



A tractable multi-partitions clustering

Vincent Vandewalle, Matthieu Marbac

► To cite this version:

Vincent Vandewalle, Matthieu Marbac. A tractable multi-partitions clustering. COMPSTAT 2018 - 23rd International Conference on Computational Statistics, Aug 2018, Iasi, Romania. hal-01956922

HAL Id: hal-01956922

<https://inria.hal.science/hal-01956922>

Submitted on 16 Dec 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.

A tractable Multi-Partitions Clustering

Vincent VANDEWALLE^{1,2}

Joint work with Matthieu MARBAC³

¹ Univ. Lille, EA2694 Santé publique: épidémiologie et qualité des soins

² Inria

³ CREST, ENSAI

COMPSTAT 2018

Friday 31th August 2018

Iasi

Outline

- 1 Extension of the variable selection to variable clustering
 - Variable selection in clustering
 - Multiple Gaussian Mixture
 - Proposed Multiple Partitions Mixture
 - Properties of the model
- 2 Parameters estimation and model selection
 - Maximum likelihood inference
 - Penalized observed-data likelihood
 - Integrated complete-data likelihood
- 3 Numerical experiments
 - NBA team data
 - Wine data
- 4 Conclusion and perspectives

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 (classifying, redundant or independant).

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

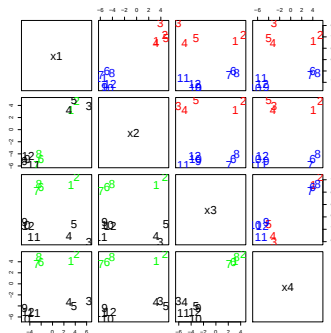
- Galimberti & *al.* (2007): First proposition of a multiple Gaussian mixture model
- Galimberti & *al.* (2017): Refinement of the previous model with ideas similar to Raftery & Dean (2006) et Maugis & *al.* (2009)

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

Illustration of a multiple partition

id	x_1	x_2	x_3	x_4	z_1	z_2
1	3.23	3.28	3.14	4.08	1	1
2	4.26	4.7	4.41	5.36	1	1
3	6.43	3.94	-6.09	-3.03	1	2
4	2.93	3.22	-5.05	-3.2	1	2
5	3.77	4.88	-3.21	-4	1	2
6	-2.03	-4.9	2.81	4	2	1
7	-2.78	-5.87	2.11	3.57	2	1
8	-2.38	-4.06	3.35	4.91	2	1
9	-4.88	-5.26	-2.86	-4.42	2	2
10	-5.01	-4.83	-3.3	-6.42	2	2
11	-3.25	-5.84	-5.23	-5.35	2	2
12	-4.28	-4.36	-3.38	-5.14	2	2



Main assumptions

- The variables can be decomposed in B independent blocks.
- The block b follows a mixture with G_b components (for $b = 1, \dots, B$), with the assumption of class conditional independence of the variables.

Notations

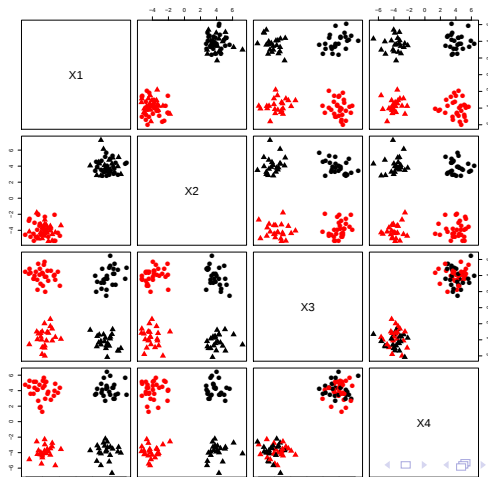
- $\omega = (\omega_j; j = 1, \dots, d)$ the repartition of the variables in blocks; $\omega_j = b$ if variable j belongs to block b .
- $\mathbf{m} = (G_1, \dots, G_B, \omega)$ defines the model
- $\Omega_b = \{j : \omega_j = b\}$ the subset of variables belonging to block b
- $\theta = (\pi, \alpha)$ model parameters
- $p(\cdot | \alpha_{jg})$ the pdf of the distribution of parameters α_{jg} .

Probability distribution function of $\mathbf{x}_i$

$$p(\mathbf{x}_i | \mathbf{m}, \theta) = \prod_{b=1}^B p(\mathbf{x}_{i\{\omega_b\}} | \mathbf{m}, \theta) \text{ with } p(\mathbf{x}_{i\{\omega_b\}} | \mathbf{m}, \theta) = \sum_{g=1}^{G_b} \pi_{bg} \prod_{j \in \Omega_b} p(x_{ij} | \alpha_{jg}),$$

Illustration :

- $n = 100$ from a MPM 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 $\mu_{j1} = 4$, $\mu_{j2} = -4$ and $\sigma_{jg} = 1$.



