



HAL
open science

Model-based variable clustering

Vincent Vandewalle, Thierry Mottet, Matthieu Marbac

► **To cite this version:**

Vincent Vandewalle, Thierry Mottet, Matthieu Marbac. Model-based variable clustering. CMStatistics/ERCIM 2017 - 10th International Conference of the ERCIM WG on Computational and Methodological Statistics, Dec 2017, London, United Kingdom. pp.1-19. hal-01691421

HAL Id: hal-01691421

<https://inria.hal.science/hal-01691421>

Submitted on 24 Jan 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Model-based variable clustering

Vincent VANDEWALLE^{1,2}, Thierry MOTTET², Matthieu MARCBAC³

¹ Université Lille 2, EA 2694

² Inria

³ ENSAI

ERCIM 2017

Sunday 17th December 2017

London

Outline

- 1 Extension of the variable selection to variable clustering
 - Variable selection in clustering
 - Multiple Gaussian Mixture
 - Proposed Multiple Mixture Model
 - Properties of the model
- 2 Parameters estimation and model selection
 - Maximum likelihood inference
 - Penalized observed-data likelihood
 - Integrated complete-data likelihood
- 3 Numerical experiments on real data
 - Mixed data framework
 - Results on real data

Data

$\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ composed of n independent observations $\mathbf{x}_i = (x_{i1}, \dots, x_{id})$ 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 solution 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 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

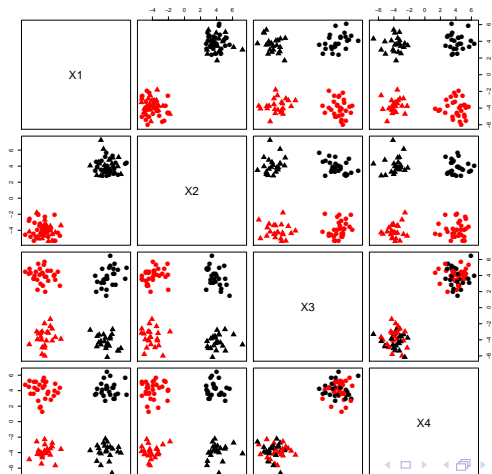
- B independent blocks.
- block b follows a Gaussian mixture with G_b components (for $b = 1, \dots, B$), with the assumption of class conditional independence of the variables
- $\omega = (\omega_j; j = 1, \dots, d)$ the repartition of the variables in blocks; $\omega_j = b$ if variable j belongs to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

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

- $\mathbf{m} = (G_1, \dots, G_B, \omega)$ defines the model
- $\Omega_b = \{j : \omega_j = b\}$ the subset of variables belonging to block b
- $\theta = (\pi, \mu, \sigma)$ model parameters
 - $\pi = (\pi_{bg}; b = 1, \dots, B; g = 1, \dots, G_b)$ the proportions with $0 < \pi_{bg}$ and $\sum_{g=1}^{G_b} \pi_{bg} = 1$
 - $\mu = (\mu_{gj}; g = 1, \dots, G_{\omega_j}; j = 1, \dots, d)$ the means
 - $\sigma = (\sigma_{gj}; g = 1, \dots, G_{\omega_j}; j = 1, \dots, d)$ the standard deviations
- $\phi(\cdot | \mu_{gj}, \sigma_{gj}^2)$ 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 $\mu_{1j} = 4$, $\mu_{2j} = -4$ 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 $\nu_m = \sum_{b=1}^B (G_b - 1) + 2G_b \text{card}(\Omega_b)$
- 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 \mathbf{x} and model \mathbf{m}

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

Observed-data likelihood for sample \mathbf{x} and model \mathbf{m}

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

Completed-data likelihood for sample \mathbf{x} and model \mathbf{m}

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

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

EM algorithm

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

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

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

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}

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

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 \mathbf{m}^* maximizing the BIC criterion which is defined by

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

where

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

Remark

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

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

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

E-step Computation of the fuzzy partitions $t_{ibg}^{[r]} := \mathbb{E}[Z_{ibg} | \mathbf{x}_i, \mathbf{m}, \theta^{[r-1]}]$, 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 $\Delta_{jb}^{[r]} = \max_{(\mu_{gj}, \sigma_{gj}^2; g=1, \dots, G_b)} \sum_{i=1}^n \sum_{g=1}^{G_b} z_{ibg} \ln \phi(x_{ij} | \mu_{gj}, \sigma_{gj}^2) - G_b \ln n$.

Integrated complete-data likelihood

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

Assumptions

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

MICL (maximum integrated complete-data likelihood) criterion

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

Thus

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

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}, ω) for B and G_1, \dots, G_B fixed

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

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

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

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

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

Mixed dataset

- $\mathbf{x}_i = (x_{i1}, \dots, x_{id})$ 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