Variable selection for model-based clustering

Matthieu Marbac (Ensai - Crest)

Joint works with:

M. Sedki (Univ. Paris-sud) and V. Vandewalle (Univ. Lille 2)

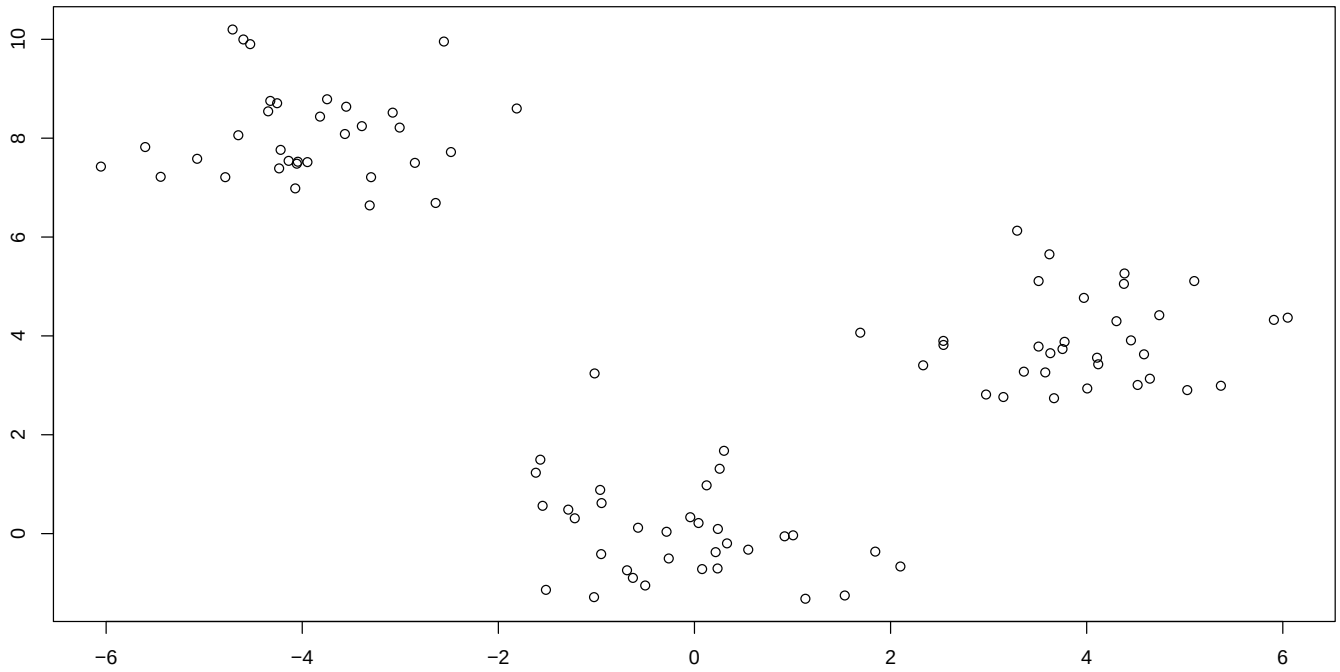
The problem

Objective:

- Estimation of a partition **z** among n observations **x**.

Notations:

- g: number of clusters.
- $\mathbf{x} = (\boldsymbol{x}_i; i = 1, \ldots, n)$: observed sample of n iid observations.
- $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})$: vector of the d features for observation i.
- $\mathbf{z} = (\boldsymbol{z}_i; i = 1, \ldots, n)$: unobserved partition.
- $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ig})$ with $z_{ik} = 1$ means that observation i belongs to cluster k.



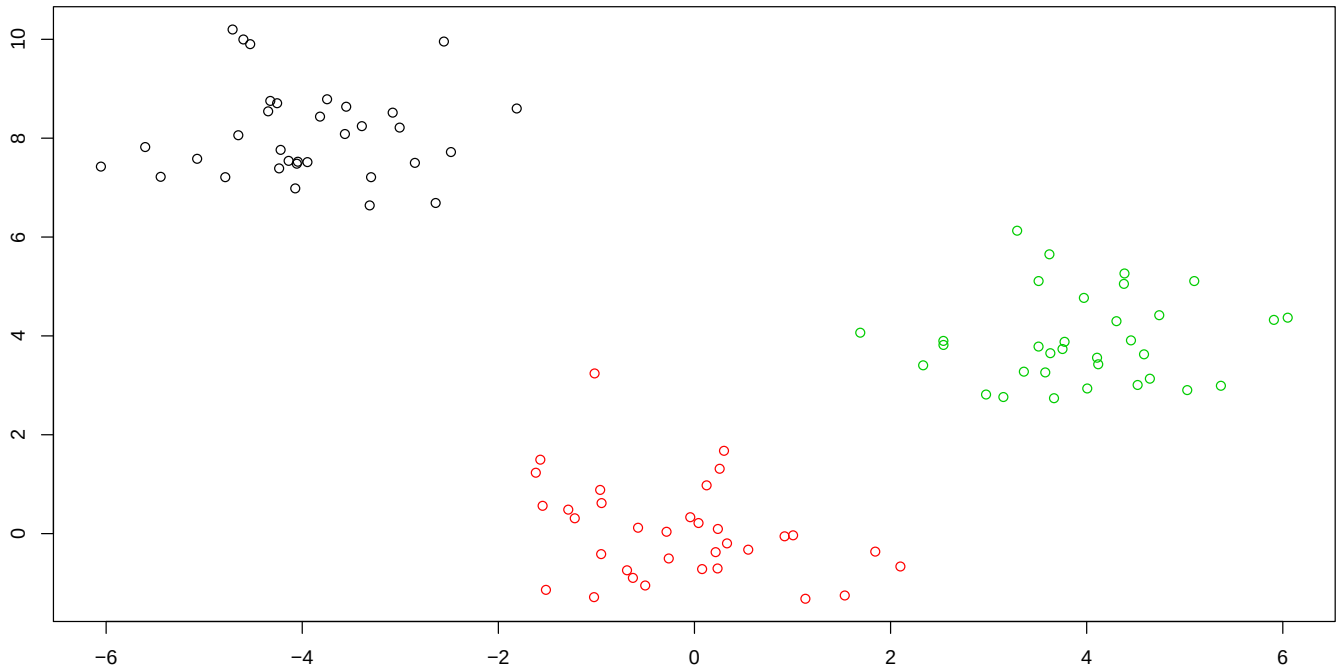
The problem

Objective:

- Estimation of a partition **z** among n observations **x**.

Notations:

- g: number of clusters.
- $\mathbf{x} = (\boldsymbol{x}_i; i = 1, \ldots, n)$: observed sample of n iid observations.
- $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})$: vector of the d features for observation i.
- $\mathbf{z} = (\boldsymbol{z}_i; i = 1, \ldots, n)$: unobserved partition.
- $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ig})$ with $z_{ik} = 1$ means that observation i belongs to cluster k.