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 $\nu_m = \sum_{b=1}^B (G_b - 1) + G_b \sum_{j \in \Omega_b} \nu_j$
- 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 $\mathbf{m}$

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

Latent partitions

- 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 $\mathbf{m}$

$$\ell(\boldsymbol{\theta}|\mathbf{m}, \mathbf{x}, \mathbf{z}) = \sum_{b=1}^B \ln p(\mathbf{z}_b|\boldsymbol{\pi}_b) + \sum_{j=1}^d \ln p(\mathbf{x}_j|\mathbf{z}_{\omega_j}, \boldsymbol{\alpha}_j),$$

where $\ln p(\mathbf{z}_b|\boldsymbol{\pi}_b) = \sum_{i=1}^n \sum_{g=1}^{G_b} z_{ibg} \ln \pi_{bg}$ and

$\ln p(\mathbf{x}_j|\mathbf{z}_b, \boldsymbol{\alpha}_j) = \sum_{i=1}^n \sum_{g=1}^{G_b} z_{ibg} \ln p(x_{ij}|\boldsymbol{\alpha}_{jg}).$

EM algorithm

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

E-step Computation of the fuzzy partitions $t_{ibg}^{[r]} := \mathbb{E}[Z_{ibg} | \mathbf{x}_{i\{b\}}, \mathbf{m}, \theta^{[r-1]}]$,
 hence for $b = 1, \dots, B$, for $g = 1, \dots, G_b$, for $i = 1, \dots, n$

$$t_{ibg}^{[r]} = \frac{\pi_{bg}^{[r-1]} \prod_{j \in \Omega_b} p(x_{ij} | \alpha_{jg}^{[r-1]})}{\sum_{k=1}^{G_b} \pi_{bk}^{[r-1]} \prod_{j \in \Omega_b} p(x_{ij} | \alpha_{jk}^{[r-1]})},$$

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

$$\pi_{bg}^{[r]} = \frac{n_{bg}^{[r]}}{n} \text{ and } \alpha_{jg}^{[r]} = \arg \max_{\alpha_{jg} \in \Theta_j} Q(\alpha_{jg} | \mathbf{x}_j, \mathbf{t}_{\omega_{jg}}^{[r]}),$$

where $Q(\alpha_{jg} | \mathbf{x}_j, \mathbf{t}_b) = \sum_{i=1}^n t_{ibg} \ln p(x_{ij} | \alpha_{jg})$.

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.
- In practice the number of blocks, the repartition of variables into blocks, and the number of classes per block are unknown.

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

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

where

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

Remark

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

Penalised completed likelihood

$$\begin{aligned}\ell_{pen}(\theta_m | m, \mathbf{x}, \mathbf{z}) &= \ell(\theta_m | m, \mathbf{x}, \mathbf{z}) - \frac{\nu_m}{2} \log n \\ &= \sum_{b=1}^B \ln p(\mathbf{z}_b | \pi_b) - \frac{G_b - 1}{2} \ln n + \sum_{j=1}^d \ln p(\mathbf{x}_j | \mathbf{z}_{\omega_j}, \alpha_j) - \frac{\nu_j G_{\omega_j}}{2} \ln n,\end{aligned}$$

Consequence

- For $\mathbf{z}_b$ fixed: possibility to re-affect each variable **individually** to the most accurate block.
- Thus computationnaly attractive

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 for $b = 1, \dots, B$, for $g = 1, \dots, G_b$, for $i = 1, \dots, n$

$$t_{ibg}^{[r]} = \frac{\pi_{bg}^{[r-1]} \prod_{j \in \Omega_b^{[r-1]}} p(x_{ij} | \alpha_{jg}^{[r-1]})}{\sum_{k=1}^{G_b} \pi_{bk}^{[r-1]} \prod_{j \in \Omega_b^{[r-1]}} p(x_{ij} | \alpha_{jk}^{[r-1]})},$$

M-step1 Updating the affectation of the variables to blocks

$$\omega_j^{[r]} = \arg \max_{\omega_j \in \{1, \dots, B\}} \left(\sum_{g=1}^{G_{\omega_j}} \max_{\alpha_{jg} \in \Theta_j} Q(\alpha_{jg} | \mathbf{x}_j, \mathbf{t}_{\omega_j g}^{[r]}) - \frac{\nu_j G_{\omega_j}}{2} \ln n \right),$$

M-step2 Updating the model parameters

$$\pi_{bg}^{[r]} = \frac{n_{bg}^{[r]}}{n} \text{ and } \alpha_{jg}^{[r]} = \arg \max_{\alpha_{jg} \in \Theta_j} Q(\alpha_{jg} | \mathbf{x}_j, \mathbf{t}_{\omega_j g}^{[r]}).$$

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 $(\mathbf{z}, \omega)$ 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 $\mathbf{z}_b^{[r]}$ such that for all $b = 1, \dots, B$

$$p(\mathbf{x}_{\{b\}}^{[r-1]}, \mathbf{z}_b^{[r]}) \geq p(\mathbf{x}_{\{b\}}^{[r-1]}, \mathbf{z}_b^{[r-1]}),$$

where $\mathbf{x}_{\{b\}}^{[r-1]} = (\mathbf{x}_j; \omega^{[r-1]} = b)$.

