathbf{x})$$

where

$$\ell_{pen}(\boldsymbol{\theta}_{\boldsymbol{m}} | \boldsymbol{m}, \mathbf{x}) = \ell(\boldsymbol{\theta}_{\boldsymbol{m}} | \boldsymbol{m}, \mathbf{x}) - \frac{\nu_{\boldsymbol{m}}}{2} \ln n.$$

Remark

$$(\boldsymbol{m}^\star, \hat{\boldsymbol{\theta}}_{\boldsymbol{m}^\star}) = \arg\max_{(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{m}})} \ell_{pen}(\boldsymbol{\theta}_{\boldsymbol{m}} | \boldsymbol{m}, \mathbf{x})$$

Combinatorial model selection through a modified the EM algorithm for B and $(G_1, \dots, G_B)$ fixed : choice of $\boldsymbol{m} \Leftrightarrow$ choice of $\boldsymbol{\omega}$

The EM algorithm to achieve $\arg\max_{(\boldsymbol{\omega},\boldsymbol{\theta})} \ell_{pen}(\boldsymbol{\theta}_{\boldsymbol{m}}|\boldsymbol{m}, \mathbf{x})$, starting from $(\boldsymbol{\omega}^{[0]}, \boldsymbol{\theta}^{[0]})$ is at iteration $[r]$:

E-step Computation of the fuzzy partitions $t_{ibg}^{[r]} := \mathbb{E}[Z_{ibg}|\boldsymbol{x}_i, \boldsymbol{m}, \boldsymbol{\theta}^{[r-1]}]$, hence

$$t_{ibg}^{[r]} := \frac{\pi_{bg}^{[r-1]} \prod_{j\in\Omega_b} \phi(x_{ij}|\mu_{gj}, \sigma_{gj}^2)}{\sum_{k=1}^{G_b} \pi_{bk}^{[r-1]} \prod_{j\in\Omega_b} \phi(x_{ij}|\mu_{kj}, \sigma_{kj}^2)},$$

M-step Maximization of the expected value of the complete-data log-likelihood over the parameters,

$$\pi_{bg}^{[r]} = \frac{n_{bg}^{[r]}}{n}, \omega_j^{[r]} = \arg\max_{b'=1,\dots,B} \Delta_{jb'}^{[r]}, \mu_{gj}^{[r]} = \frac{1}{n_{\omega_j^{[r]}g}^{[r]}} \sum_{i=1}^{n} t_{i\omega_j^{[r]}g}^{[r]} x_{ij}$$

$$\sigma_{gj}^{[r]} = \frac{1}{n_{\omega_j^{[r]}g}^{[r]}} \sum_{i=1}^{n} t_{i\omega_j^{[r]}g}^{[r]} (x_{ij} - \mu_{\omega_j^{[r]}j}^{[r]})^2,$$

with $\Delta_{jb}^{[r]} = \max_{(\mu_{gj},\sigma_{gj};g=1,\dots,G_b)} \sum_{i=1}^{n} \sum_{g=1}^{G_b} z_{ibg} \ln \phi(x_{ij}|\mu_{gj}, \sigma_{gj}^2) - G_b \ln n.$

Integrated complete-data likelihood

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}) = \int p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{m})d\boldsymbol{\theta}.$$

Assumptions

- Independence between the prior distributions
- Standard conjugate priors
- Closed form of the complete-data integrated likelihood

MICL (maximum integrated complete-data likelihood) criterion

$$\text{MICL}(\boldsymbol{m}) = \ln p(\mathbf{x}, \mathbf{z}^\star_{\boldsymbol{m}}|\boldsymbol{m}) \text{ with } \mathbf{z}^\star_{\boldsymbol{m}} = \arg\max_{\mathbf{z}_m} \ln p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}).$$

Thus

$$(\boldsymbol{m}^\star, \mathbf{z}^\star_{\boldsymbol{m}^\star}) = \arg\max_{(\boldsymbol{m}, \mathbf{z}_m)} \ln p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}).$$

Motivations

- Criteria based on the integrated complete-data likelihood are popular for model-based clustering
- Take into account the clustering purpose : model the data distribution and provide well-separated components

Optimisation of MICL over $(\mathbf{z}, \boldsymbol{\omega})$ for B and $G_1, \ldots, G_B$ fixed

Starting at the initial value $\boldsymbol{\omega}^{[0]}$, each ω_j is uniformly sampled among $\{1, \ldots, B\}$, the algorithm at iteration $[r]$ is

Partition step: find $\mathbf{z}^{[r]}$ such that for $b = 1, \ldots, B$

$$p(\mathbf{x}_{\{j|\omega_j^{[r]}=b\}}, \mathbf{z}_b^{[r]}|\boldsymbol{m}^{[r]}) \geq p(\mathbf{x}_{\{j|\omega_j^{[r]}=b\}}, \mathbf{z}_b^{[r-1]}|\boldsymbol{m}^{[r]}).$$

Model step: find $\boldsymbol{\omega}^{[r+1]}$ such that for $j = 1, \ldots, d$

$$\omega_j^{[r+1]} = \arg\max_{b \in \{1,\ldots,B\}} p(\mathbf{x}_j, \mathbf{z}_b^{[r]}|\omega_j = b)$$

Mixed dataset

- $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})$ is a vector of mixed variables (continuous, binary, count or categorical
- local independence within block $\Rightarrow$ extension to the mixed data analysis

Subset of the 1987 National Indonesia Contraceptive Prevalence Survey

1473 Indian women described by

- One continuous variable (AGE: age)
- One integer variable (Chi: number of children)
- Seven categorical variables (EL: education level, ELH: education level of the husband, Rel: religion, Oc: occupation, OcH: occupation of the husband, SLI: standard-of-living index and ME: media exposure).

Results obtained by the BIC criterion

Analysis with :

- maximum of three blocks : $B_{max} = 3$
- maximum of six components : $G_{max} = 6$

Age	Chi	EL	ELH	Rel	Oc	OcH	SLI	ME	G_1	G_2	BIC
1	1	2	2	2	1	2	2	2	6	3	-16078
1	1	2	2	2	1	2	2	2	5	3	-16081
1	1	2	2	2	2	2	2	2	4	3	-16088

Table: Best three models according to the BIC: block repartition, number of components per block and BIC values.

Adjusted Rand Index computed on the partitions obtained by blocks 1 and 2 equal to 0.01.

Results obtained by the MICL criterion

Age	Chi	EL	ELH	Rel	Oc	OcH	SLI	ME	G_1	G_2	BIC
1	1	1	1	1	2	1	1	1	4	1	-16293
1	1	1	1	1	2	1	1	1	5	1	-16301
1	1	1	1	1	1	1	1	1	4	.	-16307

Table: Best three models according to the MICL: block repartition, number of components per block and MICL values.

Conclusion

- Proposition of model-based clustering with several class variables, each one explaining the heterogeneity of a block of variables:
    - Find groups of variables producing the same clustering of the individuals
    - Interpret the clustering produced for each group of variables
- Model search performed simultaneously with parameters estimation
- Proposed model can be used in the heterogeneous data settings