The problem

The objectives

- Estimation of a classification rule (*i.e.*, estimator of $\boldsymbol{z}_i$ given $\boldsymbol{x}_i$).
- Evaluation of the risk of misclassification.
- Interpretation of the clusters.
- Estimation of the number of clusters g.
- Deal with complex data (mixed-type data, missing values, ...).

Model-based clustering

Main idea:
model the distribution of the observed data $\boldsymbol{X}$.

Generative model:

- $\boldsymbol{Z}_i \sim \mathcal{M}(\pi_1, \ldots, \pi_g)$
- $\boldsymbol{X}_i | Z_{ik} = 1 \sim \mathcal{L}_k(.)$, *e.g.*, $\mathcal{L}_k(.) = \mathcal{N}(\mu_k, \Sigma_k)$.

Mixture model:
The pdf of the observed data is

$$f(\boldsymbol{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k)$$

where $\boldsymbol{\theta}$ groups all the parameters.

Fuzzy and hard partition:

$$\mathbb{P}(Z_{ik} = 1 | \boldsymbol{X}_i = \boldsymbol{x}_i) = \frac{\pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k)}{\sum_{\ell=1}^{g} \pi_\ell f_\ell(\boldsymbol{x}_i; \boldsymbol{\theta}_\ell)}$$

The component membership of observation $\boldsymbol{x}_i$ is obtained by

$$\hat{z}_{ik^\star} = 1 \text{ if } k^\star = \arg\max_k \mathbb{P}(Z_{ik} = 1 | \boldsymbol{X}_i = \boldsymbol{x}_i)$$

Model-based clustering

Main idea:
model the distribution of the observed data $\boldsymbol{X}$.

Generative model:

- $\boldsymbol{Z}_i \sim \mathcal{M}(\pi_1, \ldots, \pi_g)$
- $\boldsymbol{X}_i | Z_{ik} = 1 \sim \mathcal{L}_k(.)$, *e.g.*, $\mathcal{L}_k(.) = \mathcal{N}(\mu_k, \Sigma_k)$.

Mixture model:
The pdf of the observed data is

$$f(\boldsymbol{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k)$$

where $\boldsymbol{\theta}$ groups all the parameters.

Fuzzy and hard partition:

$$\mathbb{P}(Z_{ik} = 1 | \boldsymbol{X}_i = \boldsymbol{x}_i) = \frac{\pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k)}{\sum_{\ell=1}^{g} \pi_\ell f_\ell(\boldsymbol{x}_i; \boldsymbol{\theta}_\ell)}$$

The component membership of observation $\boldsymbol{x}_i$ is obtained by

$$\hat{z}_{ik^\star} = 1 \text{ if } k^\star = \arg\max_k \mathbb{P}(Z_{ik} = 1 | \boldsymbol{X}_i = \boldsymbol{x}_i)$$

Model-based clustering

Main idea:
model the distribution of the observed data $\boldsymbol{X}$.

Generative model:

- $\boldsymbol{Z}_i \sim \mathcal{M}(\pi_1, \ldots, \pi_g)$
- $\boldsymbol{X}_i | Z_{ik} = 1 \sim \mathcal{L}_k(.)$, *e.g.*, $\mathcal{L}_k(.) = \mathcal{N}(\mu_k, \Sigma_k)$.

Mixture model:
The pdf of the observed data is

$$f(\boldsymbol{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k)$$

where $\boldsymbol{\theta}$ groups all the parameters.

Fuzzy and hard partition:

$$\mathbb{P}(Z_{ik} = 1 | \boldsymbol{X}_i = \boldsymbol{x}_i) = \frac{\pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k)}{\sum_{\ell=1}^{g} \pi_\ell f_\ell(\boldsymbol{x}_i; \boldsymbol{\theta}_\ell)}$$

The component membership of observation $\boldsymbol{x}_i$ is obtained by

$$\hat{z}_{ik^\star} = 1 \text{ if } k^\star = \arg\max_k \mathbb{P}(Z_{ik} = 1 | \boldsymbol{X}_i = \boldsymbol{x}_i)$$

Model-based clustering: inference

Maximum likelihood inference

From the sample $\mathbf{x} = (\boldsymbol{x}_i; i = 1, \ldots, n)$, we want $\hat{\boldsymbol{\theta}} = \text{argmax}\, \ell(\boldsymbol{\theta}; \mathbf{x})$ where

$$\ell(\boldsymbol{\theta}; \mathbf{x}) = \ln p(\mathbf{x}; \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \Big(\sum_{k=1}^{g} \pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k) \Big).$$

We consider the complete-data log-likelihood

$$\ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = \ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{g} z_{ik} \ln \Big(\pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k) \Big).$$

EM algorithm

- Introduced to deal with missing values (here $\mathbf{z}$ is missing).
- Iterative.
- Log-likelihood increases at each iteration.
- At iteration $[r]$, two steps:
  - E-step: computation of

$$t_{ik}^{[r-1]} := \mathbb{P}(Z_{ik} = 1 | \boldsymbol{x}_i, \boldsymbol{\theta}^{[r-1]}).$$

  - M-step: $\boldsymbol{\theta}^{[r]}$ maximizes the complete-data log-likelihood

$$\ln p(\mathbf{x}, \mathbf{t}^{[r-1]} | \boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{g} t_{ik}^{[r-1]} \ln \left(\pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k) \right).$$

Model-based clustering: inference

Maximum likelihood inference

From the sample $\mathbf{x} = (\boldsymbol{x}_i; i = 1, \ldots, n)$, we want $\hat{\boldsymbol{\theta}} = \text{argmax}\, \ell(\boldsymbol{\theta}; \mathbf{x})$ where

$$\ell(\boldsymbol{\theta}; \mathbf{x}) = \ln p(\mathbf{x}; \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \Big(\sum_{k=1}^{g} \pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k) \Big).$$

We consider the complete-data log-likelihood

$$\ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = \ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{g} z_{ik} \ln \Big(\pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k) \Big).$$

EM algorithm

- Introduced to deal with missing values (here $\mathbf{z}$ is missing).
- Iterative.
- Log-likelihood increases at each iteration.
- At iteration $[r]$, two steps:
  - E-step: computation of
  $$t_{ik}^{[r-1]} := \mathbb{P}(Z_{ik} = 1 | \boldsymbol{x}_i, \boldsymbol{\theta}^{[r-1]}).$$
  - M-step: $\boldsymbol{\theta}^{[r]}$ maximizes the complete-data log-likelihood
  $$\ln p(\mathbf{x}, \mathbf{t}^{[r-1]} | \boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{g} t_{ik}^{[r-1]} \ln \left(\pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k) \right).$$

Model-based clustering: model selection

Model
It is defined by the number of components, the family of the components, constraints among parameters...

Question
How can we perform model selection?

Standard approach
Define the set of competing models $\mathcal{M}$ by considering the maximum number of components $g_{\max}$. The selected model maximizes an information criterion.

Exhaustive approach
Computation of an information criterion for each model in $\mathcal{M}$.