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

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

Data description

NBA teams for the season 2016/2017 described by 16 numerical variables¹

- total minutes played (min)
- field goals made rate (fgmr)
- field goals attempted (fga)
- three-pointers made rate (3pmr)
- three-pointer attempted (3pa)
- free throws made (ftm)
- free throw attempted (fta)
- offensive rebounds (or)
- total rebounds (tr)
- assists (as)
- steals (st)
- turnovers (to)
- blocks (bk)
- personal fouls (pf)
- technical fouls (tc)
- points (pts)

Model selection with BIC

B	BIC	Time (s)	Block	G	variables
1	-1932	7	1	2	all the variables
2	-1915	47	1	3	fgmr, fga, 3pmr, 3pa, tr, as, to, pts
			2	1	min, ftmr, fta, orr, st, bk, pf, tc
3	-1909	170	1	2	fgmr, 3pmr, pf
			2	2	fga, 3pa, tr, as, st, to, pts
			3	1	min, ftmr, fta, orr, bk, tc

¹<http://www.dougstats.com/16-17RD.Team.0pp.txt>

Data description

NBA teams for the season 2016/2017 described by 16 numerical variables¹

- total minutes played (min)
- field goals made rate (fgmr)
- field goals attempted (fga)
- three-pointers made rate (3pmr)
- three-pointer attempted (3pa)
- free throws made (ftm)
- free throw attempted (fta)
- offensive rebounds (or)
- total rebounds (tr)
- assists (as)
- steals (st)
- turnovers (to)
- blocks (bk)
- personal fouls (pf)
- technical fouls (tc)
- points (pts)

Model selection with BIC

B	BIC	Time (s)	Block	G	variables
3	-1909	170	1	2	fgmr, 3pmr, pf
			2	2	fga, 3pa, tr, as, st, to, pts
			3	1	min, ftmr, fta, orr, bk, tc

¹<http://www.dougstats.com/16-17RD.Team.0pp.txt>

Model parameters

First block: offensive vs defensive teams

Three features: field goals made rate, three points made rate and personal fouls

		fmgr	3pmr	pf
offensive teams: ($\pi_{11} = 0.57$) high shooting ability low personal fouls	mean	0.468	0.371	1628.042
	sd	0.009	0.010	76.138
defensive teams: ($\pi_{12} = 0.43$) low shooting ability high personal fouls	mean	0.446	0.342	1635.503
	sd	0.005	0.010	173.701

Second block: two better statistics teams for general performances vs others

Seven features: field goal attempted, 3 points attempted, total rebounds, assists, steals, turnovers and points

		fga	3pa	tr	as	st	to	pts
GS Warriors: ($\pi_{11} = 0.07$) Houston Rockets	mean	7144	2934	3642	2281	728	1186	9481
	sd	3	372	3	211	59	3	22
Other: ($\pi_{11} = 0.93$) teams	mean	6992	2162	3564	1825	626	1091	8600
	sd	183	262	141	131	45	105	259

Third block: Six features detected as irrelevant for clustering.

Data description

Data¹

- 27 chemical and physical properties of three types of Italian wines: Barolo, Grignolino, Barbera
- Data collected during the time period of 1970–1979

Models selected by BIC

B	BIC	Time	Block	G	ARI
1	-6025.00	30	1	4	0.78
2	-5947.88	280	1	3	0.87
			2	4	0.16
3	-5921.42	1590	1	4	0.74
			2	4	0.20
			3	2	0.02
4	-5918.06	6065	1	4	0.75
			2	2	0.21
			3	3	0.02
			4	2	0.00

¹available in the package pgmm

Model interpretation

Block 1: the type of wines (ARI=0.75)

19 variables: Alcohol, Sugar-free Extract, Tartaric Acid, Uronic Acids, Alcalinity of Ash, Calcium, Magnesium, Phosphate, Total Phenols, Flavanoids, Non-flavanoid Phenols, Proanthocyanins, Color Intensity, Hue, OD280/OD315 of Diluted Wines, OD280/OD315 of Flavanoids, Glycerol, 2-3-Butanediol, Proline

	Barolo	Grignolino	Barbera
Class 1	0	45	0
Class 2	0	5	48
Class 3	58	1	0
Class 4	1	20	0

Block 2: year of production

4 variables: Fixed Acidity, Malic Acid, pH, Total Nitrogen

	Year								
	1970	1971	1972	1973	1974	1975	1976	1978	1979
Class 1	8	25	4	27	31	3	1	0	0
Class 2	1	3	3	2	14	6	16	29	5

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

Perspectives

- Consider the semi-supervised setting in the multi-partition framework for new partitions discovery: *i. e.* $\mathbf{z}_1$ known and $\mathbf{z}_2$ unknown.
- Extension to heterogeneous co-clustering by adding one level of latent variable to summarize the multi-partition by a single partition, while keeping the partition of the variables.
- Derive some k-means type multi-partition similar to Witten & Tibshirani (2010) in the variable selection framework to deal with the very high dimensional setting.