Tools
Information criteria (BIC, ICL,...) which penalize the log-likelihood.

Model-based clustering: model selection

Model
It is defined by the number of components, the family of the components, constraints among parameters...

Question
How can we perform model selection?

Standard approach
Define the set of competing models $\mathcal{M}$ by considering the maximum number of components $g_{\max}$. The selected model maximizes an information criterion.

Exhaustive approach
Computation of an information criterion for each model in $\mathcal{M}$.

Tools
Information criteria (BIC, ICL,...) which penalize the log-likelihood.

Bayesian Information Criterion: a consistent criterion

With a uniform prior for $\boldsymbol{\omega} \in \mathcal{M}$:

$$p(\boldsymbol{\omega}|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{\omega}) \text{ où } p(\mathbf{x}|\boldsymbol{\omega}) = \int p(\mathbf{x}|\boldsymbol{\omega}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\omega}) d\boldsymbol{\theta}.$$

To assess $\ln p(\mathbf{x}|\boldsymbol{\omega})$, the BIC (Schwarz G., 1978) uses a Laplace approximation:

$$\text{BIC}(\boldsymbol{\omega}) = \ell(\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}; \boldsymbol{\omega}, \mathbf{x}) - \frac{\nu_{\boldsymbol{\omega}}}{2} \ln n,$$

where $\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ is the MLE and where $\nu_{\boldsymbol{\omega}}$ denotes the number of parameters of model $\boldsymbol{\omega}$.

To summarize

- Trade-off: Accuracy / Complexity.
- Consistent criterion (Keribin C., 2000).
- The MLE is needed.
- The clustering purpose is not considered.

Integrated Complete-data Likelihood: a criterion for clustering

The ICL (Biernacki C., Celeux G., Govaert G., 2000) considers the clustering purpose:

$$\text{ICL}(\boldsymbol{\omega}) = \ln p(\mathbf{x}, \hat{\mathbf{z}}|\boldsymbol{\omega}) \text{ où } p(\mathbf{x}, \mathbf{z}|\boldsymbol{\omega}) = \int p(\mathbf{x}, \mathbf{z}|\boldsymbol{\omega}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\omega}) d\boldsymbol{\theta},$$

where $\hat{\mathbf{z}}$ is the partition given by the MAP rule with $\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$.
$p(\mathbf{x}, \mathbf{z}|\boldsymbol{\omega})$ has closed-form when the mixture components belong to the exponential family ($+$ conjugate prior).

Other, we use a Laplace approximation

$$\text{ICL}(\boldsymbol{\omega}) \simeq \text{BIC}(\boldsymbol{\omega}) + \sum_{i=1}^{n} \sum_{k=1}^{g} \hat{z}_{ik} \ln t_{ik}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}).$$

To summarize

- Trade-off: Accuracy / Complexity / Overlap of clusters.
- The clustering purpose is considered.
- The MLE is needed.
- Non-consistent criterion.
- Robustness for model misspecification.

The problem

What variables should be considered 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.

Variable selection for model-based clustering

If a mixture of two isotropic Gaussians is considered then, the minimax expected loss R_n is, up to log factors (Azizyan, Singh and Wasserman, 2013),

$$R_n \approx \kappa^2 \sqrt{\frac{d}{n}}$$

where κ^{-1} is the signal-to-noise ratio.

Now, consider the case where there are $s < d$ relevant features, then

$$\kappa \sqrt{\frac{s\sqrt{\log d}}{\sqrt{n}}} \succ R_n \succ \kappa^2 \sqrt{\frac{s \log d}{n}}$$

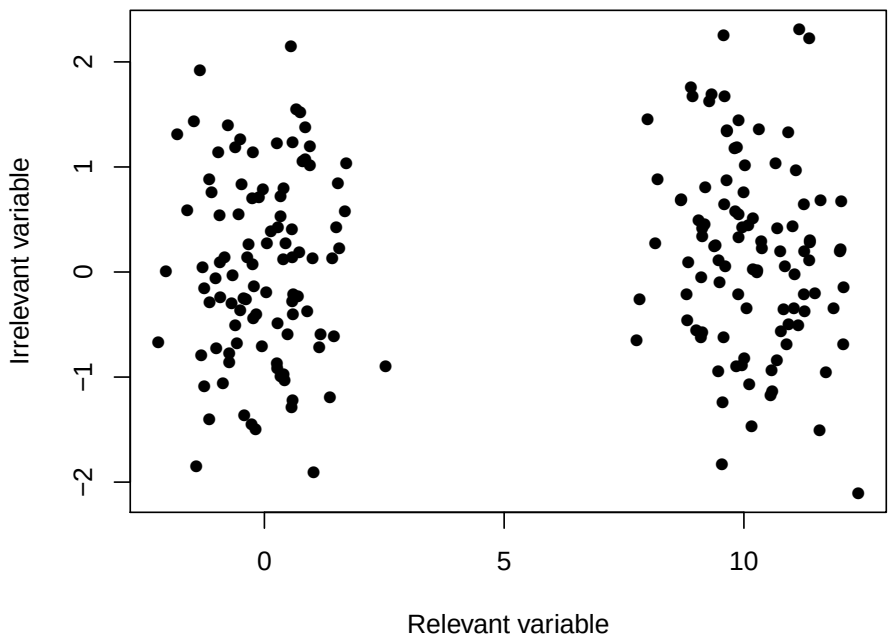
Variable selection in model-based clustering

Main idea

Only a subset of variables explains the partition (but the partition is not observed).

Advantages of variable selection

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



Variable selection in model-based clustering

Variable selection is mainly studied for Gaussian mixture

- Tadesse et al. (2005)

$$p(\boldsymbol{x}_i) = p(\boldsymbol{x}_i^I) \sum_{k=1}^{g} \pi_k p(\boldsymbol{x}_i^R | Z_{ik} = 1)$$

  $\boldsymbol{x}_i^I$ irrelevant variables and $\boldsymbol{x}_i^R$ relevant variables.

- Raftery and Dean (2006)

$$p(\boldsymbol{x}_i) = p(\boldsymbol{x}_i^I) \underbrace{p(\boldsymbol{x}_i^U | \boldsymbol{x}_i^R)}_{\text{linear regression}} \sum_{k=1}^{g} \pi_k p(\boldsymbol{x}_i^R | Z_{ik} = 1)$$

  $\boldsymbol{x}_i^R$ redundant variables (independent to the partition conditionally on $\boldsymbol{x}_i^R$.

- Maugis et al. (2009)

$$p(\boldsymbol{x}_i) = p(\boldsymbol{x}_i^I) \underbrace{p(\boldsymbol{x}_i^U | \boldsymbol{x}_i^R)}_{\text{sparse linear regression}} \sum_{k=1}^{g} \pi_k p(\boldsymbol{x}_i^R | Z_{ik} = 1)$$

Variable selection is difficult because they are many models can be considered!

Objectives

- Deal with many features
- Deal with mixed-type data (with missing values)

Solution: use simpler models for a better search

- Only two types of variables: relevant and irrelevant.
- Assumption of independence of the relevant variables given the cluster.
- Non-classifying variables are independent.
- Optimization of an information criterion (BIC or MICL).

Objectives

- Deal with many features
- Deal with mixed-type data (with missing values)

Solution: use simpler models for a better search

- Only two types of variables: relevant and irrelevant.
- Assumption of independence of the relevant variables given the cluster.
- Non-classifying variables are independent.
- Optimization of an information criterion (BIC or MICL).

Mixture with conditional independence

Conditional independence is not a big deal when the number of features is large.
Pmf of mixture model with conditional independence (Gaussian case)

$$f(\boldsymbol{x}_i|g,\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k \prod_{j=1}^{d} \phi(x_{ij}, \mu_{kj}, \sigma_{kj}^2).$$

d	3	4	5	10	20
True model	0.155	0.175	0.198	0.231	0.321
Locally indepence	0.111	0.114	0.118	0.101	0.125

Table: Misclassification rate obtained with $n = 50$ (best error 10%)

d	3	4	5	10	20
True model	0.101	0.101	0.104	0.106	0.110
Locally indepence	0.103	0.105	0.107	0.108	0.104

Table: Misclassification rate obtained with $n = 2000$ (best error 10%)

Variable selection for model-based clustering

Main idea

Only a subset of variables explains the partition (but the partition is not observed).

Twin objective

Clustering + Find the relevant variables

A variable is irrelevant for clustering if its distribution is the same among components.

Modeling the role of the variables

Model is defined by $\boldsymbol{\omega} = (\omega_j; j = 1, \ldots, d)$ with

$$\omega_j = \begin{cases} 0 & \text{if } X_j \text{ is irrelevant} \\ 1 & \text{if } X_j \text{ is relevant} \end{cases} \quad \text{thus} \quad \omega_j = \begin{cases} 0 & \text{if } \boldsymbol{\theta}_{1j} = \ldots = \boldsymbol{\theta}_{gj} \\ 1 & \text{otherwise} \end{cases}.$$

Pdf

$$f(\boldsymbol{x}_i; \boldsymbol{\omega}, \boldsymbol{\theta}) = \prod_{j \in \Omega^c} f_{1j}(x_{ij}; \boldsymbol{\theta}_{1j}) \sum_{k=1}^{g} \pi_k \prod_{j \in \Omega} f_{kj}(x_{ij}; \boldsymbol{\theta}_{kj}),$$

with $\Omega = \{j : \omega_j = 1\}$ et $\Omega^c = \{1, \ldots, d\} \setminus \Omega$.

Type of variable	$X_j \vert \boldsymbol{Z}_j = \boldsymbol{z}_j$	Priors	Jeffrey
categorical	$\mathcal{M}$	$\mathcal{D}$	yes
continuous	$\mathcal{N}$	$\mathcal{N}$ et $\mathcal{IG}$	no
count	$\mathcal{P}$	$\mathcal{G}a$	no

How to select ω (g is assumed to be known)?

Variable selection for model-based clustering

Main idea

Only a subset of variables explains the partition (but the partition is not observed).

Twin objective

Clustering + Find the relevant variables

A variable is irrelevant for clustering if its distribution is the same among components.

Modeling the role of the variables

Model is defined by $\boldsymbol{\omega} = (\omega_j; j = 1, \ldots, d)$ with

$$\omega_j = \begin{cases} 0 & \text{if } X_j \text{ is irrelevant} \\ 1 & \text{if } X_j \text{ is relevant} \end{cases} \quad \text{thus} \quad \omega_j = \begin{cases} 0 & \text{if } \boldsymbol{\theta}_{1j} = \ldots = \boldsymbol{\theta}_{gj} \\ 1 & \text{otherwise} \end{cases}.$$

Pdf

$$f(\boldsymbol{x}_i; \boldsymbol{\omega}, \boldsymbol{\theta}) = \prod_{j \in \Omega^c} f_{1j}(x_{ij}; \boldsymbol{\theta}_{1j}) \sum_{k=1}^{g} \pi_k \prod_{j \in \Omega} f_{kj}(x_{ij}; \boldsymbol{\theta}_{kj}),$$

with $\Omega = \{j : \omega_j = 1\}$ et $\Omega^c = \{1, \ldots, d\} \setminus \Omega$.

Type of variable	$X_j \vert \boldsymbol{Z}_j = \boldsymbol{z}_j$	Priors	Jeffrey
categorical	$\mathcal{M}$	$\mathcal{D}$	yes
continuous	$\mathcal{N}$	$\mathcal{N}$ et $\mathcal{IG}$	no
count	$\mathcal{P}$	$\mathcal{G}a$	no

How to select $\boldsymbol{\omega}$ (g is assumed to be known)?

Variable selection with BIC

Models in competition

$$\mathcal{M} = \{\boldsymbol{\omega};\ \boldsymbol{\omega} \in \{0,1\}^d\}.$$

g is supposed to be known (for ease of explanations).
If g is unknown, the procedure is repeated for each possible values of g.

Objective
From the sample $\mathbf{x}$, find $\boldsymbol{\omega}^\star$ maximizing the BIC

$$\boldsymbol{\omega}^\star = \underset{\boldsymbol{\omega}\in\mathcal{M}}{\operatorname{argmax}}\ \mathrm{BIC}(\boldsymbol{\omega}) \text{ with } \mathrm{BIC}(\boldsymbol{\omega}) = \ell(\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}; \boldsymbol{\omega}, \mathbf{x}) - \frac{\nu_{\boldsymbol{\omega}}}{2} \ln n,$$

where $\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ is the MLE and $\nu_{\boldsymbol{\omega}}$ is the number of parameters, for model $\boldsymbol{\omega}$.

Classical (but sub-optimal) approaches for maximizing BIC

- Exhaustive: no doable if d large $\mathrm{card}(\mathcal{M}) = 2^d$.
- Backward, forward: Raftery, A. and Dean, D. (2006).

Variable selection with BIC

Models in competition

$$\mathcal{M} = \{\boldsymbol{\omega};\ \boldsymbol{\omega} \in \{0,1\}^d\}.$$

g is supposed to be known (for ease of explanations).
If g is unknown, the procedure is repeated for each possible values of g.

Objective
From the sample $\mathbf{x}$, find $\boldsymbol{\omega}^\star$ maximizing the BIC

$$\boldsymbol{\omega}^\star = \underset{\boldsymbol{\omega}\in\mathcal{M}}{\operatorname{argmax}}\ \mathrm{BIC}(\boldsymbol{\omega}) \text{ with } \mathrm{BIC}(\boldsymbol{\omega}) = \ell(\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}; \boldsymbol{\omega}, \mathbf{x}) - \frac{\nu_{\boldsymbol{\omega}}}{2} \ln n,$$

where $\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ is the MLE and $\nu_{\boldsymbol{\omega}}$ is the number of parameters, for model $\boldsymbol{\omega}$.

Classical (but sub-optimal) approaches for maximizing BIC

- Exhaustive: no doable if d large $\mathrm{card}(\mathcal{M}) = 2^d$.
- Backward, forward: Raftery, A. and Dean, D. (2006).

Variable selection with BIC

Models in competition

$$\mathcal{M} = \{\boldsymbol{\omega};\ \boldsymbol{\omega} \in \{0,1\}^d\}.$$

g is supposed to be known (for ease of explanations).
If g is unknown, the procedure is repeated for each possible values of g.

Objective
From the sample $\mathbf{x}$, find $\boldsymbol{\omega}^\star$ maximizing the BIC

$$\boldsymbol{\omega}^\star = \underset{\boldsymbol{\omega}\in\mathcal{M}}{\operatorname{argmax}}\ \mathrm{BIC}(\boldsymbol{\omega}) \text{ with } \mathrm{BIC}(\boldsymbol{\omega}) = \ell(\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}; \boldsymbol{\omega}, \mathbf{x}) - \frac{\nu_{\boldsymbol{\omega}}}{2} \ln n,$$

where $\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ is the MLE and $\nu_{\boldsymbol{\omega}}$ is the number of parameters, for model $\boldsymbol{\omega}$.

Classical (but sub-optimal) approaches for maximizing BIC

- Exhaustive: no doable if d large $\mathrm{card}(\mathcal{M}) = 2^d$.
- Backward, forward: Raftery, A. and Dean, D. (2006).

Variable selection with BIC

Objective

$$\arg\max_{(\boldsymbol{\omega},\boldsymbol{\theta})} \ell(\boldsymbol{\theta};\boldsymbol{\omega},\mathbf{x}) - \frac{\nu_{\boldsymbol{\omega}}}{2}\ln n,$$

Penalized complete-data likelihood

$$\ell_p(\boldsymbol{\theta}|\boldsymbol{\omega},\mathbf{x},\mathbf{z}) = \sum_{j=1}^{d} v_j(\boldsymbol{\theta}_{1j},\ldots,\boldsymbol{\theta}_{gj}|\omega_j,\mathbf{x},\mathbf{z}) + \sum_{k=1}^{g}\sum_{i=1}^{n} z_{ik}\ln\pi_k - (g-1)\frac{\ln n}{2},$$

with $v_j(\boldsymbol{\theta}_{1j},\ldots,\boldsymbol{\theta}_{gj}|\omega_j,\mathbf{x},\mathbf{z})$ has closed form for $\omega_j = 1$ and $\omega_j = 0$ (and is easily maximized).

If X_j is binary, then

$$v_j(\boldsymbol{\theta}_{1j},\ldots,\boldsymbol{\theta}_{gj}|\omega_j,\mathbf{x},\mathbf{z}) = \begin{cases} \sum_{k=1}^{g}\sum_{i=1}^{n} z_{ik}\ln\left[(\alpha_{kj})^{x_{ij}}(1-\alpha_{kj})^{1-x_{ij}}\right] & \text{if } \omega_j = 1 \\ \sum_{i=1}^{n}\ln\left[(\alpha_{1j})^{x_{ij}}(1-\alpha_{1j})^{1-x_{ij}}\right] & \text{if } \omega_j = 0. \end{cases}$$

The penalized complete-data likelihood is easily maximized on $(\boldsymbol{\omega},\boldsymbol{\theta})$.

Variable selection with BIC

Objective

$$\arg\max_{(\boldsymbol{\omega},\boldsymbol{\theta})} \ell(\boldsymbol{\theta};\boldsymbol{\omega},\mathbf{x}) - \frac{\nu_{\boldsymbol{\omega}}}{2}\ln n,$$

Penalized complete-data likelihood

$$\ell_p(\boldsymbol{\theta}|\boldsymbol{\omega},\mathbf{x},\mathbf{z}) = \sum_{j=1}^{d} v_j(\boldsymbol{\theta}_{1j},\ldots,\boldsymbol{\theta}_{gj}|\omega_j,\mathbf{x},\mathbf{z}) + \sum_{k=1}^{g}\sum_{i=1}^{n} z_{ik}\ln\pi_k - (g-1)\frac{\ln n}{2},$$

with $v_j(\boldsymbol{\theta}_{1j},\ldots,\boldsymbol{\theta}_{gj}|\omega_j,\mathbf{x},\mathbf{z})$ has closed form for $\omega_j = 1$ and $\omega_j = 0$ (and is easily maximized).

If X_j is binary, then

$$v_j(\boldsymbol{\theta}_{1j},\ldots,\boldsymbol{\theta}_{gj}|\omega_j,\mathbf{x},\mathbf{z}) = \begin{cases} \sum_{k=1}^{g}\sum_{i=1}^{n} z_{ik}\ln\left[(\alpha_{kj})^{x_{ij}}(1-\alpha_{kj})^{1-x_{ij}}\right] & \text{if } \omega_j = 1 \\ \sum_{i=1}^{n}\ln\left[(\alpha_{1j})^{x_{ij}}(1-\alpha_{1j})^{1-x_{ij}}\right] & \text{if } \omega_j = 0. \end{cases}$$

The penalized complete-data likelihood is easily maximized on $(\boldsymbol{\omega},\boldsymbol{\theta})$.

Variable selection with BIC

Objective

$$\arg\max_{(\boldsymbol{\omega},\boldsymbol{\theta})} \ell(\boldsymbol{\theta};\boldsymbol{\omega},\mathbf{x}) - \frac{\nu_{\boldsymbol{\omega}}}{2}\ln n,$$

Penalized complete-data likelihood

$$\ell_p(\boldsymbol{\theta}|\boldsymbol{\omega},\mathbf{x},\mathbf{z}) = \sum_{j=1}^{d} v_j(\boldsymbol{\theta}_{1j},\ldots,\boldsymbol{\theta}_{gj}|\omega_j,\mathbf{x},\mathbf{z}) + \sum_{k=1}^{g}\sum_{i=1}^{n} z_{ik}\ln\pi_k - (g-1)\frac{\ln n}{2},$$

with $v_j(\boldsymbol{\theta}_{1j},\ldots,\boldsymbol{\theta}_{gj}|\omega_j,\mathbf{x},\mathbf{z})$ has closed form for $\omega_j = 1$ and $\omega_j = 0$ (and is easily maximized).

If X_j is binary, then

$$v_j(\boldsymbol{\theta}_{1j},\ldots,\boldsymbol{\theta}_{gj}|\omega_j,\mathbf{x},\mathbf{z}) = \begin{cases} \sum_{k=1}^{g}\sum_{i=1}^{n} z_{ik}\ln\left[(\alpha_{kj})^{x_{ij}}(1-\alpha_{kj})^{1-x_{ij}}\right] & \text{if } \omega_j = 1 \\ \sum_{i=1}^{n}\ln\left[(\alpha_{1j})^{x_{ij}}(1-\alpha_{1j})^{1-x_{ij}}\right] & \text{if } \omega_j = 0. \end{cases}$$

The penalized complete-data likelihood is easily maximized on $(\boldsymbol{\omega},\boldsymbol{\theta})$.

Variable selection with BIC

The penalized complete-data likelihood is a sum over the variables.

EM Algorithm for assessing the model and its parameters

- EM algorithm for maximizing $\ell(\boldsymbol{\theta}; \boldsymbol{\omega}, \mathbf{x}) - \frac{\nu_{\boldsymbol{\omega}}}{2} \ln n$ on $(\boldsymbol{\omega}, \boldsymbol{\theta})$.
- E-step: unchanged.
- M-step: update ω_j and $\boldsymbol{\theta}_{1j}, \ldots, \boldsymbol{\theta}_{gj}$ for each j independently.

This algorithm can be used to perform variable selection according to any IC defined by

$$\ell_p(\boldsymbol{\theta}|\boldsymbol{\omega}, \mathbf{x}) = \ell(\boldsymbol{\theta}|\boldsymbol{\omega}, \mathbf{x}) - \nu_{\boldsymbol{\omega}} c,$$

where c is a constant.

Some experiences

	Data		VarSelLCM			clustvarsel (forward/backward)		
	n	d	NRV	ARI	Time	NRV	ARI	Time
banknote	200	6	5	0.96	1	4/4	0.98/0.98	5/4
coffee	43	12	5	1.00	1	5/6	1.00/1.00	5/13
cancer	569	30	28	0.68	16	17/19	0.66/0.64	2160/14435
golub	83	3051	769	0.70	14	8/•	0.11/•	3414/•

Table: Results obtained on four benchmark datasets: number of relevant variables (NRV), Adjusted Rand Index computed on the selected model (ARI) and computing time in seconds (Time)

Asymptotic and exact criteria

Bayesian model selection
The aim is to find the model maximizing the integrated likelihood:

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

Asymptotic criterion
BIC is asymptotic because

$$\ln p(\mathbf{x}|\boldsymbol{\omega}) = \ln p(\mathbf{x}|\boldsymbol{\omega}, \hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}) - \frac{\nu_{\boldsymbol{\omega}}}{2} \ln n + \mathcal{O}(1),$$

but its approximation can leads to poor results if $n << d$.
$\Rightarrow$ An exact criterion is more relevant.

Clustering purpose
In clustering, we want to obtain well-separated classes.
$\Rightarrow$ A criterion considering the entropy between components should be used.

Variable selection with MICL

Integrated complete-data likelihood

The complete data likelihood has a closed form if conjugated priors are used

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

For the following Jeffrey's priors are used (Dirichlet with hyper-parameters 1/2).

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{\omega}) = p(\mathbf{z}) \prod_{j=1}^{d} p(\mathbf{x}_{\bullet j}|\omega_j, \mathbf{z}),$$

where $p(\mathbf{x}_{\bullet j}|\omega_j, \mathbf{z})$ has closed form for $\omega_j = 1$ and $\omega_j = 0$

If X_j is binary,

$$p(\mathbf{x}_{\bullet j}|g, \omega_j, \mathbf{z}) = \begin{cases} \left(\frac{1}{\Gamma(\frac{1}{2})^2}\right)^g \prod_{k=1}^{g} \frac{\Gamma(n_k - n_{k1} + \frac{1}{2})\Gamma(n_{k1} + \frac{1}{2})}{\Gamma(n_k + 1)} & \text{si } \omega_j = 1 \\ \frac{1}{\Gamma(\frac{1}{2})^2} \frac{\Gamma(n - \sum_{i=1}^n x_{ij} + \frac{1}{2})\Gamma(\sum_{i=1}^n x_{ij} + \frac{1}{2})}{\Gamma(n+1)} & \text{si } \omega_j = 0 \end{cases}$$

où $n_{k1} = \sum_{i=1}^n z_{ik} x_{ij}$.

Variable selection with MICL

Integrated complete-data likelihood

The complete data likelihood has a closed form if conjugated priors are used

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

For the following Jeffrey's priors are used (Dirichlet with hyper-parameters 1/2).

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{\omega}) = p(\mathbf{z}) \prod_{j=1}^{d} p(\mathbf{x}_{\bullet j}|\omega_j, \mathbf{z}),$$

where $p(\mathbf{x}_{\bullet j}|\omega_j, \mathbf{z})$ has closed form for $\omega_j = 1$ and $\omega_j = 0$

If X_j is binary,

$$p(\mathbf{x}_{\bullet j}|g, \omega_j, \mathbf{z}) = \begin{cases} \left(\frac{1}{\Gamma(\frac{1}{2})^2}\right)^g \prod_{k=1}^{g} \frac{\Gamma\left(n_k - n_{k1} + \frac{1}{2}\right)\Gamma\left(n_{k1} + \frac{1}{2}\right)}{\Gamma\left(n_k + 1\right)} & \text{si } \omega_j = 1 \\ \frac{1}{\Gamma(\frac{1}{2})^2} \frac{\Gamma\left(n - \sum_{i=1}^{n} x_{ij} + \frac{1}{2}\right)\Gamma\left(\sum_{i=1}^{n} x_{ij} + \frac{1}{2}\right)}{\Gamma\left(n + 1\right)} & \text{si } \omega_j = 0 \end{cases}$$

où $n_{k1} = \sum_{i=1}^{n} z_{ik} x_{ij}$.

Variable selection with MICL

The complete-data likelihood is

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{\omega}) = p(\mathbf{z}) \prod_{j=1}^{d} p(\mathbf{x}_{\bullet j}|\omega_j, \mathbf{z}),$$

Criterion requiring the computation of the MLE

$$\text{ICL}(\boldsymbol{\omega}) = \ln p(\mathbf{x}, \hat{\mathbf{z}}_{\boldsymbol{\omega}}|\boldsymbol{\omega}) \text{ with } \hat{\mathbf{z}}_{\boldsymbol{\omega}} = \arg\max_{\mathbf{z}} \ln p(\mathbf{z}|\boldsymbol{\omega}, \mathbf{x}, \hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}).$$

Criterion avoiding the computation of the MLE

$$\text{MICL}(\boldsymbol{\omega}) = \ln p(\mathbf{x}, \mathbf{z}^{\star}_{\boldsymbol{\omega}}|\boldsymbol{\omega}) \text{ with } \mathbf{z}^{\star}_{\boldsymbol{\omega}} = \arg\max_{\mathbf{z}} \ln p(\mathbf{z}|\boldsymbol{\omega}, \mathbf{x}).$$

Variable selection with MICL

Objective

Maximization of MICL implies the maximization on $(\boldsymbol{\omega}, \mathbf{z})$ of

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{\omega}) = p(\mathbf{z}) \prod_{j=1}^{d} p(\mathbf{x}_{\bullet j}|\omega_j, \mathbf{z}).$$

$\Rightarrow$ Model selection without parameter estimation!

Algorithm for assessing the model

An iterative algorithm of optimization is used, where iteration $[r]$ is:

- **Partition step (iterative):**
  Find $\mathbf{z}^{[r]}$ with
  $$\ln p(\mathbf{x}, \mathbf{z}^{[r]}|\boldsymbol{\omega}^{[r]}) \geq \ln p(\mathbf{x}, \mathbf{z}^{[r-1]}|\boldsymbol{\omega}^{[r]}).$$
- **Model step (explicit):**
  Find $\boldsymbol{\omega}^{[r+1]} = \underset{\boldsymbol{\omega} \in \mathcal{M}}{\operatorname{argmax}} \ln p(\mathbf{x}, \mathbf{z}^{[r]}|\boldsymbol{\omega})$, where
  $$\omega_j^{[r+1]} = \underset{\omega_j \in \{0,1\}}{\operatorname{argmax}} p(\mathbf{x}_{\bullet j}|\omega_j, \mathbf{z}^{[r]}).$$

Properties

- The objective function increases at each iteration.
- Convergence to local optimum (like for EM algorithm).

Variable selection with MICL

Objective

Maximization of MICL implies the maximization on $(\boldsymbol{\omega}, \mathbf{z})$ of

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{\omega}) = p(\mathbf{z}) \prod_{j=1}^{d} p(\mathbf{x}_{\bullet j}|\omega_j, \mathbf{z}).$$

$\Rightarrow$ Model selection without parameter estimation!

Algorithm for assessing the model

An iterative algorithm of optimization is used, where iteration $[r]$ is:

- **Partition step (iterative):**
  Find $\mathbf{z}^{[r]}$ with
  $$\ln p(\mathbf{x}, \mathbf{z}^{[r]}|\boldsymbol{\omega}^{[r]}) \geq \ln p(\mathbf{x}, \mathbf{z}^{[r-1]}|\boldsymbol{\omega}^{[r]}).$$
- **Model step (explicit):**
  Find $\boldsymbol{\omega}^{[r+1]} = \underset{\boldsymbol{\omega} \in \mathcal{M}}{\operatorname{argmax}} \ln p(\mathbf{x}, \mathbf{z}^{[r]}|\boldsymbol{\omega})$, where
  $$\omega_j^{[r+1]} = \underset{\omega_j \in \{0,1\}}{\operatorname{argmax}} p(\mathbf{x}_{\bullet j}|\omega_j, \mathbf{z}^{[r]}).$$

Properties

- The objective function increases at each iteration.
- Convergence to local optimum (like for EM algorithm).

Variable selection with MICL

Objective

Maximization of MICL implies the maximization on $(\boldsymbol{\omega}, \mathbf{z})$ of

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{\omega}) = p(\mathbf{z}) \prod_{j=1}^{d} p(\mathbf{x}_{\bullet j}|\omega_j, \mathbf{z}).$$

$\Rightarrow$ Model selection without parameter estimation!

Algorithm for assessing the model

An iterative algorithm of optimization is used, where iteration $[r]$ is:

- **Partition step (iterative):**
  Find $\mathbf{z}^{[r]}$ with
  $$\ln p(\mathbf{x}, \mathbf{z}^{[r]}|\boldsymbol{\omega}^{[r]}) \geq \ln p(\mathbf{x}, \mathbf{z}^{[r-1]}|\boldsymbol{\omega}^{[r]}).$$
- **Model step (explicit):**
  Find $\boldsymbol{\omega}^{[r+1]} = \underset{\boldsymbol{\omega} \in \mathcal{M}}{\operatorname{argmax}} \ln p(\mathbf{x}, \mathbf{z}^{[r]}|\boldsymbol{\omega})$, where
  $$\omega_j^{[r+1]} = \underset{\omega_j \in \{0,1\}}{\operatorname{argmax}} p(\mathbf{x}_{\bullet j}|\omega_j, \mathbf{z}^{[r]}).$$

Properties

- The objective function increases at each iteration.
- Convergence to local optimum (like for EM algorithm).

Details on the partition step

At iteration $[r]$: $\mathbf{z}^{[r]}$ is initialized $\mathbf{z}^{(0)} = \mathbf{z}^{[r-1]}$. S iterations are made where iteration (s) is composed of two steps:

- **Sampling:** $i^{(s)} \sim \mathcal{U}\{1, \ldots, n\}$.
- **Optimization:** hold $\mathcal{Z}^{(s)} = \{\mathbf{z} : \boldsymbol{z}_i = \boldsymbol{z}_i^{(s-1)}, \ \forall i \neq i^{(s)}\}$ and find $\mathbf{z}^{(s)}$ such that

$$\mathbf{z}^{(s)} = \text{argmax}_{\mathbf{z} \in \mathcal{Z}^{(s)}} \ln p(\mathbf{x}, \mathbf{z} | \boldsymbol{\omega}^{[r]}).$$

Stopping criterion : $\ln p(\mathbf{x}, \mathbf{z} | \boldsymbol{\omega}^{[r]})$ **does not increase after** S **iterations**

Summary

- Two fast and simple steps.
- $\ln p(\mathbf{x}, \mathbf{z} | \boldsymbol{\omega})$ increases at each iteration.
- Many random starting points are needed.

Details on the partition step

At iteration $[r]$: $\mathbf{z}^{[r]}$ is initialized $\mathbf{z}^{(0)} = \mathbf{z}^{[r-1]}$. S iterations are made where iteration (s) is composed of two steps:

- **Sampling:** $i^{(s)} \sim \mathcal{U}\{1, \ldots, n\}$.
- **Optimization:** hold $\mathcal{Z}^{(s)} = \{\mathbf{z} : \boldsymbol{z}_i = \boldsymbol{z}_i^{(s-1)}, \ \forall i \neq i^{(s)}\}$ and find $\mathbf{z}^{(s)}$ such that

$$\mathbf{z}^{(s)} = \text{argmax}_{\mathbf{z} \in \mathcal{Z}^{(s)}} \ln p(\mathbf{x}, \mathbf{z} | \boldsymbol{\omega}^{[r]}).$$

Stopping criterion : $\ln p(\mathbf{x}, \mathbf{z} | \boldsymbol{\omega}^{[r]})$ **does not increase after** S **iterations**

Summary

- Two fast and simple steps.
- $\ln p(\mathbf{x}, \mathbf{z} | \boldsymbol{\omega})$ increases at each iteration.
- Many random starting points are needed.

Some experiences

	Data		VarSelLCM (BIC / MICL)			clustvarsel (forward/backward)		
	n	d	NRV	ARI	Time	NRV	ARI	Time
banknote	200	6	5/5	0.96/0.96	1/1	4/4	0.98/0.98	5/4
coffee	43	12	5/5	1.00/1.00	1/1	5/6	1.00/1.00	5/13
cancer	569	30	28/15	0.68/0.75	18/24	17/19	0.66/0.64	2160/14435
golub	83	3051	769/553	0.70/0.79	16/15	8/•	0.11/•	3414/•

Table: Results obtained on four benchmark datasets: number of relevant variables (NRV), Adjusted Rand Index computed on the selected model (ARI) and computing time in seconds (Time)

EDEN

Data

- Obseity prevention
- Data collected on 1436 individuals between 2 and 5 years
- 44 features (continuous and categorical, with missing values) describing:
  - Dietary intake and eating habits
  - Physical activity measures
  - Sedentary behavior measures
  - Sleep measures

Results

- 17/44 features are detected as discriminative with two clusters.
- The discriminative features mainly describe dietary intake and eating habits.
- Whats about the other families of features?

EDEN

Data

- Obseity prevention
- Data collected on 1436 individuals between 2 and 5 years
- 44 features (continuous and categorical, with missing values) describing:
  - Dietary intake and eating habits
  - Physical activity measures
  - Sedentary behavior measures
  - Sleep measures

Results

- 17/44 features are detected as discriminative with two clusters.
- The discriminative features mainly describe dietary intake and eating habits.
- Whats about the other families of features?

Multiple partitions in model-based clustering

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

Main assumptions

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

Notations

- $\boldsymbol{\omega} = (\omega_j; j = 1, \ldots, d)$ the repartition of the variables in blocks; $\omega_j = b$ if variable j belongs to block b.
- $\boldsymbol{m} = (g_1, \ldots, g_B, \boldsymbol{\omega})$ defines the model
- $\Omega_b = \{j : \omega_j = b\}$ the subset of variables belonging to block b

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

$$p(\boldsymbol{x}_i; \boldsymbol{m}, \boldsymbol{\theta}) = \prod_{b=1}^{B} p(\boldsymbol{x}_{i\{b\}}; \boldsymbol{m}, \boldsymbol{\theta}) \text{ with } p(\boldsymbol{x}_{i\{b\}}; \boldsymbol{m}, \boldsymbol{\theta}) = \sum_{k=1}^{g_b} \pi_{bk} \prod_{j \in \Omega_b} p(x_{ij} | \boldsymbol{\theta}_{jk}),$$

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

Main assumptions

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

Notations

- $\boldsymbol{\omega} = (\omega_j; j = 1, \ldots, d)$ the repartition of the variables in blocks; $\omega_j = b$ if variable j belongs to block b.
- $\boldsymbol{m} = (g_1, \ldots, g_B, \boldsymbol{\omega})$ defines the model
- $\Omega_b = \{j : \omega_j = b\}$ the subset of variables belonging to block b

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

$$p(\boldsymbol{x}_i; \boldsymbol{m}, \boldsymbol{\theta}) = \prod_{b=1}^{B} p(\boldsymbol{x}_{i\{b\}}; \boldsymbol{m}, \boldsymbol{\theta}) \text{ with } p(\boldsymbol{x}_{i\{b\}}; \boldsymbol{m}, \boldsymbol{\theta}) = \sum_{k=1}^{g_b} \pi_{bk} \prod_{j \in \Omega_b} p(x_{ij} | \boldsymbol{\theta}_{jk}),$$

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

Main assumptions

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

Notations

- $\boldsymbol{\omega} = (\omega_j; j = 1, \ldots, d)$ the repartition of the variables in blocks; $\omega_j = b$ if variable j belongs to block b.
- $\boldsymbol{m} = (g_1, \ldots, g_B, \boldsymbol{\omega})$ defines the model
- $\Omega_b = \{j : \omega_j = b\}$ the subset of variables belonging to block b

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

$$p(\boldsymbol{x}_i; \boldsymbol{m}, \boldsymbol{\theta}) = \prod_{b=1}^{B} p(\boldsymbol{x}_{i\{b\}}; \boldsymbol{m}, \boldsymbol{\theta}) \text{ with } p(\boldsymbol{x}_{i\{b\}}; \boldsymbol{m}, \boldsymbol{\theta}) = \sum_{k=1}^{g_b} \pi_{bk} \prod_{j \in \Omega_b} p(x_{ij} | \boldsymbol{\theta}_{jk}),$$

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

$$p(\boldsymbol{x}_i; \boldsymbol{m}, \boldsymbol{\theta}) = \prod_{b=1}^{B} \left(\sum_{k=1}^{g_b} \pi_{bk} \prod_{j \in \Omega_b} p(x_{ij} | \boldsymbol{\theta}_{jk}) \right)$$

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).
- Natural extension to the heterogeneous data setting.
- Model identifiability is directly obtained from the identifiability of Gaussian mixture with local independence (Teicher, 1963, 1967).
- Model selection (*i.e.*, find the blocks of variables) can be carried out via modified EM algorithm (BIC) or via the optimization of the integrated complete-data likelihood (MICL).

EDEN (continued)

Data

- Obesity prevention
- Data collected on 1436 individuals between 2 and 5 years
- 44 features (continuous and categorical, with missing values) describing:
  - Dietary intake and eating habits
  - Physical activity measures
  - Sedentary behavior measures
  - Sleep measures

Results

- Best model considers 3 blocks of variables.
- 17/44 features are grouped into block 1 (mainly dietary intake and eating habits) with two components.
- 7/44 features are grouped into block 2 (mainly physical activity measures and sedentary behavior measures) with two components.
- 20/44 features are grouped into block 3 which contains non-discriminative variables.

EDEN (continued)

Data

- Obesity prevention
- Data collected on 1436 individuals between 2 and 5 years
- 44 features (continuous and categorical, with missing values) describing:
  - Dietary intake and eating habits
  - Physical activity measures
  - Sedentary behavior measures
  - Sleep measures

Results

- Best model considers 3 blocks of variables.
- 17/44 features are grouped into block 1 (mainly dietary intake and eating habits) with two components.
- 7/44 features are grouped into block 2 (mainly physical activity measures and sedentary behavior measures) with two components.
- 20/44 features are grouped into block 3 which contains non-discriminative variables.

Model-based clustering

Take home message

- Conditional independence facilitates variable selection...
- ... but if d is small and n large enough, use more complex models!

VarSelLCM

- Available on CRAN: `https://cran.r-project.org/web/packages/VarSelLCM/`
- Tutorial: `http://varsellcm.r-forge.r-project.org/`
  - Clustering of mixed data (continuous, categorical and integer) with missing values
  - Variable selection (BIC or MICL)

References

- Marbac, M. and Sedki, M. Variable selection for model-based clustering using the integrated complete-data likelihood. Statistics and Computing, 27 (4), 2017
- Marbac, M. and Sedki, M. VarSelLCM: an R/C++ package for variable selection in model-based clustering of mixed-data with missing values. Bioinformatics, 2018
- Marbac, M. and Patin, E. and Sedki, M. Variable selection for mixed data clustering: Application in human population genomics. Journal of Classification, (forthcoming)
- Marbac, M., and Vandewalle, V. A tractable Multi-Partitions Clustering. Computational Statistics and Data Analysis, (forthcoming)